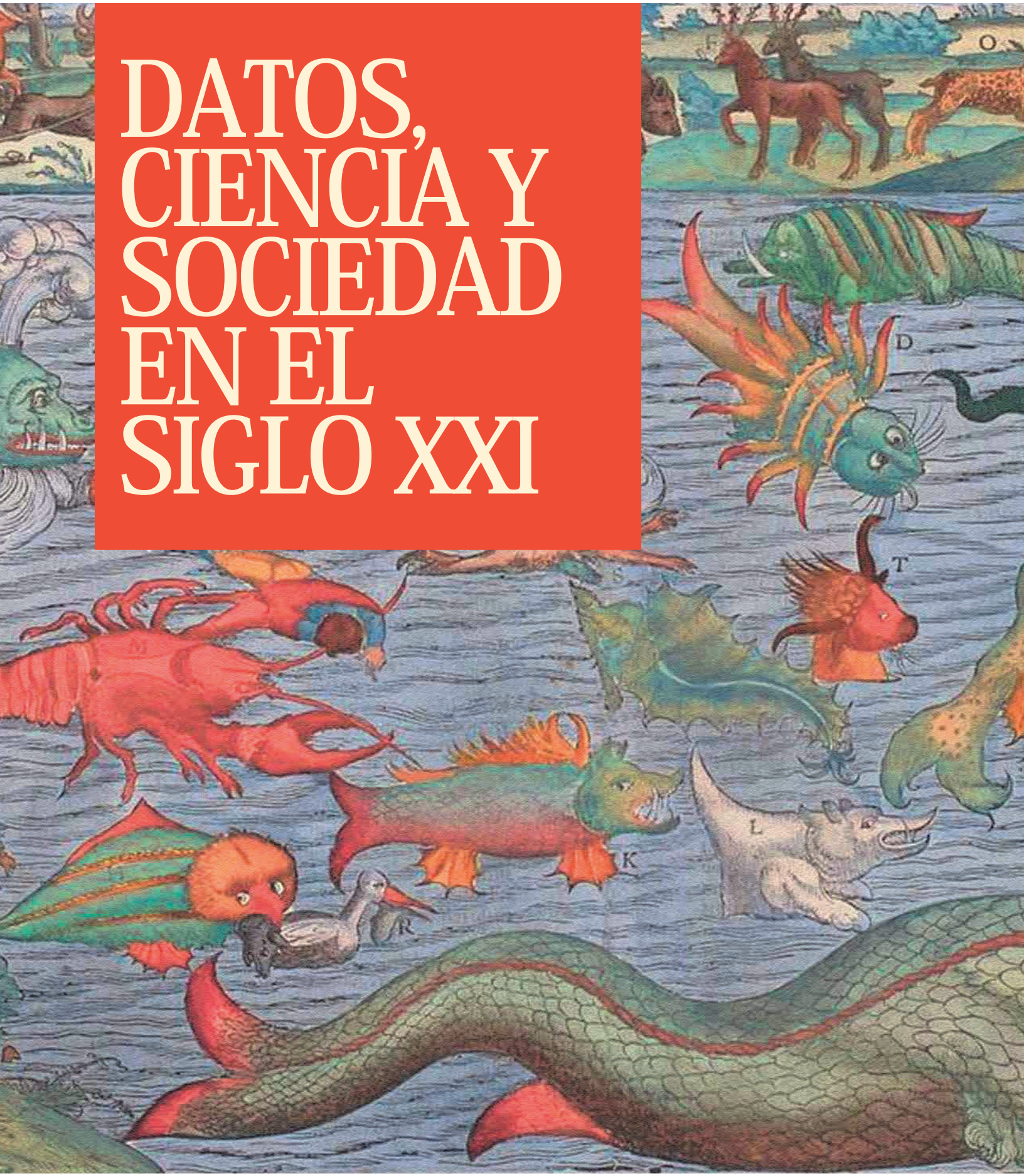
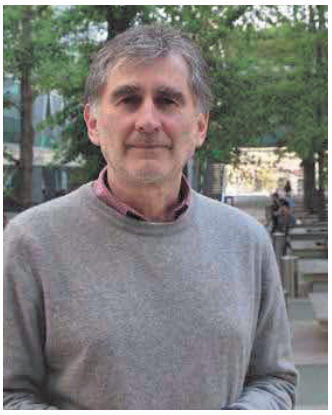
DATOS, CIENCIA Y SOCIEDAD EN EL SIGLO XXI





CLAUDIO GUTIÉRREZ

Profesor Titular, Departamento de Ciencias de la Computación, Universidad de Chile. Investigador Senior del Núcleo Milenio Centro de Investigación de la Web Semántica (CIWS). Ph.D. Computer Science, Wesleyan University; Magíster en Lógica Matemática, Pontificia Universidad Católica de Chile; Licenciatura en Matemáticas, Universidad de Chile. **Líneas de investigación:** Fundamentos de la Computación, Lógica aplicada a la Computación, Bases de Datos, Semántica de la Web, Máquinas Sociales.

cgutierr@dcc.uchile.cl

"Los fundamentos de la experiencia han sido hasta ahora inexistentes o muy débiles; ni se ha buscado ni hecho todavía una recolección o provisión de particulares, capaz o de alguna manera adecuada, ya sea en número, tipo o certeza, de informar al intelecto. [...] La historia natural no contiene nada que se haya investigado de manera adecuada, nada verificado, nada contado, nada pesado, nada medido" (F. Bacon, 1620. Aforismo XCVIII).

La noción de "dato" ("datos") proviene del latín *datum* que significa algo dado. Los primeros usos del término en un contexto científico datan de la mitad del siglo XVII, y aluden, como en la cita de Bacon arriba, a algo que caracterizaría la investigación científica. En su uso moderno, el concepto refiere a colecciones de mediciones e información factual que forma la base de la investigación, el razonamiento, la evidencia que soporta a éste, etc. Con el advenimiento de los computadores hacia mediados del siglo XX se incorporó un nuevo sentido al tradicional. Entonces también se comenzó a entender por datos a las entidades abstractas básicas sobre las que esas nuevas máquinas trabajaban. Sin embargo ambos sentidos, el científico y el computacional, permanecieron confinados hasta hace poco en comunidades técnicas.

La popularización del término "datos" en las portadas de revistas y en cualquier informe que quisiera ser considerado científico es relativamente reciente. Dos metáforas tienen mucha responsabilidad por este milagro: primero, la noción de *"diluvio de datos"* y luego la de "*big data*".

Aunque la imagen del *diluvio de datos* es poderosa (la sociedad y los humanos inundados con datos), la noción es fuertemente engañosa. Primero, sugiere algo producido por un otro, tradicionalmente un dios castigador, o por poderes naturales fuera de nuestro control. Segundo, convierte los actuales niveles de datos en una catástrofe, dándoles una connotación de inseguridad. En resumen, presenta los datos como algo negativo ante lo cual nosotros solamente pudiéramos reaccionar, o a lo más defendernos.

La noción de *big data* (como muchos conceptos del inglés, sin buena traducción) es menos engañosa. Evita la connotación negativa explícita y resalta una de las principales características del fenómeno: su tamaño. En el área de la computación se acuñó el término tempranamente, a principios de la década de 1990, pero el bombo publicitario en los negocios que la popularizó es más reciente. En las ciencias y la investigación el término comenzó a ser ampliamente adoptado recién en el siglo XXI. Mi preocupación con esta noción es que ella aún representa el fenómeno como algo externo e inalcanzable para nosotros. De hecho, mucha gente habla (y piensa) sobre *big data* como un ente ("el" *big data*) oscuro y fantasmal, un sujeto lejos del control de la humanidad, con el que habría que aliarse para sacarle provecho. Hasta es posible leer "el" *big data* como una nebulosa creciendo en torno nuestro esperando ser domada, pero dispuesta a aplastarnos si no la tomamos en serio.

TORRENTES DE DATOS

Retengamos lo esencial: hay una enorme cantidad de datos luchando por, directa o indirectamente, capturar nuestra atención. Una pregunta natural es por qué este despliegue *hoy*. Históricamente, ha habido tsunamis inundando las capacidades simbólicas y semánticas del ser humano. Sin duda la adopción de la escritura y de medios para preservarla deben haber transformado radicalmente las formas tradicionales de interactuar con la información. Más tarde, la imprenta debe haber producido similares remezones entre la población culta. Y más recientemente, los diarios, revistas y la ubicua tecnología de impresión, sumado a la radio y la televisión, agobiaron a la gente con información. En la década de 1930, José Ortega y Gasset, en un discurso al Congreso Internacional de Bibliotecarios, hablaba del "libro furioso" y expresaba así sus preocupaciones sobre este fenómeno:

"Hay ya demasiados libros. Aun reduciendo sobremanera el número de temas a que cada hombre dedica su atención, la cantidad de libros que necesita ingerir es tan enorme que rebosa los límites de su tiempo y de su capacidad de asimilación. [...] La cultura que había libertado al hombre de la selva primigenia, le arroja de nuevo en una selva de libros no menos inextricable y ahogadora. [...] Hay aquí, pues, un drama: el libro es imprescindible en estas alturas de la historia, pero el libro está en peligro porque se ha vuelto un peligro para el hombre" (Misión del Bibliotecario, 1935).

El libro, según Ortega, había llegado a ser un peligro para el hombre. Escuchamos hoy día quejas muy similares sobre los datos. ¿Qué está ocurriendo en el fondo? El fenómeno puede ser parafraseado usando un conocido texto sobre el cambio social: *En un cierto estadio de desarrollo, las fuerzas materiales de la sociedad comienzan a producir más material simbólico que el que las relaciones sociales existentes pueden digerir. De formas de desarrollo de la cultura, esas relaciones se transforman en cadenas que las constriñen. Comienza entonces una era de turbulencia de la información.* El problema que enfrentamos hoy día es que la captura, la producción y la oferta de datos sobrepasa con creces las capacidades humanas y sociales para manipularlos, procesarlos y entenderlos.

Desarrollemos algo más esta idea. Los datos que nos agobian no son los millones de unidades de datos entendibles (lo que aterraba a Ortega), sino que la comprensibilidad de la unidad misma. En otras palabras, no hay libros ilegibles. Cada uno de ellos fue pensado y escrito para ser leído por un humano (aún algunos mamotretos como *En busca del tiempo perdido*). El problema que Ortega alertaba era la casi infinita cantidad de libros que se publicaban. El problema principal, entonces, era la cantidad (y por ello su propuesta de solución era acorde: limitar la producción de libros). Por otra parte, el problema con los datos es que muchas veces la unidad misma (la fuente de datos, el *dataset*) no es inteligible por un humano. Pero además, hay una cantidad casi infinita de *datasets*. El problema es ahora doble: la cantidad y la cualidad.

Una metáfora de la física puede ayudar en este punto. La mecánica tradicional es una disciplina a *escala humana*, en el sentido que permite la interacción directa de la gente con ella. Una bicicleta es un artefacto que podemos entender, reparar, transformar casi enteramente por nosotros mismos. Por el contrario, la química primero, y luego la física atómica, cruzaron la barrera de los objetos que los humanos podemos tocar, y se sitúan en un espacio más allá donde nuestros sentidos incorporados no nos ayudan. Hoy día

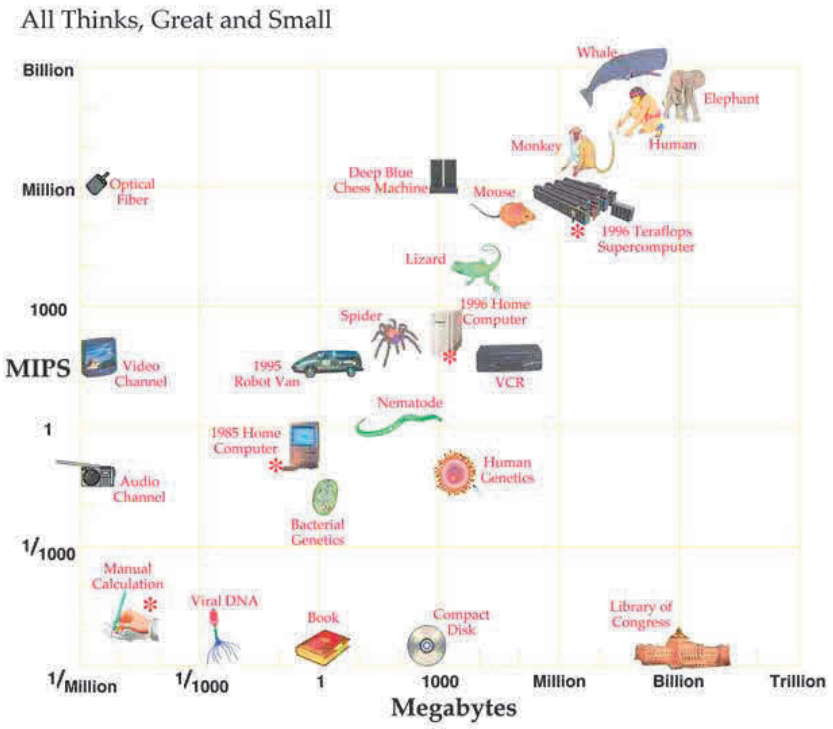


FIGURA 1.
GRÁFICO QUE MUESTRA LOS ALCANCES Y LIMITACIONES DE LA MEMORIA, Y LA CAPACIDAD DE PROCESAMIENTO DEL HUMANO COMPARADO CON OTRAS ESPECIES Y OBJETOS. EN EL EJE HORIZONTAL, CAPACIDAD DE ALMACENAMIENTO EN MEGABYTES. EN EL EJE VERTICAL, CAPACIDAD DE PROCESAMIENTO EN MILLONES DE INSTRUCCIONES POR SEGUNDO (IMAGEN DE HANS MORAVEC).

los avances tecnológicos en medios (poder computacional, memoria, redes, sensores, comunicación, etc.) han incrementado dramáticamente la capacidad de *captura* de datos (sensores, telescopios, Web, etc.); de *producción* de datos (computadores, juegos, media, LHC, etc.); de *almacenamiento* de datos (memorias, medios de almacenamiento, nube, etc.); de *análisis* de datos (técnicas estadísticas, redes neuronales, *deep learning*, etc.). En una frase: el límite que hoy estamos sobrepasando es el de las capacidades humanas para entender y manipular este vasto mundo de objetos simbólicos llamados datos **(Figura 1).**

A pesar de las quejas de Ortega, hasta hace poco nosotros los humanos podíamos lidiar con todo tipo de objetos simbólicos a nuestro alrededor: textos, fotografía, música, películas. Pero este mundo simbólico está creciendo tan rápido que escapa a nuestras capacidades humanas y sociales "dadas", y por lo tanto, sentimos que un oscuro y amenazador, pero fundamentalmente inentendible mundo paralelo crece frente a nosotros. Pero es bueno recalcarlo: no es que ese mundo simbólico no existiera antes. Existía y era muy vasto. Pero era esencialmente volátil. Nadie era capaz de congelarlo ni capturarlo ni menos procesarlo. Contemplar una tormenta es, desde el punto de vista de proceso de información, un proceso extraordinariamente complejo. Pero de ella quedaba a lo más un vago recuerdo. La novedad esencial que enfrentamos hoy, es que es posible "materializar" gran parte de ese proceso en la forma de datos. De alguna manera, las tecnologías digitales de captura, procesamiento, análisis y visualización nos han hecho conscientes de ese mundo con que interactuábamos solo en vivo (ver algunos números en la **Tabla 1**). Mi hipótesis es que este nuevo escenario ha tornado obsoleto los modelos conceptuales que teníamos para enfrentar el mundo simbólico. Entre los mayores desafíos está la noción de escala (Gibson, Ostrom, Ahm, 2000).

EL CONCEPTO DE DATOS (DATA)

Las relaciones entre las nociones de datos, información y conocimiento son complejas y sutiles. Pero para lo que sigue, nos bastará asumir la premisa ampliamente aceptada que los datos son, en cierto sentido, el punto de partida, la base, los bloques básicos, de la información y el conocimiento. Bajo esta hipótesis analizaremos el concepto de datos.

1. Al nivel más básico y abstracto, los datos son una distinción, esto es, un signo de una falta de uniformidad en el mundo externo. Como lo plantea Luciano Floridi (2015), los datos son una fractura en la fábrica del ser, y solo pueden ser planteados como un enganche externo a nuestra información. No son nunca accesados o elaborados independientemente de un cierto nivel de abstracción. Carolina Haythornthwaite (en Zins, 2007) apunta a lo mismo de otra manera: los datos son la más pequeña unidad recolectable asociada a un fenómeno. Normalmente, los datos se encuentran en colecciones que son reunidas para monitorear un proceso, para evaluar una situación y/o obtener un referente para un fenómeno. En resumen, los datos son el estrato más básico en el mundo simbólico. Los datos no tienen significado por sí mismos, pero son la fuente del significado.

2. Por datos entenderemos datos materializados (almacenados digitalmente), esto es, símbolos una vez que han sido congelados materialmente (digitalizados). Desde este aspecto los datos, como los tratamos en este artículo, son parte del mundo "objetivo". Los datos son colecciones materiales de símbolos. Éste es el espíritu de las siguientes definiciones de datos que pueden encontrarse en algunos diccionarios: "Información en forma numérica que puede ser transmitida o procesada digitalmente" (Merriam-Webster), o "las cantidades, caracteres, o símbolos sobre las que un computador realiza sus operaciones, que pueden ser almacenadas o transmitidas en la forma de señales eléctricas o almacenadas en medios magnéticos, ópticos, o mecánicos" (Oxford). En resumen, a pesar de su ambigüedad tecnológica entre lo material y lo inmaterial, los datos son materiales.

3. Las distinciones que definen los datos suponen un contexto implícito. Esta red de significados no es enunciada explícitamente, esto es, no es especificada en el dato mismo. Esto permite variadas interpretaciones de los mismos datos desde diferentes puntos de vista para explorar nuevas dimensiones. Un buen ejemplo es una fotografía. Con alta probabilidad, el fotógrafo la tomó con alguna intención en mente. Pero las futuras generaciones pueden usarla para "ver" dimensiones que no estaban presentes en el foco original del fotógrafo. Lo anterior no significa que usualmente se incluyan algunos contextos explícitos en la forma de metadatos, esto es, datos adicionales que dan información o indican

escala humana	Byte B	~	10^0	un caracter
	Kilo KB	~	10^3	texto escrito
	Mega MB	~	10^6	imágenes, música
	Giga GB	~	10^9	videos
más allá de lo humano	Tera TB	~	10^{12}	Biblioteca Congreso USA
	Peta PB	~	10^{15}	*Data center* grande
	Exa EB	~	10^{18}	Todas las palabras habladas
	Zetta ZB	~	10^{21}	Cantidad de datos globales

TABLA 1.
TAMAÑOS DE DATOS Y ESCALA HUMANA.

relaciones en los datos crudos. En resumen, los datos tienen significado, aunque no siempre explícito. Aunque los datos sean recolectados o construidos con algún objetivo en mente, permiten diversas interpretaciones y pueden apoyar múltiples tesis.

En nuestra área, la ciencia de los datos, los datos son el punto de partida. Nuestra tarea no es aclarar el estatus ontológico de los datos, sino entender sus propiedades, sus "modos de combinación", y ojalá obtener un modelo conceptual para ellos. Para gente que se dedica a obtener, a almacenar, a cuidar, a curar, a procesar, a analizar, a visualizar, los datos son simplemente algo dado, como lo indica su acepción latina. Nuestro asunto en este punto no es la posible semántica que puede ser destilada de los datos, sino los datos como elemento material. Usando el contrapunto entre el mundo de los *bits* y el de los átomos popularizada por Nicolás Negroponte en *Ser Digital*, trabajamos en el mundo de los *bits*, un mundo tan material como el de los átomos, pero con una significación social radicalmente diferente, como veremos.

Aprovechando la oposición *bit*-átomo, otra metáfora puede ayudar a aclarar las relaciones entre esos dos mundos:

$$\frac{\text{Datos}}{\text{Mundo virtual}} = \frac{\text{Átomos}}{\text{Mundo material}}$$

Tensionando esta asociación de ideas, la ciencia de los datos debiera ser la química del mundo virtual. Las ciencias de la información y el conocimiento trabajan con este material, pero a diferentes niveles de agrupamiento y abstracción.

DATOS CIENTÍFICOS Y DE INVESTIGACIÓN

La noción de datos de investigación, bajo los términos de experiencia, hechos, observación, evidencia, etc. tiene una larga historia.

"Observación" en su sentido científico ya se menciona por Aristóteles; Bacon argumenta su relevancia para la investigación; y la conciencia de las sutilezas de sus conexiones con el conocimiento se remontan a los comienzos del siglo XX. Sin embargo, es solo al despuntar el siglo XXI que los datos comienzan a ser pensados como motor de la ciencia **(Figura 2).** El Premio Turing Jim Gray escribía en 2007:

"Originalmente, solo había ciencia experimental, y luego hubo ciencia teórica, con las leyes de Kepler, las leyes del movimiento de Newton, las ecuaciones de Maxwell, y así sucesivamente. Entonces, para muchos problemas, los modelos teóricos crecieron de manera demasiado complicada para ser resueltos analíticamente, y la gente tuvo que comenzar hacer simulaciones. Estas simulaciones nos han acompañado a través de gran parte de la última mitad del milenio anterior. Hoy día, estas simulaciones están generando muchos datos, junto con un tremendo incremento en datos de las ciencias experimentales. [...] El mundo de las ciencias ha cambiado, y no hay duda acerca de ello. Las técnicas y tecnologías para esta ciencia intensiva en datos son tan diferentes de la ciencia computacional que vale la pena distinguirla como un nuevo paradigma de exploración científica" (Gray, 2007).

Este cambio empujado por las fuerzas materiales de la sociedad está produciendo fuertes cambios sociales, en particular está dando un valor prominente a los datos. El argumento funciona como sigue.

Desde la revolución industrial ha habido conciencia del creciente rol de la ciencia en la economía, pero es solo recientemente que la ciencia ha comenzado a jugar un rol central en ella, como lo reconoce la OCDE:

"El término `economía basada en el conocimiento´ resulta de un pleno reconocimiento del rol del conocimiento y la tecnología en el crecimiento económico. El conocimiento encarnado en los seres humanos (como `capital humano´) y en la tecnología, ha sido siempre

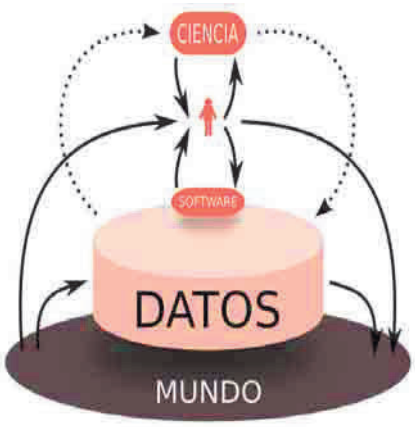


FIGURA 2.
LOS DATOS CONFORMAN HOY CASI UNA RÉPLICA DEL MUNDO MATERIAL Y DE ESA MANERA TRANSFORMAN EL MODO DE HACER CIENCIA.

central para el desarrollo económico. Pero solo en los últimos pocos años se ha reconocido su importancia relativa, justo cuando esa importancia está creciendo. Las economías de la OCDE son hoy más fuertemente dependientes de la producción, distribución y uso del conocimiento que nunca antes" (OCDE, 1996).

De esta afirmación y de la premisa establecida por Jim Gray ("la ciencia hoy está fuertemente basada en los datos"), se sigue la conclusión: *los datos son la nueva materia prima del nuevo proceso de producción.* Una versión algo más alegórica de esta conclusión es: "Los datos son el nuevo petróleo" *(data is the new oil)* **(Figura 3).**

A medida que los datos juegan un rol central en la economía, su proceso de producción comienza a caer bajo la presión por la eficiencia. La división del trabajo comienza a afectar su ciclo –captura, curatoría, análisis, visualización– que tradicionalmente era hecho por la misma persona o equipo (Tycho Brahe / Copérnico son más bien una excepción). Los científicos y sus colaboradores diseñan el experimento o el proceso de recolección de datos (Von Humboldt, Darwin, Mendel, Pasteur, etc.). En particular, hoy hay una tendencia creciente a separar los usos y la producción/recolección de datos. De esta manera, los datos comienzan a adquirir cierto grado de autonomía.

FIGURA 3.
LOS DATOS SON EL NUEVO PETRÓLEO. PERO, ¿TIENES TÚ LOS RECURSOS PARA REFINARLOS? (FUENTE: HTTPS://SUCCESSFLOW.CO.UK).

Otra faceta relevante de los datos científicos es el viejo, pero actual debate, sobre el estatus epistémico de la observación versus la experimentación. La primera, un proceso donde el observador actúa sin tocar, sin alterar, sin preguntar, a su objeto de estudio. El segundo es un producto directo de la manipulación del objeto para extraer lo que sea necesario. James Bogen presenta un elocuente ejemplo: "Mirar un fruto en una vid y contemplar su color y forma sería observarlo. Extraer su jugo y aplicarle reactivos para probar la presencia de compuestos de cobre sería hacer un experimento" (Bogen, 2017). La diferencia, si hubiera una absoluta, es sutil. Uno puede establecer una contraparte en términos computacionales como la pregunta: ¿Datos estáticos o dinámicos? ¿Datos a granel o APIs? La discusión es relevante no solo para cómo recolectar o producir datos, sino sobre todo, para determinar cómo almacenarlos y cómo entregarlos a los usuarios finales.

Hoy podemos "exponer" datos vivos, "observaciones", en la forma de APIs a través de cámaras, sensores, etc. Estas fuentes están convirtiéndose en fuentes muy relevantes de datos. Hay un interés creciente en tecnologías diseñadas para procesar esos datos en vivo, *"online"*, esto es, como *streams* de datos. Hay ya muchos ejemplos de uso masivo, como valor de cambio de monedas, canales meteorológicos, noticias en vivo, etc.

Por último, queremos llamar la atención a las borrosas diferencias entre datos "científicos" y "comunes". Los datos vienen de diversas fuentes y en muchas formas. Hablamos de datos científicos para referirnos a aquellos recolectados sistemáticamente en el marco de una investigación científica. Hoy hay muchas empresas y organizaciones de datos, fuera de lo que uno podría considerar proyectos o instituciones "científicas", particularmente en el ámbito social (que están entre los datos más populares, valiosos y... sucios). Tweets, identidades y comportamiento de usuarios en redes sociales, huellas sociales de muchos tipos, imágenes y videos personales, etc., están entre los datos más valiosos. Cada día se hace más difícil trazar una clara línea divisoria entre datos científicos y el "resto". Al final, todos los datos son recolectados con algún propósito (nadie ocuparía tiempo, energía o recursos para recolectar datos que no tendrían algún, aunque sea difuso o lejano, objetivo).

EL CARÁCTER SOCIAL DE LOS DATOS

Hemos aprendido que los datos están en todas partes; que los datos son relevantes; que tienen valor. No es sorprendente, entonces, que agencias internacionales, gobiernos, comunidades y empresas estén concibiendo formas de acercarse a, hacerse o simplemente tomar ventaja de, este nuevo bien.

Como vimos, los datos son un recurso esencial para el desarrollo del conocimiento científico, y como tal, relevante para la comprensión de nosotros como humanos, para el desarrollo de nuestras sociedades, y para la satisfacción de necesidades y deseos personales. Por otra parte, como el "nuevo petróleo", esto es, como un bien económico, los datos están bajo las tensiones de las categorías de la economía y del poder.

Pero los datos tienen características muy propias. Un enfoque ingenuo sería tratarlos de manera similar a como se concibe el conocimiento, que concebido como bien, tiene diferencias notables con los bienes materiales tradicionales: es no-excluible y no-sustraible. El entonces economista jefe del Banco Mundial, Joseph Stiglitz lo caracterizaba así al despuntar el siglo XXI:

"Un bien público tiene dos propiedades críticas, consumo no-competitivo (el consumo de un individuo no quita el de otro) y no-exclusivo (es difícil, si no imposible excluir a un individuo de gozar el bien). El conocimiento es un bien público que requiere soporte a nivel global (Stiglitz, 1998).

Si cambiamos "conocimiento" por datos obtenemos un programa para los datos como bien público. Es en efecto, el programa de muchos gobiernos y agencias internacionales. Por ejemplo, el foco del Banco Mundial es hacer los datos accesibles a particulares para "permitir a los hacedores de políticas y grupos de defensa o presión tomar decisiones bien informadas y medir mejoras con más precisión" (Banco M undial, 2010).

Basado en lineamientos similares, la OCDE define su programa para acceso abierto, definido en sus *Principios de acceso a datos de investigación* como sigue: "Apertura significa acceso en términos iguales para la comunidad internacional de investigación al menor costo posible, preferiblemente a no más que el costo marginal de diseminación. Acceso abierto a datos de investigación usando

fondos públicos debiera ser fácil, oportuno, amigable, y preferiblemente basado en Internet" (OCDE, 2007). En estos principios están basadas la ola de políticas de transparencia e interoperabilidad para los gobiernos tan populares estos días.

Un buen ejemplo de estas iniciativas en el área científica es la política de datos abiertos de la *National Science Foundation* del Gobierno norteamericano. Ella establece que las agencias deben fomentar la apertura "en los niveles permitidos por la ley y sujeto a la privacidad, confidencialidad, seguridad y otras restricciones válidas". Definen datos abiertos como "datos públicos estructurados en una forma que sean totalmente accesibles y usables". Y argumentan su importancia pues "los datos que son abiertos, disponibles y accesibles ayudarán a incentivar la innovación e informar cómo las agencias debieran conducir sus programas para servir mejor las necesidades públicas". Establecen siete principios de consistencia para apertura de datos: que sean públicos, accesibles, describibles, reusables, completos, oportunos y administrados después de publicados.

Todas esas iniciativas de apertura de organizaciones de poder fueron motivadas e impulsadas por la presión de diversas comunidades conocidas como movimiento de datos abiertos (*open data*). Su noción de datos abiertos está tomada esencialmente del mundo de las comunidades de *"open source"* y *"open access"*. La "traducción" de esas nociones en el mundo de los datos conlleva los mismos temas y desafíos (no más, no menos) que en esos campos. El *Open Data Handbook* lo define así: "Datos abiertos son datos que pueden ser libremente usados, reusados y redistribuidos por cualquiera, sujetos solo, a lo más, a los requerimientos de atribución y compartir. Como podemos ver, hay aquí una concepción más amplia que la del Banco Mundial, la OCDE y las agencias internacionales y de gobiernos, cuyas agendas de apertura de datos están gatilladas por preocupaciones económicas.

CONSIDERACIONES FINALES: MÁS ALLÁ DEL ACCESO

A pesar de los avances sociales que estas políticas para datos han traído, aún quedan asuntos muy importantes por definirse.

La mayoría de los enfoques usados para abordar la noción de "datos abiertos", asocian implícitamente "datos" con conocimiento e información. Y ciertamente, una de las mayores amenazas para estas últimas es efectivamente el "cercado" (*enclosure*) de ellas (en la forma de patentes y copyrights). Nótese que un supuesto clave de este argumento es que el "bien" bajo amenaza de cercado es algo que está listo para ser consumido. Luego, el objetivo último sería el acceso a ese bien, que permitiría que cualquiera lo consuma. Esa premisa vale para datos simples, como planillas de cálculo y listas, cuyo mejor ejemplo son muchos datos de transparencia gubernamental. Pero esto no vale para la mayoría de los datos hoy, y en particular para lo que se llama usualmente *big data*. El acceso en este caso es solamente un primer paso en el ciclo de datos que incluye recolección, curatoría, análisis, y visualización. Los recursos y tecnología necesarios para almacenar y curar esos datos, para analizarlos, y finalmente visualizarlos o usarlos, son gigantescos, usualmente fuera del alcance de la persona u organización común. Los datos no existen aisladamente, realmente los datos forman un ecosistema **(Figura 4**). El desafío de la escala muestra las limitaciones del enfoque centrado solo en el acceso.

Es aquí que el marco conceptual de los *comunes* (*commons*) viene al rescate. Como Charlotte Hess y Elinor Ostrom afirman,

"Las preguntas esenciales de cualquier análisis de comunes son inevitablemente sobre equidad, eficiencia y sustentabilidad. La equidad refiere a asuntos de apropiación justa o igual, y contribución a, el mantenimiento de un recurso. La eficiencia habla de la producción, gestión y uso óptimo de los recursos. La sustentabilidad mira los resultados sobre el largo plazo" (Hess & Ostrom, 2006).

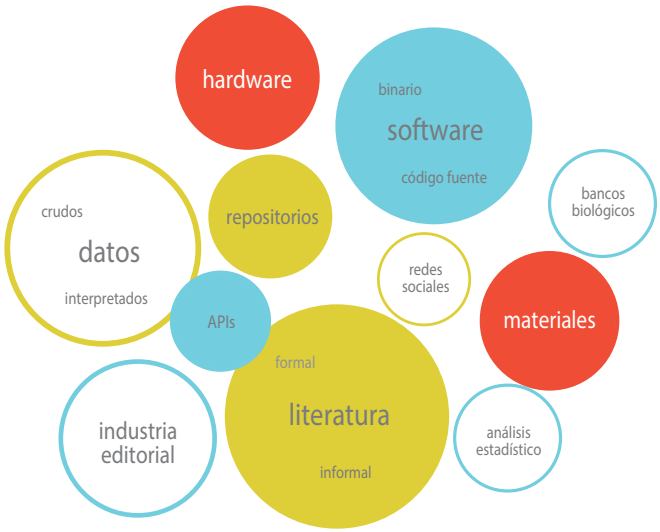


FIGURA 4.
ECOSISTEMA DE DATOS: NO SOLO INVOLUCRA DATOS, SINO EL SOFTWARE Y EL HARDWARE NECESARIOS PARA PROCESARLOS, Y UN CONJUNTO DIVERSO DE BIENES (IMAGEN DE PUNEET KISHOR).

Debido a los enormes tamaños y complejidad de los datos, pensar acerca de los datos como comunes implica incluir el ciclo completo de los datos como un común. Los datos son un recurso compartido (y producido) por grupos de personas. En su faceta intangible, es un bien claramente no-exclusivo y no-competitivo, muy similar al conocimiento: compartirlo no significa casi ningún esfuerzo; consumirlo no sustrae de esa posibilidad a otros. El problema aparece cuando una considera su faceta material. Y aquí surgen todos los problemas de un bien material, y es donde la teoría de los comunes aparece en todo su potencial: cómo resolver los problemas de cercado, polución, degradación, y no sustentabilidad de un bien común.

Imagen de Tim Berners-Lee (modificada).

Los datos llegaron a nuestras sociedades para quedarse. Ya vimos que los datos son el nuevo petróleo. La alegoría puede extenderse para incluir la historia del petróleo sobre la tierra y lo que ella nos enseña: los posibles conflictos que la apropiación de este nuevo recurso puede traer. Dependerá de nosotros, humanos, definir qué queremos de este nuevo petróleo y cómo podemos usarlo para mejorar nuestras vidas y nuestras sociedades.

La discusión que debiéramos abrir hoy es cómo nos gustaría gestionar y gobernar este nuevo bien, incluido el cómo generarlo, accesarlo, almacenarlo, curarlo, procesarlo, analizarlo y entregarlo. Los comunes ofrecen intuiciones muy adecuadas para abordar estos temas. ■

REFERENCIAS

[1] F. Bacon, *The New Organon*. Edit. L. Jardine, M. Silverthorne. Cambridge Univ. Press, 2000.

[2] Banco Mundial. World Bank Open Data Initiative. World Bank, 4/30/2010. data.worldbank.org

[3] J. Bogen. Theory and Observation in Science (Version Mar 28, 2017). (In Stanford Encyclopedia of Philosophy). https://plato.stanford.edu/entries/science-theory-observation/

[4] L. Floridi. Semantic Conceptions of Information SEP (Version Jan 7, 2015) https://plato.stanford.edu/entries/information-semantic/

[5] C. C. Gibson, E. Ostrom, T. K. Ahm. The concept of scale and the human dimensions of global change: a survey. Ecological Economics, 32 (2000), 217-239.

[6] J. Gray. Jim Gray on eScience: A Transformed Scientific Method. (Basado en una transcripción de una charla dada por Jim Gray al NRC-CSTB1 en Mountain View, California, el 11 de enero de 2007). En: T. Hey, S. Tansley, K. Tolle. The Fourth Paradigm. Data-Intensive Scientific Discovery. Microsoft Research, 2009.

[7] Ch. Hess, E. Ostrom (Ed.) Understanding Knowledge as a Commons From Theory to Practice. MIT Press 2006.

[8] National Science Foundation. Open Data at NSF. https://www.nsf.gov/data/

[9] OECD. La economía basada en el conocimiento. (The Knowledge-based economy.) OECD, París 1996.

[10] OECD. OECD Principles and Guidelines for Access to Research Data from Public Funding, 2007.

[11] Open Data Foundation. What is Open Data? Open Data Handbook. http://opendatahandbook.org/guide/en/what-is-open-data/

[12] J. Ortega y Gasset. La misión del bibliotecario. (1935). Edición digital, Consejo Nacional para la Cultura y las Artes, México, 2005.

[13] J. Pearl, R. Dechter. Learning Structure From Data: A Survey. Proc. COLT'89, pp. 230-244.

[14] D. Rosenberg. Data Before the Fact. American Historical Association, 2012.

[15] J. Stiglitz, Knowledge as a Global Public Good. In: Global Public Goods: International Cooperation in the 21st Century, 1998.

[16] Ch. Zins. Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology 58, 2007. pp. 479-493.