Datos abiertos en un mundo de grandes datos. Un acuerdo internacional ICSU-IAP-ISSC-TWAS *

Geoffrey Boulton, Simon Hodson, Dominique Babini, Jianhui Li, Tshilidzi Marwala, Maria G. N. Musoke, Paul F. Uhlir y Sally Wyatt **

1. El mundo de los grandes datos (*big data*)

La revolución digital de décadas recientes es un evento histórico mundial tan profundo y más penetrante que la introducción de la imprenta. Ha creado una explosión sin precedentes en la capacidad de adquirir, almacenar, manipular y transmitir instantáneamente grandes y complejos volúmenes de datos, con profundas implicaciones para la ciencia.[1] La velocidad del cambio es formidable. En 2003 los científicos declararon que el mapeo del genoma humano estaba completo. Llevó más de diez años y costó un billón de dólares; hoy se tarda apenas unos días y cuesta una pequeña fracción de dicho monto (mil dólares). Los grandes volúmenes de datos (*big data*), de donde emanan flujos sin precedentes de datos desde y hacia los sistemas computacionales, y los datos amplios (*broad data*), en los que numerosos conjuntos de datos pueden ser semánticamente vinculados para crear significados más profundos, son los motores de esta revolución, ofreciendo nuevas oportunidades a las ciencias naturales, sociales y humanas.

* La versión extendida de este acuerdo internacional está disponible en inglés en: http://www.icsu.org/science-international/accord.

** *Geoffrey Bouton*: Universidad de Edimburgo, Escocia; presidente de CODATA y del grupo de trabajo que redactó este acuerdo internacional. *Simon Hodson*: director ejecutivo de CODATA. *Dominique Babini*: CLACSO y Universidad de Buenos Aires, Argentina. *Jianhui Li:* Academia China de Ciencias-CNIC. *Tshilidzi Marwala*: Universidad de Johannesburgo, Sudáfrica. *Maria G. N. Musoke*: Universidad de Makerere, Uganda. *Paul F. Uhlir*: Academia Nacional de Ciencias de Estados Unidos. *Sally Wyatt:* Universidad de Maastricht, Holanda, y eHumanities-KNAW.

1. La palabra “ciencia” se utiliza para referirse a la organización sistemática del conocimiento que se puede explicar racionalmente y aplicar en forma confiable. El concepto “ciencia” se utiliza, como en la mayoría de los idiomas distintos del inglés, para incluir todos los ámbitos, incluyendo las humanidades y las ciencias sociales, así como las disciplinas STEM (ciencia, tecnología, ingeniería y medicina, por sus siglas en inglés).

2. Las oportunidades

Las oportunidades científicas de este mundo rico en datos residen en descubrir patrones que hasta ahora han estado fuera de nuestro alcance; en vincular y correlacionar mejor los diferentes aspectos de los sistemas para entender su comportamiento; en caracterizar la complejidad; y en la iteración entre descripciones del estado de sistemas complejos y simulaciones que pronostican su comportamiento dinámico. Hay muchas áreas de investigación donde estas capacidades son profundamente relevantes: en predicción meteorológica y climática; en la comprensión del funcionamiento del cerebro; en el comportamiento de la economía global; en la evaluación de la productividad agrícola; en las previsiones demográficas; en historias a desentrañar; y en muchos de los desafíos globales contemporáneos como los del cambio ambiental, las enfermedades infecciosas y la migración masiva, que requieren combinar conocimientos y datos de muchas disciplinas.

3. Los desafíos

Aprovechar estas oportunidades plantea serios desafíos a la forma en que la ciencia se ejecuta y se organiza. Los datos abiertos son el elemento común que lo hace posible.

3.1. El imperativo de los datos abiertos

El rol fundamental de la investigación financiada con fondos públicos es agregar al acervo de conocimiento y comprensión que son esenciales para el discernimiento humano, la innovación y el bienestar social y personal. Las tecnologías y los procesos de la revolución digital proporcionan un poderoso medio a través del cual la productividad y la creatividad científica se pueden mejorar permitiendo que los datos y las ideas fluyan en forma abierta, rápida y generalizada a través de la interacción en red de muchas mentes. Si esta revolución social en la ciencia ha de realizarse, es vital que adoptemos, como posición por defecto, que los datos financiados con fondos públicos sean accesibles públicamente y reutilizables cuando se completa el proyecto de investigación a través del cual se han recolectado.

3.2. Mantener la autocorrección

La apertura de la evidencia de aseveraciones científicas (los datos) es la base del progreso científico. Permite que la lógica de un argumento pueda ser examinada y que la reproducibilidad de las observaciones o experimentos puedan ser comprobados, apoyando o invalidando de ese modo las aseveraciones. Cuando se publica un documento haciendo una afirmación científica, es esencial que los datos probatorios, los metadatos relacionados que permiten el re-análisis y los códigos utilizados en la manipulación por computadora se abran al mismo tiempo al escrutinio para asegurar que se mantiene el proceso vital de autocorrección. Recientes demostraciones en varias disciplinas de altas tasas de no-reproducibilidad de los resultados de los trabajos publicados enfatizan la necesidad crucial de revitalizar procesos de datos abiertos para un mundo de grandes datos. La apertura no es, sin

embargo, suficiente. Los datos deben ser inteligentemente abiertos, lo que significa que deben ser: descubribles, accesibles, inteligibles, evaluables y reutilizables.

3.3. Adaptar el razonamiento científico

Muchas de las complejas relaciones que ahora tratamos de capturar a través de grandes datos (*big data*) o datos amplios (*broad data*) enlazados se encuentran más allá de la capacidad analítica de muchos métodos estadísticos clásicos. Requieren enfoques matemáticos más profundos, incluyendo métodos topológicos para asegurar que las inferencias extraídas de grandes datos y datos amplios sean válidas. El análisis computacional intensivo en datos y la capacidad de las computadoras de aprendizaje automático (*machine-learning*) están muy presentes y tienen importantes implicaciones para el descubrimiento científico. La complejidad de los patrones que las máquinas son capaces de identificar no son fácilmente captados por los procesos cognitivos humanos, lo que plantea cuestiones profundas sobre la interfaz hombre-máquina y lo que puede significar ser un investigador en el siglo XXI.

3.4. Restricciones éticas

El principio de datos abiertos tiene implicaciones éticas para los investigadores y para los sujetos investigados. Puede parecer que resta valor a los intereses individuales de los investigadores que generan los datos, de modo que es necesario desarrollar nuevas formas de reconocer y recompensar su contribución. La privacidad de los sujetos a los cuales se refieren necesita ser protegida. En un régimen de intercambio abierto en el que los datos son transmitidos por sus creadores, hay pérdida de control sobre su uso futuro, al tiempo que los procedimientos de anonimización han demostrado ser incapaces de garantizar la seguridad de los registros personales.

3.5. Participación global abierta

Los grandes datos y los datos abiertos tienen un gran potencial para beneficiar a los países con menos recursos, y en especial a los menos desarrollados. Sin embargo, los países menos desarrollados suelen tener sistemas nacionales de investigación con pocos recursos disponibles. Si no pueden participar en investigación basada en grandes datos y datos abiertos, la brecha podría crecer exponencialmente en los próximos años. Ellos no podrán recolectar, almacenar y compartir datos, participar en la investigación global, contribuir como socios plenos en los esfuerzos globales sobre el cambio climático, el cuidado de la salud y la protección de los recursos, y no podrán beneficiarse plenamente de tales esfuerzos, donde las soluciones globales sólo se lograrán si hay una participación internacional. Por lo tanto, las naciones emergentes y desarrolladas tienen ambas un claro y directo interés en ayudar a movilizar plenamente el potencial científico de los países menos desarrollados y de esta manera contribuir al logro de los objetivos de desarrollo sostenible de la Organización de las Naciones Unidas.

3.6. Aprovechar la oportunidad

La apertura efectiva de datos sólo puede ser realizada si hay acción sistémica a nivel

personal y disciplinario, nacional e internacional. Aunque la ciencia es una actividad internacional, se realiza dentro de distintivos sistemas nacionales de responsabilidad, organización y gestión, todos los cuales necesitan responder a la oportunidad. Los financiadores y las instituciones de investigación deben financiar e implementar procesos que aligeren la carga que significa para los investigadores abrir en forma inteligente los datos y apoyar los procesos de datos abiertos.

Un número creciente de comunidades de investigación han descubierto los beneficios de compartir datos en campos tan variados como la lingüística, la bioinformática y la cristalografía química, y han hecho grandes avances en el logro de beneficios en sus disciplinas a través de la colaboración internacional para facilitar el acceso y el uso de datos abiertos.

Las responsabilidades también corresponden a organismos internacionales como el *Committee on Data for Science and Technology* (CODATA) del *International Council for Science* (ICSU), su *World Data System* (WDS) y la *Research Data Alliance* (RDA), para promover y apoyar desarrollos de los sistemas y procedimientos que garanticen a nivel internacional el acceso, la interoperabilidad y la sustentabilidad de los datos.

3.7. Ciencia abierta y conocimiento público

La idea de "ciencia abierta" se ha desarrollado en reconocimiento de la necesidad de fortalecer el diálogo y el compromiso de la comunidad científica con la sociedad en general para hacer frente a muchos problemas actuales mediante la formulación recíproca de los temas y el diseño, ejecución y aplicación colaborativa de la investigación. Hay, por supuesto, límites legítimos a la apertura, tales como la preocupación de la necesidad de proteger la seguridad, privacidad y propiedad mediante mecanismos aplicados juiciosamente. También hay tendencias compensatorias hacia la privatización del conocimiento que están en contradicción con la ética de la investigación científica y la necesidad básica de la humanidad de utilizar las ideas libremente. Si la iniciativa científica no debe hundirse bajo tales presiones, se requiere un compromiso firme de la comunidad científica global con los principios de datos abiertos, información abierta y conocimiento abierto.

4. Los principios de datos abiertos

Tal es la importancia y magnitud de los desafíos para la práctica de la ciencia en la revolución de datos que *Science International* considera oportuno promover la siguiente declaración de principios de datos abiertos.

4.1. Responsabilidades

4.1.1. Los científicos

Los científicos financiados con fondos públicos tienen la responsabilidad de contribuir al bien público a través de la creación y comunicación de nuevos conocimientos, en los cuales los datos asociados son parte intrínseca. Ellos deben hacer que esos datos

estén disponibles abiertamente a los demás, después de su producción y tan pronto como sea posible, en formas que permitan que los datos puedan ser reutilizados y utilizados con otros propósitos.

Los datos que proporcionan evidencia de las afirmaciones científicas publicadas deben hacerse disponibles y públicos al mismo tiempo, así como abiertos de manera inteligente.[2] Esto debe permitir que la lógica de relación entre los datos y las afirmaciones pueda ser rigurosamente analizada y la validez de los datos comprobada por replicación de experimentos u observaciones. En la medida de lo posible, los datos deben ser depositados en repositorios bien gestionados y confiables, con bajas barreras de acceso.

4.1.2. Las instituciones de investigación y las universidades
Tienen la responsabilidad de crear un entorno de apoyo para los datos abiertos. Esto incluye brindar capacitación en gestión, preservación y análisis de datos y el soporte técnico pertinente, incluyendo servicios de biblioteca y de gestión de datos. Las instituciones que emplean a los científicos, y los organismos que los financian, deben desarrollar incentivos y criterios de promoción para aquellos involucrados en los procesos de datos abiertos. Es necesario un consenso a nivel nacional sobre tales criterios, e idealmente a nivel internacional, para facilitar pautas deseables de movilidad de los investigadores. En el espíritu actual de internacionalización, las universidades y otras instituciones científicas en los países desarrollados deben colaborar con sus contrapartes en los países en desarrollo para movilizar las capacidades de uso intensivo de los datos.

4.1.3. Los editores
Tienen la responsabilidad de poner a disposición de los evaluadores los datos durante el proceso de revisión, de requerir acceso abierto inteligente a los datos al mismo tiempo que la publicación que los utiliza, y exigir las referencias y citas completas de esos datos. Los editores también tienen la responsabilidad de poner a disposición el registro científico para su posterior análisis mediante el suministro abierto de los metadatos y el acceso abierto para minería de textos y datos.

4.1.4. Los organismos de financiación
Deben considerar los costos de los procesos de datos abiertos en los proyectos de investigación como parte intrínseca del costo de su realización, y deben proporcionar recursos y políticas adecuados para la sostenibilidad a largo plazo de la infraestructura y repositorios. La evaluación del impacto de la investigación, particularmente los indicadores que involucran métricas de citación, debe tomar en cuenta la contribución de los creadores de datos.

4.1.5. Las asociaciones profesionales, sociedades científicas y academias
Deben desarrollar directrices y políticas de datos abiertos y promover las

2. Véase la versión completa del documento en: http://www.icsu.org/science-international/accord.

oportunidades que los datos abiertos ofrecen, de forma tal que reflejen las normas epistémicas y las prácticas de sus miembros.

4.1.6. Bibliotecas, archivos y repositorios
Tienen responsabilidad en el desarrollo y prestación de servicios y normas técnicas para los datos, de tal forma que aseguren su disponibibilidad para quienes deseen utilizarlos, y para que los datos sean accesibles en el largo plazo.

4.2. Los límites de la apertura

Los datos abiertos deben ser la posición por defecto para la ciencia financiada con fondos públicos. Las excepciones deben limitarse a cuestiones de privacidad, de seguridad y de uso comercial en el interés público. Las excepciones propuestas deben justificarse caso por caso, y no como exclusión general.

4.3. Prácticas habilitantes

4.3.1. Citación y procedencia
En publicaciones académicas, cuando los investigadores utilizan datos creados por otros, éstos deben ser citados con referencia a su autor, a su procedencia y a un identificador digital permanente.

4.3.2. Interoperabilidad
Tanto los datos de investigación como los metadatos que permiten la evaluación y reutilización de los datos deben ser interoperables en la mayor medida posible.

4.3.3. Reutilización no restrictiva
Si los datos de investigación no están ya en el dominio público, deben ser etiquetados como reutilizables por medio de una renuncia a los derechos o una licencia no restrictiva que deje en claro que los datos pueden ser reutilizados sin mayor requisito que el reconocimiento al productor.

4.3.4. Capacidad de vinculación
Los datos abiertos deben, siempre que sea posible, estar vinculados con otros datos basados en su contenido y contexto, con el fin de maximizar su valor semántico.