Explorando en el big data. El ejemplo de datificación de la encuesta de hogares de Chile

Dra. Paulina Benítez
Investigadora independiente
paulinabenitez14@gmail.com

Resumen

La ingente cantidad de datos (texto, imágenes, sonido, números, videos) que hay en el ciberespacio se duplica cada tres años siguiendo su propia Ley de Moore (Ford, 2016). Esos han sido creados por empresas, organizaciones públicas (que aportan datos de encuestas, actas, informes, documentos oficiales, leyes) y por las personas.
La ponencia presenta los primeros resultados del estudio exploratorio de big data de la serie de encuestas de hogares de Chile (Encuesta CASEN) del período 1990-2017 mediante la datificación de los libros de códigos de las encuestas. *Datificar un fenómeno es plasmarlo en un formato cuantificado para que pueda ser tabulado y analizado* (Mayer-Schonberger y Cukier, 2015). El trabajo muestra el análisis temporal de las preguntas CASEN y los cambios que fueron experimentando. Además, describe la huella digital que deja el Estado en las preguntas de la encuesta: preguntas que nacen, mueren y permanecen. La dinámica expresa los intereses y preocupaciones del Estado en esas materias. La datificación otorga un nuevo valor a los datos originales CASEN, facilita el acceso, permite desarrollar capacidades de uso de los datos. Y puede contribuir a las políticas públicas al momento de evaluarlas.

Palabras clave: big data, datificación, visualización, huella digital, uso de datos públicos.

Abstract

The huge amount of data (text, images, sound, numbers, videos) in cyberspace doubles every three years following its own Moore's Law (Ford, 2016). Those have been created by companies, public organizations (providing data from surveys, minutes, reports, official documents, laws) and by individuals.
The paper presents the first results of the exploratory big data study of Chile's household survey series (CASEN) from 1990-2017 by datifying survey codebooks. To datify a phenomenon is to capture it in a quantified format so that it can be tabulated and analyzed (Mayer-Schonberger and Cukier, 2015). The work shows the temporal analysis of the CASEN questions and the changes they underwent. In addition, it describes the fingerprint left by the State in the survey questions: questions that are born, die and remain. The dynamics express the interests and concerns of the State in these matters. The datification gives a new value to the original CASEN data, facilitates access, allows the development of data use capabilities. And it can contribute to public policies when evaluating them.

Keywords: big data, datification, visualization, digital footprint, use of public data.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Introducción

El vertiginoso desarrollo de las tecnologías digitales con consecuencias en la vida social, política, económica de los países generó un cambio socio tecnológico a gran escala. Entre sus efectos reconocemos la ingente cantidad de datos publicados en la web y, aquellos datos que provienen de la Internet de las cosas (IOT), desde sensores instalados en edificios, cultivos agrícolas, fábricas, automóviles, máquinas, artefactos electrónicos domésticos, etc. Todos corresponden a la noción de big data (datos masivos, *macrodatos* según Mayer-Shönberger y Cukier, 2015) que experimentó un crecimiento exponencial en el último tiempo. Recientemente, en el ámbito de las Ciencias Sociales, y en contexto de pandemia, la irrupción de big data quedó expresado en el creciente número de iniciativas colaborativas en la red. En los años 2020, 2021 fuimos testigos de cursos, seminarios, capacitaciones y mesas de trabajo de congresos científicos realizados. El big data (datos masivos) en este trabajo comprenden no sólo los datos abiertos de la web sino también los que están disponibles en distintas fuentes públicas y privadas. En definitiva, se trata del conjunto de datos que proveen fuentes nuevas de datos (Becker, 2018) a los que prestar atención. El tema no es nuevo en los estudios sociales. Bolstanki y Chiapello (1999) conceptualizaron *el nuevo espíritu del capitalismo* con metodología aplicada a las fuentes de datos del conjunto de libros de gestión empresarial y la creación de un corpus de datos cuyo análisis fue asistido por el programa informático Prospero@. Los autores describieron la utilidad del procedimiento para examinar las categorías relacionadas cada uno de los *mundos* analíticos establecidos (Bolstanki y Chiapello ([1999], 2010)[1].

En Chile existe un abundante corpus de información sobre materias públicas (datos públicos) amparado en un conjunto de disposiciones legales y regulaciones. La Ley de Transparencia (20.285) del año 2008[2], la Ley de datos públicos de 2010. La primera regula el acceso de los chilenos/as a la información pública, esto es, permite acceder a la información de los organismos de la administración del Estado. En segundo lugar, los datos abiertos o datos de formato abierto (*open data*) corresponden a *aquellos donde las especificaciones del software están disponibles para cualquier persona de forma gratuita, así cualquiera puede usar las especificaciones en su propio software sin ninguna limitación en su reutilización que fuere*

[1] Ver Apéndices, pp. 659-680 en Bolstanki y Chiapello ([1999], 2010).
[2] Texto de ley disponible en https://www.bcn.cl/leychile/navegar?idNorma=276363&idParte=

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

impuesta por derechos de propiedad intelectual (*Open Data Handbook*)[3]. Un ejemplo en materia de datos públicos es el repositorio de datos abiertos del Estado chileno *DatosGob*[4]. El sitio web menciona que *las instituciones publican sus datos de manera sencilla y transparente, en formatos abiertos con la lógica de un catálogo centralizado y ordenado, de rápida búsqueda y fácil uso.* Asimismo, desde 1999 están las normas que atienden la seguridad y privacidad de los datos de las personas, Ley 19.628 de Protección de la Vida Privada[5] que también se aplica al tratamiento que otorgan los organismos públicos a esos datos.

Las fuentes de datos públicos poseen distintos objetivos. Por ejemplo, la Comisión Nacional de Energía[6] (CNE), en cumplimiento de su papel de regulador estatal en la materia, provee semanalmente en el sitio web *Bencina en línea* los precios de combustibles de los servicentros del país. El Ministerio de Educación mantiene el Sistema de Información de Educación Superior (SIES) que pone a disposición los datos emanados por el sistema de educación terciaria, entre los datos están, la admisión a la educación universitaria, los resultados del sistema (características de egresados, datos de empleabilidad y más). En el sistema de políticas sociales nacionales una fuente de información es la Encuesta de Caracterización Socioeconómica Nacional (en adelante CASEN). En general, podemos señalar que las políticas públicas constituyen importantes fuentes de datos.

En el terreno de la producción de información para la toma de decisiones el Estado no es una entidad rápida a la hora de recolectar datos, procesarlos y publicar los resultados. En el último tiempo cuando se habla de los problemas del desarrollo de Chile, es decir, de los resultados obtenidos por la estrategia (modelo) de desarrollo vigente de las últimas cuatro décadas, se hace referencia al papel del Estado, su eficacia y eficiencia. Es habitual escuchar que el Estado es lento, que la gestión de las instituciones públicas va quedando atrás respecto a los problemas de la ciudadanía. Sobre la eficacia estatal de producir y distribuir datos que alimenten las

[3] La noción está disponible en https://datos.gob.cl/about. Consultado el 19 de mayo de 2022.
[4] En enlace https://datos.gob.cl/about. Consultado el 19 de mayo de 2022.
[5] Disponible en https://www.bcn.cl/leychile/navegar?idNorma=141599&idParte=864270
[6] En el marco el papel regulador del Estado la Comisión Nacional de Energía aplica el Mecanismo de Estabilización de Precios de Combustibles (MEPCO), creado por la Ley Nº20.765 (09.07.2014) con el fin de establecer un mecanismo que estabiliza los precios internos de venta de combustibles (Ley N°18.502). El mecanismo opera a través de incrementos y rebajas a impuestos específicos a los combustibles establecidos por la ley. Ellos se modifican sumando al componente base de la ley un componente variable (positivo o negativo) determinado para cada combustible: gasolina automotriz, petróleo diésel, gas natural comprimido y gas licuado de petróleo.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

decisiones que resuelven los problemas sociales, surgen voces (principalmente del mundo político y académico) que plantean la necesidad de modernizar el Estado para enfrentar los nuevos desafíos del desarrollo nacional.

Los datos públicos utilizados en dar forma a las estadísticas sociales y económicas del país y el acceso han apoyado predominantemente al trabajo de los profesionales de la administración pública y las estadísticas sociales y económicas, y la labor especializada de investigación. Esa mirada insinúa que el despliegue del esfuerzo público hacia los usos descritos parece ser insuficiente y, al mismo tiempo el acceso a los datos almacenados en repositorios abre un conjunto de posibilidades en el marco del desarrollo del país.

La ponencia plantea que el acceso y uso a datos públicos de las personas interesadas y los profesionales insertos en ambientes locales de planificación que enfrentan limitaciones en el uso de datos para el estudio de problemas colectivos y la toma de decisiones puede mejorar y/o ser complementado con procesos de datificación. A esta situación me enfrenté en la docencia de Planificación Social cuando quise utilizar los datos de las encuestas CASEN con estudiantes universitarios. El beneficio de disponer de datos disminuye significativamente por la falta de estandarización. En el caso de los estudiantes universitarios los datos debieron ser reprocesados para ser utilizados. En la elaboración de mi tesis de doctorado observé el valor que tenían las preguntas de las encuestas (2000 - 2013), en el proceso de estudio pude advertir la escasa velocidad de las instituciones del Estado para capturar fenómenos de rápido desarrollo. Por ejemplo, la evolución de la penetración de dispositivos tecnológicos e Internet en el hogar nacional y los que portan los individuos (Benítez, 2020). En ese momento, el tiempo dedicado a la tarea fue excesivo para aumentar el ámbito de búsqueda

La perspectiva de este trabajo es que aún falta progresar en el aspecto central de mejorar el acceso y uso de datos abiertos. La política requiere ampliar la base de usuarios e ir más allá de aquellos que ya los emplean. La idea se funda en aumentar *las capacidades humanas, en lugar reemplazar a la gente con métodos computacionales en la toma de decisión* (Anscombe, 1973). Lo anterior va unido al impulso de la transformación digital (datos + nubes nuevas herramientas de productividad y de colaboración), que fue acelerada durante la pandemia, y contribuyó a darnos cuenta de la necesidad de optimizar el trabajo con big data, esto es, hacer más con los datos disponibles.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Dado lo anterior, poner a disposición en repositorios los datos de los Cuestionarios, Libros de Códigos, Manuales de Usuario de la encuesta CASEN no es suficiente para darles uso. Los datos requieren de *datificación* para ser analizados y complementados con nueva información. Cuando el dato es *datificado* admite ser registrado, analizado, reorganizado. No existe un término adecuado para nombrar du transformación desde el registro. *Datificar un fenómeno es plasmarlo en un formato cuantificado para que pueda ser tabulado y analizado* (Mayer-Schonberger y Cukier, 2015, p.100). El ciclo vital del dato (huella digital) es una variable proxy del comportamiento del Estado en estas materias. Una vez datificados los datos se complementan con la visualización, Few (2004) señala que la visualización constituye el despliegue gráfico de una información abstracta para los propósitos de dar sentido al análisis de datos y su comunicación.

Además, los datos experimentan transformaciones temporales y revelan que nacen, mueren y permanecen dejando una huella digital (Hilbert, 2013). En general, los datos suelen presentar problemas como consecuencia de la dinámica que tienen. Dichos problemas afectan su calidad y requieren ser resueltos. Ello genera un valor adicional a la datificación en el caso de las encuestas, porque permiten seguir el ciclo de vida de las preguntas y observar la huella digital de la entidad que las produce.

El trabajo desarrolla el análisis temporal de las preguntas CASEN y los cambios que fueron experimentando mediante el uso de la datificación de los Libros de Códigos de las encuestas. La ponencia presenta los primeros resultados del estudio exploratorio de big data de la encuesta de hogares de Chile, 1990-2017, 2020, e ilustra cómo a partir de la *datificación* de los datos CASEN se obtienen nuevas categorías de variables. *La datificación* otorga nuevo valor a los datos originales. La datificación y la visualización de los datos CASEN procuran mostrar el potencial que tienen para facilitar el acceso y desarrollar capacidades en el uso de datos. Por último, la ponencia describe la huella digital que deja el Estado en las preguntas de esta encuesta (cuáles nacen, mueren y cuáles permanecen). Su dinámica expresa los intereses y preocupaciones del Estado en |estas materias. Lo anterior, sugiere la posibilidad de generar nuevas soluciones a necesidades colectivas: innovar.

Las páginas siguientes presentan la metodología utilizada, los principales resultados de la datificación de datos CASEN. La discusión subraya la huella digital estatal en la encuesta y los

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

beneficios de datificar y visualizar datos para un número más amplio y diverso de usuarios en la línea de ampliar sus capacidades. Por último, se describe la conclusión principal.

Materiales y métodos

El big data de trabajo está compuesto por las bases de datos, cuestionarios, libros de códigos, Manual del investigador, Manual de trabajo de campo de la encuesta CASEN[7]. Ocasionalmente estos datos masivos incluyen notas técnicas de los ajustes metodológicos y las bases de datos complementarias de ingresos. La Encuesta CASEN es el instrumento que recoge *información para caracterizar la situación de los hogares y de la población, especialmente de aquella en situación de pobreza y de grupos definidos como prioritarios por la política social en aspectos demográficos, educación, salud, vivienda e ingresos. Estimar la cobertura, focalización y distribución del gasto fiscal de los principales programas sociales del alcance nacional, para evaluar su impacto en el hogar, en términos del ingreso adicional que les significa y el efecto en la distribución del mismo* (Metodología CASEN 2009, Ministerio de Desarrollo Social, 2010). En consecuencia, en términos generales, la encuesta se compuesta de los Módulos de Educación, Salud, Vivienda, Ocupación (Empleo, Trabajo), Ingresos y datos demográficos de los hogares.

Los resultados de cada encuesta[8] constituyen un conjunto de documentos técnicos y base de datos publicados por el Ministerio de Desarrollo Social y Familia (MDSF) en el sitio web *Observatorio Social*[9]. Allí se almacenan las Estadísticas, Metodología, Cuestionarios, Bases de datos de las encuestas del periodo 1990 – 2000. En 2003 se agregó la categoría Resultados. En 2013 la encuesta modificó de manera importante la medición de la pobreza lo cual sumó documentos con complementos metodológicos. Los archivos del período se disponen en formatos Excel (Estadísticas), PDF y archivos comprimidos en formato Zip. Y las bases de datos en STATA y SPSS. Los resultados de las encuestas generalmente se presentan en las secciones de Estadísticas (planillas de datos Excel) y de Resultados (documentos PDF con tablas de datos y gráficos en presentaciones PPT). Éstos exhiben los tópicos de módulos y los principales indicadores sociales, por ejemplo, la distribución de ingresos (deciles, quintiles de

[7] Disponibles en el sitio web Observatorio Social del Ministerio de Desarrollo Social y Familia (MDSF), Enlace http://observatorio.ministeriodesarrollosocial.gob.cl/
[8] Encuesta CASEN efectuadas, años 1990,1992, 1994, 1996, 1998, 2000, 2003, 2006,2009, 2011, 2013, 2015, 2017, 2020 (CASEN en pandemia). Los resultados se publican un año después.
[9] http://observatorio.ministeriodesarrollosocial.gob.cl/

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

ingresos; Coeficiente de Gini) y nivel de pobreza. La Metodología se enseña en el Manual de Trabajo de Campo y del investigador. Y el facsímil del Cuestionario presenta las preguntas de la encuesta. Las Bases de datos contienen el Libro de Códigos de bases de datos (principal y complementarios cuando los hay), más las bases de datos de la encuesta respectiva.

El componente básico de la data son las variables que agrupan las preguntas de los módulos temáticos en los cuestionarios de encuestas y los indicadores.

Los datos usados en este trabajo corresponden a los Libros de Códigos de la encuesta, aplicada bianual o trianualmente, en el período 1990-2017. La última encuesta data del año 2020 (CASEN en pandemia) implementada con modificaciones respecto a las versiones anteriores: menor cantidad de preguntas, una muestra menor y nueva modalidad aplicada (encuesta telefónica). Dicha consideración fundamenta el énfasis del trabajo en el período 1990-2017. El trabajo empleó las bases de datos en formato SPSS. Y, por razones prácticas se decidió procesar las bases de datos en Excel y no usar el paquete estadístico R. Cabe subrayar que la tarea de datificación de la serie CASEN aún está en desarrollo lo que puede provocar ajustes futuros en los resultados. Por lo cual los resultados presentados en esta ponencia podrían variar discretamente próximamente.

¿Cómo se datificó en este trabajo? Se datificaron los libros de códigos de la serie en una base de datos. A la variable CASEN se agregaron las categorías de tiempo (año), tipo, módulo, a quien se dirige la pregunta, nombre interno y ámbito. La datificación de los datos CASEN originales y las nuevas categorías otorgaron valor adicional a los datos de la encuesta que facilita la búsqueda de variables, permite generar series de tiempo; rehacer informes de resultados oficiales e identificar variables asociadas con informes oficiales nacionales, internacionales y rehacer artículos de investigación, entre otros. Estas opciones posibilitan ampliar el universo de potenciales usuarios de datos.

El procedimiento de la coincidencia (*match*) de las preguntas que presenta la ponencia en la sección de resultados de la huella digital refiere a un proceso manual que usó la descripción como guía para establecer las coincidencias. Cuando hubo dudas se analizaron las opciones de las preguntas cuya consulta fue corroborada con los datos de las encuestas. Cabe señalar, que hubo catorce preguntas sin descripción. Finalmente, los resultados obtenidos con el

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

procedimiento fueron contrastados con los que entrega el Observatorio Social del ministerio social (MDSF).

La experiencia del usuario

Los interesados en obtener información socioeconómica vigente (nacional, regional) de la encuesta CASEN con frecuencia utilizarán aquella que está disponible en los Resultados de la última versión publicados por el MDSF. Cuando el objetivo del usuario es acceder y usar más de una encuesta deberá dar lectura a las preguntas del Cuestionario y, probablemente, el Libro de Códigos o Manual de usuario[10]. Qué ocurre con los interesados en los datos CASEN que no están publicados en los Resultados ni las Estadísticas cuyos tópicos pueden encontrarse en las Bases de Datos de la encuesta. Y con aquellos que desean combinar datos de dos o más encuestas. Por último, qué ocurre con otro tipo de combinaciones de datos y/o con la generación de otros indicadores a partir de los datos originales.

La labor de datificación identificó una serie de problemas relacionados con las preguntas CASEN de la serie. Los datos presentan algunos errores cuya única forma de resolverlos es la lectura de los libros de códigos de las encuestas que interesan al usuario. Se constataron problemas que se refieren al nombre, descripción y codificación. Otros se asocian a las variables e indicadores que no permiten usar (directamente) la información de la encuesta y por ende obtener mayor provecho de los datos. En algunas encuestas se aprecian diferencias entre la descripción de los libros de códigos y la de los cuestionarios. El conjunto de problemas descritos[11] no permite que los datos puedan ser utilizados de manera simple por el usuario. Por lo cual datificar las encuestas facilita bastante el uso de los datos y agrega valor a los datos originales. Respecto al análisis de datos cabe decir que los estudios indican que un porcentaje importante del proceso, más del 50% del tiempo, se ocupa en tareas de adecuación y/o preparación de los datos para su examen.

[10] Si el usuario requiere profundizar más, además de los anteriores, revisará el Manual del Investigador.

[11] Una breve descripción detalla algunos problemas específicos que exhiben los datos para ilustrar la situación. En el *nombre de la pregunta*: un nombre igual de la pregunta no representa lo mismo en encuestas de años diferentes. En otras ocasiones sucede lo contrario. En la *descripción de la pregunta*: distintas variables pueden describir a una misma pregunta. La diferencia ocurre en la descripción de la variable de la pregunta y el código. El ejemplo: Pregunta CASEN 1998, Módulo de Vivienda: variable v44 y v38. Pregunta ¿En qué año recibió el subsidio o vivienda? Ejemplo de Indicador CASEN 2013, Módulo Ingresos. El indicador denominado Independientes *principal – Efectivo*, se expresa de manera idéntica en cuatro descripciones de distintos indicadores y0701ch, y0703, y0702, y0701c. Los problemas entre descripción y códigos como los descritos son muchos en la serie CASEN.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

En suma, los cuestionarios CASEN no constituyen un conjunto homogéneo de preguntas. Los inconvenientes mencionados afectan la precisión de la información y presentan dificultades al usuario cuando su objetivo es acceder a más de una encuesta. Frente a las dificultades, la resolución vía la lectura de libros de códigos de las encuestas que interesan al usuario con la inversión de trabajo y tiempo que implica desalienta la acción la mayoría de las veces.

Resultados

La introducción planteó que datificar un fenómeno consiste en expresarlo *en un formato cuantificado* para que pueda ser tabulado y analizado. El primer resultado de la sección entrega el formato cuantificado de la serie de encuestas CASEN. El segundo muestra los procesos con los datos que permiten la reproducción de informes oficiales, el análisis de datos y la visualización. Finalmente, se presenta el resultado que ha posibilitado la datificación, el análisis del ciclo de vida de las preguntas de las encuestas que se ha denominado huella digital.

El formato cuantificado

El primer elemento de resultado, el formato cuantificado, explica las categorías usadas en la datificación para el uso más eficiente (sencillo y rápido) de los datos CASEN. En cada encuesta las variables se identificaron con el nombre de la variable y descripción ambas son propias de CASEN. A estos identificadores se asociaron las categorías: Año, Tipo, Módulo, A quién aplica, Nombre interno y Ámbito. El *Año* describe el año de ejecución de la encuesta. *Tipo* indica si el dato de la encuesta es un identificador de ella (IE), representa una pregunta (P) o un indicador (I) que se construye en base a las preguntas de la encuesta. *Módulo* refiere al tema que agrupa las preguntas de la encuesta. *A quien aplica* describe el sujeto al que se aplica la pregunta, las preguntas y módulos CASEN especifican la edad y/o el rol en el hogar del encuestado. Estas variables se extrajeron de los libros de códigos de las encuestas. El proceso de datificación continuó con la incorporación de otras dos variables: Nombre (código) interno y Ámbito. La primera es un nombre que asocia a todas las variables equivalentes en la serie de encuestas. Ámbito es una categoría que se adoptó del análisis de la forma como el gobierno clasifica la información de esta encuesta. Su fuente es el Observatorio Social[12] del MDSF.

[12] Disponible en http://observatorio.ministeriodesarrollosocial.gob.cl/

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Tabla 1. Vista parcial del formato cuantificado Modulo Educación.

Variable	Descripción	Año	Tipo	Módulo	A quién aplica	Nombre interno	Ámbito
e1	Sabe leer y escribir	1990	P	Educación	Personas de 15 y más años	e1	Caracterización de la población (CP)
e1	Sabe leer y escribir	1992	P	Educación	15 ó más	e1	CP
e1	Sabe leer y escribir	1994	P	Educación	15 ó más	e1	CP
e1	Sabe leer y escribir	1996	P	Educación	15 ó más	e1	CP
e1	¿Sabe leer y escribir?	1998	P	Educación	15 o más	e1	CP
e1	Sabe leer y escribir	2000	P	Educación	15 o más	e1	CP
E1	¿Sabe leer y escribir?	2003	P	Educación	15 años o más	e1	CP
E1	¿Sabe leer y escribir?	2006	P	Educación	15 años o más	e1	CP
E1	¿Sabe leer y escribir? (15 años y más)	2009	P	Educación	15 años o más	e1	CP
e1	e1. ¿sabe leer y escribir?	2011	P	Educación	Personas de 15 años o más	e1	CP
e1	e1. ¿Sabe leer y escribir?	2013	P	Educación	Personas de 15 años o más	e1	CP
e1	e1. ¿Sabe leer y escribir?	2015	P	Educación	Personas de 15 años o más	e1	CP
e1	e1. ¿Sabe leer y escribir?	2017	P	Educación	Personas de 15 años o más	e1	CP

Fuente: Elaboración propia.

El formato cuantificado presentado permite hacer la búsqueda de variables por cualquier combinación de sus categorías (campos). Por ejemplo, las variables de una encuesta o todas las variables del Módulo Educación y las de otros módulos. La Tabla 1 muestra el resultado de todas las variables cuyo código interno es *e1*.

El primer análisis que surgió de la datificación es la vista general de la cantidad de variables según año, tipo y módulo de pertenencia de la Tabla 2. La cantidad total de preguntas fue creciendo asimismo las variables del período. Considérese que la tarea de datificación CASEN del período está en pleno desarrollo, por lo tanto, podría haber algunas variaciones discretas en el número total de los elementos de la tabla en el futuro cercano.

Tabla 2. Encuestas CASEN (1990-2017) por Módulo y categorías de datificación.

		Tipo			Módulos CASEN			
Variables	**Año**	**Identificación**	**Preguntas**	**Indicadores**	**Educación**	**Salud**	**Ocupación[1]**	**Vivienda[2]**
158	1990	16	106	36	10	42	12	31
194	1992	18	139	37	30	55	18	36
216	1994	19	163	33	27	56	22	58
193	1996	16	138	39	22	49	27	39
201	1998	17	144	40	19	53	22	45
285	2000	19	223	42	35	75	34	52
306	2003	15	249	42	35	71	0	44
348	2006	12	276	60	36	56	29	57
357	2009	17	269	71	37	65	0	51
394	2011	19	279	96	51	82	0	52

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

600	2013	11	517	72	57	87	0	82
776	2015	21	672	83	57	98	0	66
807	2017	20	697	90	60	91	0	0

Fuente: Elaboración propia.
Notas:
(1) Módulo Ocupación también asume el nombre de Empleo, Trabajo en las encuestas. Ocupación incluyen los ingresos del hogar. En las encuestas con el módulo Empleo, Trabajo los ingresos generalmente van en módulos distintos. El valor 0 en el 2003, 2009-2017 indica que ese año Ocupación cambió de nombre. Por ejemplo, 2003 el módulo fue Empleo e Ingresos del trabajo.
(2) Modulo Vivienda de 2017 tiene el nombre de Vivienda y Entorno.

Tabulación y análisis de los datos

El objetivo de la datificación es la tabulación y el mejor análisis de los datos. A continuación, se presenta la reproducción de documentos oficiale. Ésta es posible hacer por medio de plantillas que permiten al usuario verificar la información, hacer filtros y rehacer el resultado completo si lo desea. La plantilla de admite el manejo simple.

No obstante, el objetivo de datificar no se agota en la reproducibilidad de documentos oficiales porque además permite disponer de los datos base para corroborar y/o profundizar el análisis. Esas tareas se pueden apoyar en las herramientas que tienen las aplicaciones de procesamiento de datos de uso habitual. En este caso es Excel, la herramienta cuenta con *tablas dinámicas*, filtros, funciones matemáticas y estadísticas de fácil manejo para apoyar las tareas.
La Tabla 3 reproduce el resultado del Cuadro 1 del Módulo de Educación[13] de la población mayor de quince años según sexo de la Región I que sabe leer y escribir y la población analfabeta por zona urbana y rural de CASEN 1990[14]. La reproducción se hizo leyendo las variables de *sexo (sexo*); *edad (edad)*; *z (zona), r (región); expr (factor de expansión regional), e1(sabe leer y escribir)*. La selección estuvo apoyada en el proceso de datificación anteriormente descrito. Los valores del cuadro se obtuvieron mediante operaciones que sumaron *expr* filtradas por zona, región, sexo, edad, sabe leer y escribir (*z, r, sexo, edad, e1* respectivamente*)*. La reproducción obtenida es 100% fidedigna.

Tabla 3. Reproducción. Población mayor de 15 años que sabe y no sabe leer y escribir. Región I, CASEN 1990.

Cuadro 1 - Reconstruido (variables usadas: edad, *r, z, sexo, expr*)

[13] Los cuadros de resultados del Módulo Educación CASEN 1990 suman veintisiete planillas en total.
[14] Disponible en Enlace http://observatorio.ministeriodesarrollosocial.gob.cl/encuesta-casen-1990

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

POBLACIÓN DE 15 AÑOS Y MÁS SEGÚN ANALFABETISMO REGIÓN, ZONA Y SEXO

Región I

Zona	Sexo	Sabe leer y escribir				Total	
		Sí		No			
		Número	%	Número	%	Número	%
Urbano	Hombre	99.941	98,5	1.518	1,5	101.459	100,0
	Mujer	110.131	97,7	2.580	2,3	112.711	100,0
	Total	210.072	98,1	4.098	1,9	214.170	100,0
Rural	Hombre	6.793	95,8	299	4,2	7.092	100,0
	Mujer	5.690	93,2	415	6,8	6.105	100,0
	Total	12.483	94,6	714	5,4	13.197	100,0
Total	Hombre	106.734	98,3	1.817	1,7	108.551	100,0
	Mujer	115.821	97,5	2.995	2,5	118.816	100,0
	Total	222.555	97,9	4.812	2,1	227.367	100,0

Fuente: Elaboración propia.
Nota:
El Cuadro 1 mantuvo el formato y distribución de datos del cuadro original.

A partir del cuadro de resultados reelaborado del módulo Educación se pueden hacer varias combinaciones de variables. La Tabla 4 ilustra muy bien esa característica. La tabla de datos se obtuvo con la *tabla dinámica,* una herramienta que admite trabajar con los datos de manera sencilla, como se mencionó la datificación facilita el uso de datos pues permite hacer la clasificación con los filtros de Excel y no con módulos estadísticos.

Tabla 4. Región I. Población mayor de 15 años que sabe y no sabe leer y escribir por zona (urbana, rural) y Quintil de Ingreso Regional, CASEN 1990.

Suma de expr	**Etiquetas de columna**						
	Sí		**Total Sí**	**No**		**Total No**	**Total general**
Etiquetas de fila	**Urbano**	**Rural**		**Urbano**	**Rural**		
I	41526	2698	44224	1629	108	1737	45961
II	47058	2748	49806	415	238	653	50459
III	39993	2596	42589	1032	138	1170	43759
IV	41958	2462	44420	415	161	576	44996
V	37666	1956	39622	607	69	676	40298
Total general	**208201**	**12460**	**220661**	**4098**	**714**	**4812**	**225473**

Fuente: Elaboración propia.
Nota: La tabla mantuvo el formato de elaboración de la tabla dinámica.

La Tabla 4 muestra el análisis de las variables: zona (*z*) urbana, rural; quintiles de ingreso autónomo regional (*quatr*), Región I e ilustra una de las combinaciones posibles de las variables de análisis, en este caso agregó la distribución por quintiles de ingreso autónomo.

Completamos la descripción del potencial de análisis de la datificación con la visualización de los datos. La visualización es un recurso gráfico que permite fijar rápidamente algunas

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

características del fenómeno analizado y/o mostrar diferencias relevantes (que saltan a la vista) del fenómeno cuando se desean comparar dos o más categorías de análisis. La visualización puede realizarse porque se dispone de todos los datos de la encuesta datificados. De este modo puede hacerse una rápida selección de variables que interesa trabajar, luego llevar los respectivos datos de la visualización. El Diagrama 1 muestra la cantidad de personas por el tipo de establecimiento educativo (variable *e9*) según zona y sexo. Se observa/n:

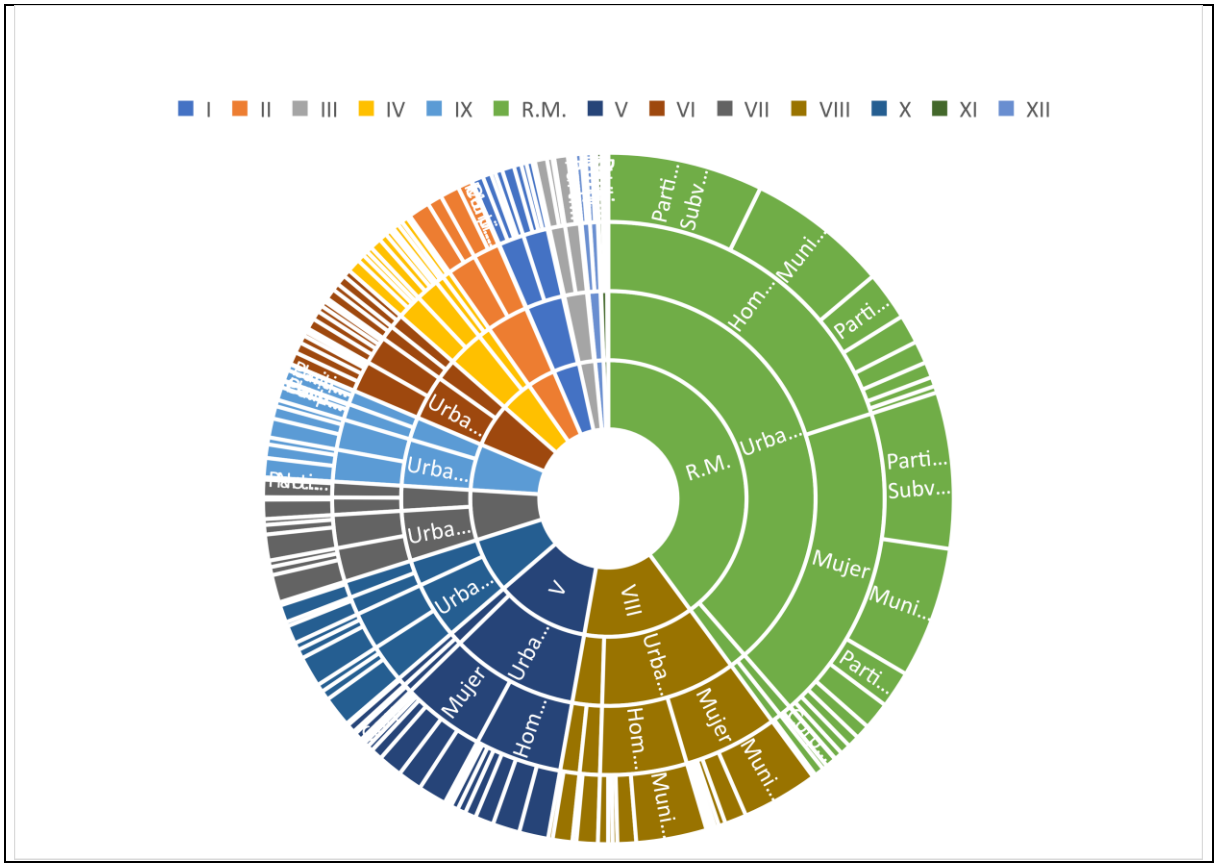
a) La magnitud que tiene la Región Metropolitana (R.M.).

b) Las tres regiones con la mayor concentración de población en zonas urbanas en 1990 eran (que conforman las áreas metropolitanas más importantes de Chile) la Región Metropolitana (Gran Santiago); V Región (Gran Valparaíso) y VIII Región (Gran Concepción).

c) La educación Municipal (pública) tiene un lugar destacado en el sistema educativo del país en varones y mujeres de zonas urbanas.

d) En la R.M. y V Región se advierte una proporción de Educación Particular[15](Privada).

Diagrama 1. Regiones de Chile. *Establecimiento educacional* por zona y sexo de la población, CASEN 1990.



Fuente: Elaboración propia.

[15] CASEN 1990 la denominó *Particular No Subvencionada*, esto es, educación impartida por entidades educativas privadas.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Huella digital. La vida de las preguntas y el Estado retratado

En el ciberespacio no es sólo el lugar donde están almacenados los datos masivos, también denota la manera como "los productores de datos" los presentan. Por lo tanto, los datos acumulados en repositorios en el espacio digital constituyen la huella digital de las entidades que los generan, indican la interacción que estos efectúan en el ciberespacio revelando sus intereses, preferencias, preocupaciones, prioridades, etc.

Entonces, el registro de las preguntas de las encuestas CASEN en el ciberespacio son una muestra de las preocupaciones e intereses del Estado chileno por un conjunto de problemas socioeconómicos de la población. La huella digital expresa el ciclo vital de las preguntas, ese rastro digital es una variable proxy del comportamiento del Estado en estas materias, de manera que el comportamiento del Estado queda retratado en estos datos.

El análisis se circunscribe al Módulo de Educación, la Tabla 5 muestra las veces que una pregunta apareció en la serie de encuestas.

Tabla 5. Cantidad de veces que las preguntas de Educación han estado presentes en la encuesta.

Nro. Veces	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	
Nro. Preguntas	28	10	22	9	5	3	4	7	8	0	2	6	3	Total 107

Fuente: Elaboración propia.

Se han formulado 107 preguntas de Educación en la serie de encuestas CASEN (1990-2017). Solamente, un grupo pequeño de ellas se han mantenido durante el período, en cambio muchas preguntas fueron realizadas pocas veces. Esto indica que existe una dinámica importante en el ciclo de vida de las preguntas, por lo tanto, hubo cambios en la perspectiva del Estado frente a la materia.

Tabla 6. Cantidad de preguntas del Módulo de Educación en el período 1990-2017.

Año	1990	1992	1994	1996	1998	2000	2003	2006	2009	2011	2013	2015	2017	Total
Preguntas/ año	10	30	27	22	19	35	35	36	37	51	57	57	60	476

Fuente: Elaboración propia.

La Tabla 6 presenta la cantidad de preguntas del Módulo de Educación de la serie CASEN. Las diez preguntas originales se convirtieron en sesenta en la encuesta 2017. En su recorrido se aprecian tres subperíodos. De 1990 a 1998, el número de preguntas fluctuó entre 19-30. En el

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

segundo subperíodo 2000-2009 las preguntas se estabilizaron en 35-37. En el subperíodo de 2011-2017 la cantidad de preguntas se ubicó en el rango 51-60.

La Tabla 7 describe el detalle del subperíodo de 1990-1998, los datos muestran el ciclo de vida de todas las preguntas realizadas. Se observa que hubo preguntas que estuvieron presentes en todo el subperíodo, mientras otras nacen y no se mantuvieron. Finalmente, hubo otras que surgieron al final del subperíodo, en 1998.

Tabla 7. Preguntas CASEN del Módulo de Educación, subperíodo 1990-1998.

Variable	Año					
	1990	1992	1994	1996	1998	Total
e1	1	1	1	1	1	5
e2	1	1	1	1	1	5
e3	1	1	1	1	1	5
e4	1	1	1	1	1	5
e6	1	1	1	1	1	5
e15		1	1	1	1	4
e16		1	1	1	1	4
e34		1	1	1	1	4
e43		1	1	1	1	4
e44		1	1	1	1	4
e46		1	1	1	1	4
e45		1	1		1	3
e20			1	1	1	3
e10	1	1	1			3
e47		1	1	1		3
e67		1	1	1		3
e69		1	1	1		3
e71		1	1	1		3
e72		1	1	1		3
e74		1	1	1		3
e70		1	1	1		3
e73		1	1	1		3
e8	1	1	1			3
e7	1	1	1			3
e35		1	1			2
e14		1	1			2
e31		1				1
e51					1	1
e66					1	1
e48					1	1
e49					1	1
e50					1	1
e42					1	1

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

e38		1				1
e22				1		1
e39		1				1
e18			1			1
e68				1		1
e37		1				1
e5	1					1
e9	1					1
e36		1				1
Total	**10**	**30**	**27**	**22**	**19**	**108**

Fuente: Elaboración propia.

Por lo tanto, se mantienen las preguntas relacionadas con las características de educación de la población: sabe leer y escribir, cobertura, nivel de estudios y tipo de enseñanza. En el año 1992, la encuesta incorporó elementos que permiten normalizar la dependencia del establecimiento educativo y las becas (ayudas económicas); que al final del subperíodo aparecen incorporadas según Educación Básica y Secundaria. Además, surgió un conjunto de preguntas que buscó determinar los montos de las *becas de matrícula* y de *crédito fiscal* (en la educación terciaria). Finalmente, existe un conjunto de preguntas que aparecen por una vez y corresponden a preguntas que no se mantuvieron en la encuesta.

El comportamiento descripto concuerda con los ámbitos que el Estado chileno entrega periódicamente en los resultados de la encuesta a través del *Observatorio Social* del MDSF. Los ámbitos son: *Caracterización de la población, Cobertura de beneficios educativos, Educación General Básica, Educación general Media, Educación Parvularia, Educación Superior, Financiamiento a la Educación Superior*.

CONCLUSIÓN

La ponencia plantea el valor de la *datificación* de los datos públicos para incrementar su disponibilidad efectiva en la sociedad, de modo que más personas interesadas accedan a los datos en forma simple, tal que los datos se transformen *en una plataforma donde los demás puedan organizarse y crear nuevo valor* (Tapscott y Williams, 2012, p. 478).

La datificación (en tanto formato cuantificable) de las encuestas CASEN consistió en una tabla que integra la información básica de las preguntas (nombre, descripción, año), la información complementaria (módulo y a quien aplica) más un código interno para integrar las palabras similares y un campo para describir el ámbito de uso de las preguntas. La datificación queda incluida en una herramienta de productividad personal que permite que, la selección de las

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

preguntas, la lectura de los datos respectivos y el análisis sean realizados de manera muy simple en una computadora personal.

Los datos que publica el Estado vía datificación pueden aumentar significativamente su potencial de apoyo a las políticas públicas, no sólo porque permiten un mayor uso, sino porque también es posible rastrear la huella que deja en el proceso; al igual que la gran mayoría de las organizaciones (sociales, productivas) y personas que colocan contenidos en el ciberespacio. En el análisis de las redes sociales digitales se ha constatado que las personas rehacen (copian) sus redes sociales físicas (Benítez, 2020; Winocur, 2010; Castells, 2005). Siempre las personas y las organizaciones dejan huellas de lo que hacen, Vallejo (2021) [16]señala que los antiguos copistas de libros dejaron su huella en los errores, omisiones y palabras intercambiadas de la copia.

La ponencia muestra el ciclo de vida de las preguntas de las encuestas CASEN forma parte de la huella digital del Estado chileno, las preguntas CASEN representan las necesidades de las instituciones del estado y, presentan una mirada indirecta del Estado y la Sociedad Civil. Además, la dinámica de cambio que presentan las preguntas (las que nacen y mueren) y las que se mantienen en el período, muestra los cambios en las prioridades del Estado y, también los problemas que fue teniendo al abordar algunos temas socioeconómicos.

En resumen, cuando el Estado publica sus datos en el ciberespacio deja una huella en su forma y contenido. La huella registrada en el ciberespacio permite obtener una visión indirecta del Estado actual y, por lo tanto, contribuye con un complemento importante al momento de evaluar las políticas públicas. El análisis de las preguntas del Módulo de Educación de la serie CASEN de la ponencia muestra que el Estado chileno enfrentó tres períodos distintos (1990-1998; 2000-2009 y 2011-2017) en la materia, constituyen resultados que contribuyen o pueden contribuir a dicho proceso.

Referencias

Anscombe F.J. (1973). *Graphs in Statistical Analysis. The American Statistician*, Vol. 27, No. 1. (Feb, 1973), pp. 17-21.

Becker H. (2018). Datos, pruebas e ideas. Por qué los científicos sociales deberían tomárselos más en serio y aprender de sus errores. Siglo XXI editores.

[16] Vallejo, I. (2021, p.87). En el libro *El infinito en un junco*.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

Benítez P. (2020). Sociedad y TIC: la difusión de la idea de educación como derecho social en la acción colectiva universitaria chilena del año 2011.Tesis de doctorado. Facultad de Ciencias Sociales, Universidad de Buenos Aires. Disponible en:

http://repositorio.sociales.uba.ar/items/browse?advanced%5B0%5D%5Belement_id%5D=41&advanced%5B0%5D%5Btype%5D=is+exactly&advanced%5B0%5D%5Bterms%5D=Fil%3A+Ben%C3%ADtez%2C+Paulina.+Universidad+de+Buenos+Aires.+Facultad+de+Ciencias+Sociales.+Buenos+Aires%2C+Argentina&sort_field=added&sort_dir=d

Bolstanki L. y Chiapello E. ([1999], 2010). El nuevo espíritu del capitalismo. Akal, Primera reimpresión.

Castells M. (2005). *Internet y la sociedad red*. En De Moraes D. (Coordinador). Por otra comunicación. Los media, globalización cultural y poder. Primera edición. Icaria Editorial s.a. España.

Few S. (2004). *Show me the numbers. Designing tables and graphs to enlighten*. Analytics Press.

Ford M. (2016) El auge de los robots. La tecnología y la amenaza de un futuro sin empleo. Paidós. Primera Edición. Ciudad Autónoma de Buenos aires.

Hilbert M. (2013). *Big data for development: for information- to fnowledge societies*.
UN ECLAC & Annenberger School Communication, University of Southern California.
Disponible en http://ssrn.com/abstract=22055145

Mayer-Shönberger V. y Cukier H. (2015). Big Data. La revolución de los datos masivos. Turner Publicaciones S.L. Segunda Edición.

Ministerio de Desarrollo Social y Familia, Observatorio Social. Disponible en http://observatorio.ministeriodesarrollosocial.gob.cl/

Tapscott D. y Williams A.D. (2012). Macrowikinomics. Nuevas fórmulas para impulsar la economía mundial. Paidós. Primera Edición. Buenos Aires.

Vallejo I. (2021). El infinito en un junco. La invención de los libros en el mundo antiguo. Penguin Random House Grupo Editorial, Tercera Edición, Buenos Aires.

Winocur R. (2010). Robinson Crusoe ya tiene celular. Primera reimpresión. Siglo XXI editores. Primera reimpresión. México.

Páginas web consultadas, Enlaces

http://www.bencinaenlinea.cl/web2/ Consultado 24.09- 30.11 de 2021

https://www.bcn.cl/leychile/navegar?idNorma=276363&idParte= Consultado el 19 de mayo de 2022.

https://datos.gob.cl/about. Consultado el 19 de mayo de 2022.

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/

https://datos.gob.cl/about. Consultado el 19 de mayo de 2022.

https://www.bcn.cl/leychile/navegar?idNorma=141599&idParte=864270 Consultado el 19 de mayo de 2022.

https://www.hacienda.cl/areas-de-trabajo/politicas-macroeconomicas/mepco Consultado el 30 de noviembre de 2021.

http://observatorio.ministeriodesarrollosocial.gob.cl/ Consultado el 19.03-19.05 de 2022

Ensenada, provincia de Buenos Aires, 5, 6 y 7 de diciembre de 2022
ISSN 2250-8465 - web: http://jornadassociologia.fahce.unlp.edu.ar/