Acta Botánica Cubana

OPEN ACCESS

ISSN 2519-7754 RNPS 2402
www.revistas.geotech.cu/index.php/abc

‖ ARTÍCULO DE REVISIÓN ‖

Vol. 219, No. 2 (julio-diciembre 2020): 35 - 45

Los datos pueden ser más importantes que las publicaciones: hay que hacerlos valer

Data can be even more important that papers: let enforce their value

Dennis Denis Ávila

Facultad de Biología, Universidad de La Habana, Cuba.

Correspondencia: dda@fbio.uh.cu

Recibido: 03 de junio de 2020

Aceptado: 30 de agosto de 2020

Conflicto de intereses: El autor declara que no existen conflictos de intereses.

https://eqrcode.co/a/tGKF1W

RESUMEN

Los datos en una investigación son considerados como las materias primas para la extracción de información o conocimiento científico que sustenta las publicaciones y la toma de medidas. Existe la creencia de que cuando la investigación termina y se publica, los datos utilizados pierden su valor y tienden a desecharse o almacenarse, invisibles e inaccesibles para otros investigadores. Esta es una visión superficial que esconde la realidad de que los datos son uno de los productos primarios más valiosos de la ciencia, con múltiples valores adicionales que permiten afirmar que pueden llegar a ser más importantes incluso que las propias publicaciones. Pero para que se pueda expresar este valor, debe seguirse una serie de aspectos metodológicos que incluyen desde la estructuración apropiada de las matrices o ficheros, su documentación a través de metadatos, la selección y asignación de las licencias, el identificador digital o su publicación como Artículo de datos. En la presente revisión se discuten algunas de estas prácticas para revalorizar los datos de las investigaciones ecológicas y poder explotarlos en toda su extensión, dado el desconocimiento generalizado de la infraestructura existente y los métodos para hacerlos disponibles, recuperables y utilizables por tiempo indefinido. Para ello se responden las preguntas básicas: ¿por qué y para qué? ¿cómo? ¿dónde? y ¿cuándo? Se termina haciendo un llamado al cam bio desde la visión tradicional que se enfoca solo en el análisis de los datos. Se requiere una nueva visión que ponga más énfasis en la organización de los datos y su documentación para hacerlos públicos, conjuntamente con los artículos.

Palabras clave: ciencia abierta, compartición de datos, metadatos, repositorios

ABSTRACT

Research data are considered the raw material for acquiring scientific knowledge for articles and decision-making process. There is a believe that once research finished and papers are published, data loss value and can be disposal or keep in forgotten corners, invisibles and inaccessible for any other scientist. This is a shallow view that ignores the reality that data is one of the most valuable results in science, with additional values that raise them to even more important place than papers. But for getting this value a series of steps and procedures should be followed from matrix structure or file ordering, documentation with proper metadata, selection and assignment of licenses and digital identification or their publication as data papers. In current review, some of these practices are discussed to reinforce ecological research data values, due to the generalized ignorance on current available infrastructure and method to secure their available, retrieval, and use for all time. To do that, I answer the basic questions of why? why for? how? where? and when? I ended by making a call for a vision change from current focus on data. A new emerging vision includes more efforts directed to Data organization and documentation for data sharing, along with the papers.

Keywords: data sharing, metadata, open Science, repositories

INTRODUCCIÓN

Los datos en una investigación son considerados como las materias primas para la extracción de información o conocimiento científico que sustenta las publicaciones y la toma de medidas. Generalmente está bien clara la noción de que información no es equivalente a datos, y que lo que se publica es la información. Por ello, los análisis, visualizaciones gráficas y tablas conforman los aspectos centrales de los artículos científicos de investigación. Sin embargo, prevalece la noción implícita de que cuando un proyecto de investigación finaliza, y se escriben y publican los artículos o informes, los datos utilizados ya pierden su valor de uso y pueden desecharse (Goodman *et al.*, 2014) o almacenarse en algún rincón de una gaveta o entre las carpetas digitales de una computadora, invisibles e inaccesibles para otros investigadores.

Claerbout y Karrenbach (1992) mencionan que una investigación científica no es solo el producto visible, que es el artículo o la presentación en un evento, sino que incluye todos los elementos del contexto asociado: datos, códigos, métodos y programas utilizados. La visión que solo da valor a los artículos es superficial y esconde la realidad de que los datos son uno de los productos primarios más valiosos de la ciencia (Martone, 2014). Esta visión ignora sus múltiples valores adicionales que permiten sustentar la tesis de que, incluso, pueden llegar ser más importantes que las propias publicaciones. Entre las razones que permiten fundamentar esto están las siguientes:

- Los datos garantizan la reproducibilidad de los estudios publicados.
- Los datos son pólizas de seguro contra sospechas de fraudes o de conductas inadecuadas (Fiedler y Schwarz, 2016).
- Un conjunto de datos puede ser analizado por diferentes vías, distintas al método utilizado en la publicación, para corroborar los resultados por triangulación (Munafo y Smith, 2018).
- Los datos de un estudio pueden sustentar o formar parte de otros estudios futuros, incluso algunos para los que no se habían concebido desde un inicio. Un mismo conjunto de variables puede servir para probar aspectos ecológicos distintos, nutriendo así la ciencia del futuro (Editorial Nature-Communications, 2018; Specht *et al.*, 2018).
- Los datos no envejecen como las publicaciones que, con el tiempo, pierden valor relativo ya sea por otros descubrimientos, por la aparición de nuevos métodos analíticos o por el propio proceso de evolución de la Ciencia. Los datos, sin embargo, siempre mantienen su valor explicativo para las condiciones, método, lugar y momento en que fueron tomados.
- Los datos permiten el desarrollo de meta-análisis integradores que vencen las limitaciones logísticas para grandes tamaños de muestra en un único estudio (Crowther *et al.*, 2010).
- Los datos de múltiples fuentes permiten el desarrollo y comprobación de teorías integradoras, que no se pueden fundamentar solo con datos locales.
- Las series de datos a largo plazo permiten descubrir y comprender patrones ecológicos y evolutivos complejos, en amplias escalas temporales y espaciales (Kuebbing *et al.*, 2018).
- Los datos, si son compartidos públicamente, pueden aumentar la visibilidad de los artículos originales e incrementar sus tasas de citación (Piwowar *et al.*, 2007)

De estos aspectos, el primero es fundamental. La validez epistemológica de los resultados científicos se garantiza por las bases metodológicas de la ciencia (Pigliucci y Boudry, 2013). Un aspecto esencial del método científico es la repetibilidad: un estudio se considera válido si es repetible. El resultado de una investigación se asume como verdadero si, al ser repetida por otros investigadores, se llegan a las mismas conclusiones. Sin embargo, en la actualidad la gran mayoría de los estudios científicos no pueden ser repetidos, por diversas causas (Baker, 2016). Ante la imposibilidad logística de volver a hacer la toma de datos de las investigaciones publicadas para comprobarlas, una alternativa que permite evaluar la calidad de un estudio es su reproducibilidad: la posibilidad de verificar los resultados repitiendo los análisis expuestos, con los mismos datos que fundamentaron la publicación. Pero esto implica que es esencial que esos datos estén disponibles (Rodríguez-Sánchez *et al.*, 2016). Los problemas asociados a estos aspectos de replicabilidad y reproducibilidad han sido identificados como una de las crisis principales que enfrenta la Ciencia en la actualidad (Peng, 2011), y la Ecología no es ajena a ella (Schnitzer y Carson, 2016).

Por desgracia, en el contexto actual, el número de casos identificados de conductas inadecuadas y fraudes científicos ha llevado a un estado de desconfianza generalizada (Kornfeld y Titus, 2016), y las dificultades en reproducibilidad, robustez y generalización de los resultados se atribuye a la simple visión de los autores como descuidados, sesgados o desconfiables, sin considerar los múltiples factores sinérgicos que pueden conducir a ello. Por estas razones, garantizar la reproducibilidad a partir de compartir los datos y códigos empleados en un estudio, es una garantía de transparencia y calidad (Piwowar *et al.,* 2007; Vandewalle, 2012). Los principales criterios de reproducibilidad se enfocan en los datos y sus características: disponibilidad, visibilidad, validación, formato, metadatos, seguridad y licencias. Esto ha llevado a la corriente actual que se conoce como "Ciencia Abierta" (*Open Science*), cuyo principio central es la total transparencia del trabajo científico donde el acceso a los datos desempeña un papel fundamental. Por tanto, se hace preciso un cambio en el tratamiento de los datos desde la visión tradicional que se enfoca solo en el análisis que se hace de ellos hacia una donde los investigadores ponen más énfasis en su organización, documentación y procesado para hacerlos públicos, conjuntamente con los artículos.

Existe una serie de pasos y condiciones de estandarización a cumplir para otorgar a los datos el protagonismo que merecen, como recursos de identidad propia, publicables y citables. Para esto existen diferentes recomendaciones y buenas prácticas que aportan consejos sobre la gestión de los datos y cómo compartirlos, sobre todo en el caso de la Ecología (Kervin *et al.,* 2013; Goodman *et al.*, 2014; Michener, 2015a). Sin embargo, a pesar de haberse reconocido las ventajas de hacer abiertos los datos esta práctica no se ha incorporado plenamente a la actividad cotidiana (Culina *et al.,* 2018).

La preservación de los datos y los esfuerzos por hacerlos publicables, almacenables y reusables en repositorios oficiales aumenta el valor de la información que se genera de ellos debido a los procesos de control de calidad a los que son sometidos. Esto repercute en una serie de beneficios, no solo para el propio investigador, sino para la comunidad científica en general (Specht *et al.,* 2018). Por estas razones, en la presenta comunicación se profundiza en algunas de las prácticas que permiten dar valor a los datos de las investigaciones ecológicas para explotarlos en toda su extensión, ya que hay todavía desconocimiento entre la comunidad de investigadores de temas ambientales sobre la infraestructura existente y los métodos a seguir para que un científico pueda, por tiempo indefinido, hacer disponibles, recuperables y utilizables los datos de sus artículos. Para ello se responderán las preguntas más básicas en relación a este tema: ¿por qué y para qué? ¿cómo? ¿dónde? y ¿cuándo?

DESARROLLO

¿Por qué y para qué?

Estas dos preguntas están muy relacionadas y se enfocan no solo en las ventajas ya mencionadas que para la ciencia brinda el manejo adecuado y la publicación de los datos de las investigaciones, sino en los beneficios que trae directamente para los investigadores que lo hacen. Las ventajas de compartir datos están bien identificadas y concretadas (e.g., Whitlock, 2011; Torres-Salinas *et al.,* 2012; Hampton *et al.,* 2015; Soranno, 2019).

Entre los deseos que son comunes a todos los científicos están el poder acceder a fondos para hacer sus investigaciones, poder publicarlas en revistas del mayor impacto posible y obtener reconocimiento profesional por ello. En este sentido, cada vez más se está solicitando la inclusión de un buen Plan de Manejo de Datos en los proyectos de solicitud de financiamiento. Esto ya es un requisito, por ejemplo, para los proyectos presentados a la Fundación Nacional para la Ciencia (*National Science Foundation*) de Estados Unidos (Michener y Jones, 2012; Michener, 2015b). Un Plan de Manejo de Datos es un documento que detalla cómo se tratarán los datos durante el proyecto y qué sucederá con ellos una vez se termine. Incluye todos los pasos del ciclo de vida de los datos: desde el descubrimiento, recolecta, organización, control de calidad, documentación, utilización, preservación y publicación (Michener, 2015). Pero incluso sin llegar a un plan formal, la solicitud de inclusión de criterios de disponibilidad de los datos se encuentra entre los aspectos actualmente solicitados, tanto para proyectos como para la publicación en revistas importantes (Stodden *et al.,* 2013).

La evolución en las políticas de las revistas científicas muestra una tendencia general hacia una mayor disponibilidad de datos (Stodden, 2013), no solo por su valor *per se*, sino como medida para fomentar la reproducibilidad. La editorial del volumen 515 de

Nature (2014) fue titulada: "*Journal Suniteforre Producibility*". Esto incluye el acceso a los datos, ya sea para la revisión o de forma ilimitada, y la adición a cada artículo de un "*data availability statement*". La revista *Molecular Ecology* plantea en las instrucciones a los autores de forma explícita que "los artículos con archivado de datos y código son más valiosos para investigaciones futuras, por lo que, a igualdad de condiciones, se les dará mayor prioridad para su publicación". Requerimientos similares se implementaron en las revistas de la *American Association for the Advancement of Science* (AAAS) y *Science Translational Medicine* desde 2013 y en *Science* desde 2014 (McNutt, 2014). Un paso importante se dio cuando varias revistas claves en el campo de la evolución y la ecología adoptaron la política llamada *Joint Data Archiving Policy*, la cual introducía como requisito indispensable para la publicación del artículo la accesibilidad a los datos que sus tentaban la investigación (Whitlock, 2011). O sea, para publicar en revistas importantes el manejo adecuado y la apertura de los datos es, sino indispensable, una ventaja para los autores que lo hacen.

En relación al reconocimiento de los autores que publican sus datos, ya desde hace años, la *National Science Foundation* expandió su guía de méritos para incluir mayor rango de productos científicos como los códigos, programas y datos, y no solo evaluar a los científicos por las publicaciones arbitradas. Se ha demostrado que compartir los datos abiertamente promueve la visibilidad y aumenta el impacto personal, así como las tasas de citación de los autores (Piwowar *et al.*, 2007; Vandewalle, 2012). Como parte de una iniciativa para la generación y empleo de *Big Data* se ha creado un *Data Discovery Index* (DDI) para permitir la búsqueda y acceso a datos primarios, con lo cual surge una nueva métrica de uso de un artículo que mide su impacto sin depender de las citas que reciba, y es a través de las cantidades de descargas de los datos asociados. Todas estas políticas están conduciendo a una mayor generalización de la ciencia abierta y con ello aumenta la disponibilidad de datos, con lo cual se acelera el progreso científico.

¿Dónde?

Los lugares para publicar o almacenar datos científicos se han ido desarrollando progresivamente, no solo para hacerlos abiertos sino incluso para mantener copias privadas con mayor seguridad. Vines *et al.* (2014) estimaron que la disponibilidad de los datos tiende a reducirse en el tiempo a una tasa anual del 17%. En relación a su publicación adjuntos a los artículos, suelen incluirse en los "Materiales Suplementarios", la versión modernizada de los "Anexos", y que son cada vez más comunes en los artículos científicos.

Sin embargo, una publicación eficiente de los datos va más allá de ponerlos como material suplementario en los propios sitios de las revistas, ya que estos son insuficientes en términos de integración y recuperabilidad. Si la información está ligada al sitio de una revista específica no se integra automáticamente con otros conjuntos para sustentar un conocimiento científico global sobre un tema (Michener, 2015). Los datos incluidos como forma de material suplementario también adolecen de otras características que limitan su uso, como la alta variedad de formatos utilizada por los autores, la falta de estándares que armonicen distintos conjuntos de datos y la escasez de herramientas apropiadas de bús queda que posibiliten su localización (Santos *et al.,* 2005). En relación a este último aspecto, sin embargo, un paso importante ha sido la aparición del buscador de datos de *Google* (*Google Data Search*) que se enfoca en indexar y hacer descubribles estos elementos científicos. Por ello, publicar los datos como materiales suplementarios, aunque es válido, no debe sustituir a su depósito en un repositorio científico público. De estos existe una alta variedad y la gran mayoría son gratuitos, porque precisamente abogan por la apertura libre de los datos.

Existen desde repositorios generalistas que almacenan datos de diversas ramas científicas, como *Dryad* (https://datadryad.org) o *FigShare* (http://www.figshare.org) hasta otros más especializados en ciertos tipos de datos como GBIF (https://www.gbif.org) para datos primarios de biodiversidad o *GeneBank* (http://www.ncbi.nlm.nih.gov/genbank/) para secuencias genéticas. *Figshare*, por ejemplo, garantiza el servicio de publicación y conservación de datos bajo una licencia *Creative Commons*, de forma ilimitada siempre y cuando sean de uso público, dando beneficios claves como la seguridad, persistencia, visualizaciones, indexaciones, asignación gratuita de identificadoras digitales (DOI) y el seguimiento del impacto con estadísticas de visitas y uso (*altmetrics*). En el campo de la Biología vegetal, destaca la *Plant Trait Database TRY* (http://www.try-db.org), que desde su origen en 2007 hasta la actualidad, ha logrado recopilar y brinda libremente el acceso a 10 billones de registros sobre rasgos vegetales, con lo cual se ha descrito como el archivo más exhaustivo, disponible, que existe sobre este tipo de dato de plantas (Fraser, 2020).

Estos repositorios son la mejor forma que existe en la actualidad de asegurar la persistencia de los datos a largo plazo (White *et al.,* 2013; Hart *et al.,* 2016). Existe un registro internacional de repositorios de datos de investigación (http://www.re3data.org) y herramientas para ayudar al investigador a identificar el repositorio más apropiado para ubicar sus datos, como *Open DOAR* (http://v2.sherpa.ac.uk/opendoar/) y *Repository Finder* (https://repositoryfinder.datacite.org/). También hay revistas que proporcionan en sus políticas de datos, listados de repositorios recomendados.

¿Cómo?

El "cómo" lograr extraer el valor pleno de los conjuntos de datos incluye varios elementos relacionados, como la estructuración apropiada de las matrices o ficheros, su documentación a través de metadatos, la selección y asignación de las licencias y la identificadora digital (DOI), estas dos últimas para proteger el derecho de autor. Una variante válida que también está cobrando fuerza es su publicación en forma de Artículo de datos (*Data Papers*).

El primer paso siempre es el control de la calidad de los datos, que lamentablemente es frecuentemente obviado. Este incluye un conjunto de buenas prácticas en la conformación de las matrices y bases de datos (Wickham, 2014) que resulta crucial para facilitar su análisis posterior, ya que cualquier error en esta etapa se propagará hasta los resultados finales de las investigaciones (Michener y Jones, 2012). El sitio *Data Horror Histories* (https://www.lcrdm.nl/horror) publica una serie de ejemplos de sucesos reales relacionados con desastres en investigaciones producto del mal manejo de los datos.

Las dos prácticas asociadas más recomendadas para asegurar esta calidad son: el control de versiones (*data-versioning*) y las prácticas de cuidado de datos (*data-care*). Estas dos se encuentran entre las 10 principales cosas que Carly Strasser (2014), en su blog en el UC3 (*University of California Curation Center*), compilaba como las más importantes actualmente para los jóvenes investigadores, como pasos para promover el cambio hacia una ciencia más abierta y transparente. Esto trae a colación la implementación y manejo de sistemas para el control de versiones. Por la importancia que tienen los datos, ninguna medida es poca para protegerlos y ello incluye mantener versiones previas ante cualquier modificación de importancia que permita en caso de error recuperar la matriz previa. La opción de deshacer no es válida a largo plazo. El sistema más universal de control de versiones, problemático y poco recomendado, consiste en guardar copias de los ficheros con distintos nombres. Para ayudar a los investigadores se han desarrollado programas automatizados que se encargan de monitorizar los cambios realizados en cualquier fichero, registrando quién hizo qué cambio, cuándo y por qué (Blischak *et al.,* 2016). Los sistemas más conocidos son el *GitHub*, *LabHub*, *CVS*, *Vesta*, *Mercurial* o *Subversion*. Otras herramientas como *Google Docs*, *Dropbox*, *Overleaf* o *Authorea* representan un buen avance al respecto, pero están más enfocadas a la escritura colaborativa de manuscritos que a las matrices de datos primarios.

Para mantener los datos y asegurar su valor de uso, la primera y más universal regla de la organización de datos es la consistencia. La consistencia debe mantenerse en:

- Nombres y códigos para todas las variables: a lo largo de una investigación, en todas las matrices y ficheros las variables deben mantener un mismo nombre, escrito exactamente igual, así como los códigos en su contenido.
- El símbolo para identificar valores faltantes, que debe ser único (los que emplean R suelen usar NA) y evitarse usar números para ello (como los -9999 que utiliza el ESRI ArcGis).
- El tipo de dato de la variable: en una misma columna nunca se deben incluir celdas con números y otras con letras, simultáneamente.
- Los identificadores de casos: si se refiere al mismo individuo, no deben aparecer en una matriz como "153" y en otra como "individuo153", o como "ind-153F" o "I-153".
- En el formato y orden de los datos en múltiples ficheros: las columnas deben mantener el mismo orden entre ficheros para facilitar la combinación o intercambio de datos.
- Los nombres de ficheros: debe seguirse un esquema consecutivo para nombrarlos, sobre todo que siempre mantengan un mismo orden en el explorador de ficheros.
- En los formatos de horas y fechas: si se emplea el formato de fecha día-mes-año no usar en otros datos mes-día-año, o si las horas son en formato militar (24H) o formato de 12h.

- En las frases o notas aclaratorias: Si hay variables de notas usar las mismas frases y cuidar las palabras exactamente escribirlas igual (no abreviar unas y otras no, o usar unas con mayúsculas y otras no).

Todas estas regularidades conllevan a una propiedad muy importante en la actualidad: los datos mantienen un lenguaje que permite que sean leídos sin errores, tanto por seres humanos como por programas de computación.

Una buena costumbre debe ser la de crear un diccionario de los datos, que es simplemente emplear la primera hoja del *Excel* de datos o el primer fichero de la serie, para anotar las características de los datos: nombres de variables, códigos empleados… de forma tal que si en un momento no se recuerda como se ha estado poniendo se busque ahí. Esto aumenta la eficiencia del uso de los datos y permite que otros los puedan entender con facilidad también. Siempre debe pensarse que los datos pueden terminar siendo usados para otros propósitos y no solo para el que fueron obtenidos inicialmente (Hampton *et al.,* 2015; Díaz-Delgado, 2016) y que deben hacerse de manera que sean leíbles y comprensibles por un programa de computación y no solo a través de la visión e interpretación humana.

Este diccionario de datos sería una forma de metadato, que se refiere a toda aquella información que describe el conjunto de datos y resulta necesaria para su correcta interpretación (Michener *et al.,* 1997). Los metadatos multiplican la utilidad de los datos al favorecer su reutilización (Fegraus *et al.,* 2005; Alonso y Valladares, 2006; Rüegg *et al.,* 2014). El depósito de los datos en un sitio público no es suficiente para asegurar su reutilización (Roche *et al.,* 2015) ya que si carecen de una adecuada documentación, se obstaculiza su re-uso o evaluación crítica (Goodman *et al*., 2014).

Aunque los metadatos pueden ser tan simples como un fichero de texto o la hoja inicial del *Excel*, es más conveniente utilizar un sistema estándar que facilite la validación, integración y síntesis de los datos de manera automatizada (Alonso y Valladares, 2006). Para facilitar el intercambio, la integración y síntesis de los datos se han desarrollado diversos estándares de metadatos, en función del propósito y la disciplina científica. En Ecología existe el estándar llamado *Ecological Metadata Language* (EML) que se utiliza, por ejemplo, en las redes de seguimiento ecológico a largo plazo (LTER, *Long-Term Ecological Research*; Vanderbilt *et al.,* 2015), el *Darwin Core* (Wieczorek *et al.,* 2012) e INSPIRE, para datos espaciales. El empleo de estos formatos internacionales garantiza una integración a los sistemas y bases globales del conocimiento.

Existen varias herramientas para crear o editar metadatos. Son conocidos el *Morpho* (http://knb.ecoinformatics.org/morphoportal.jsp) y DEIMS (https://data.lter-europe.net/deims/). En el entorno de programación R han aparecido varios paquetes con este objetivo como el *eml* (Boettiger *et al.,* 2019) y *emld* (Boettiger, 2019) para la creación de metadatos, y el paquete *EMLAssemblyLine* (Smith, 2019) que incorpora además un flujo de trabajo. Asimismo, para facilitar la tarea a los investigadores, algunas iniciativas como la *Environmental Data Initiative*, ofrecen plantillas para generar los metadatos de un conjunto de datos de acuerdo al estándar EML. Esta documentación completa de los datos implica un cambio en las prácticas de los investigadores, poniendo énfasis en su posible reutilización y concibiéndolos como un elemento clave en la producción de conocimiento, sin limitarse a la publicación de un artículo científico.

Unos conjuntos de estos requerimientos de los datos se resumieron en los llamados criterios *FAIR data*, establecidos en 2016 para una gestión eficaz de los datos (Wilkinson *et al.,* 2016). *FAIR*, en inglés significa justo, apropiado o imparcial, pero en este caso es el acrónimo de *Findables* (encontrables), *Accesible* (asequibles), *Interoperative* (inter-operativos) y *Reusable* (reusables). Que los datos sean "encontrables" se refiere a que los datos y los metadatos puedan ser descubiertos y recuperados por la comunidad científica de forma automatizada. Un conjunto de datos no se considera disponible solo con el enunciado público de la disposición de los autores o tenedores de ofrecérselos a quienes lo soliciten personalmente. Para que sean realmente abiertos o disponibles se requiere que estén identifica dos de forma única y persistente, que estén descritos en detalle y que estén indexados. El identificador único, global y eternamente persistente, más recomendado es el DOI (*Digital Object Identificator*). Para que los datos sean accesibles, una vez encontrados, es necesario que se publiquen utilizando protocolos estandarizados. La utilización de estándares de inter cambio permite que los datos sean interoperables, y su estructuración y documentación asegura que sean reutilizables, al igual que una licencia de uso clara y accesible, como las *Creative Commons*.

Más allá de publicar los datos con sus metada tos en un repositorio científico online puede valorarse la opción de escribir un artículo de datos (*Data Paper)* sobre ese conjunto de datos. Los *Data Paper* son una modalidad de publicación científica de reciente origen, en la cual se documenta detalladamente uno o varios conjun tos de datos accesibles, describiendo el contexto en el que fueron generados y su contenido. Estas publicaciones no siguen el esquema típico de un artículo científico, ya que su propósito no es exponer un resultado sino describir conjuntos de datos de forma entendible y estructurada para divulgarlos y facilitar su uso. Como cualquier otro artículo científico, es una publicación que da visibilidad al contenido y reconocimiento académico a sus autores (Chavan y Penev, 2011). Al ser sometidos a revisión por pares, igual que cualquier otro artículo de investigación actúa como una garantía de accesibilidad y calidad de los datos y metadatos (Costello *et al.,* 2013). En los últimos años, ha aumentado el número de revistas que incluyen esta modalidad de publicación o que se dedican exclusivamente a ellos (*data journals*) como la revista *Biodiversity data Journal.*

¿Cuándo?

El momento adecuado para publicar los datos es una decisión del investigador. La tendencia usual es mantener los datos en repositorios científicos desde el inicio, pero en modo privado hasta que son publicados los artículos para los cuales fueron tomados. En algunos casos, puede aplicarse un tiempo de embargo posterior a la publicación que retrasa su apertura, para dar tiempo a la salida de otros artículos relacionados. Pero potencialmente pueden ser publicados desde el momento de la culminación de su toma y manejo, ya que, al estar descritos, referenciados y protegidos bajo una licencia de uso, actúan como un pre-registro de la investigación (Nosek *et al*., 2018) y pueden generar expectativas que aumenten la tasa de citación y el grado de inmediatez de la publicación cuando salga.

Retos y perspectivas

En la mayoría de las publicaciones de Ecología durante toda la historia de esta ciencia no se han tenido en cuenta los aspectos relacionados con la replicabilidad, reproducibilidad o la publicación de datos. La gran mayoría de los estudios de biodiversidad no son replicables ni reproducibles, y en muchos casos los datos se han perdido o no están disponibles, y la probabilidad de acceso a estas fuentes primarias disminuye exponencialmente con el tiempo transcurrido desde la publicación (Vines *et al.,* 2014). Ello contribuye de forma negativa al desarrollo de las ciencias ambientales, por la pérdida de posibilidades que acarrea y el nivel de incertidumbre o desconfianza que se ha generado sobre la validez de lo ya publicado (Fraser *et al*., 2018).

A niveles nacionales e institucionales debieran desarrollarse políticas explícitas y sólidas en relación al manejo de los datos, que se extienda no solo a los proyectos de investigación sino a las propias agencias de financiamiento, centros académicos o de investigación, revistas y editoriales científicas. Sin embargo, esto no es tan sencillo y las medidas impuestas no siempre conducen a resultados positivos. Esto ratifica los dos retos que Reichman *et al.* (2011) señalan para la Ciencia Abierta: el tecnológico y el social. El primero se relaciona con la propia complejidad inherente a la información ecológica, y el segundo con las reticencias sociales presentes a la hora de compartir.

Si bien la opinión generalizada, es la de reconocer racionalmente la importancia de la apertura, la realidad es que son muchos los científicos que aún se niegan a hacerlo. Según Pinowar (2011), en estudios de expresión genética, los autores con más tendencias a compartir sus datos son aquellos con experiencias previas en el empleo de datos compartidos, los que tienden a publicar en revistas de acceso abierto o con políticas sólidas hacia la apertura de datos o los que tuvieron, con mayor frecuencia, financiamientos del Instituto Nacional de Salud de EUA (NIH), que estipula la publicación de los datos de forma obligatoria. Investigaciones previas encontraron asociación entre el género del investigador y la frecuencia de compartir datos (Blumenthal *et al.,* 1997). Otros estudios han identificados como más proclives a compartir datos a investigadores de mediana experiencia o a jóvenes *millenials* que han desarrollado su personalidad en un entorno de apertura digital generalizada, promovida por las redes sociales y para los cuales los conceptos de *altmetrics* son naturalmente aceptados.

El rechazo a compartir datos puede derivarse de un malentendido sentido de propiedad sobre estos por parte de investigadores o grupos de investigación (Caldera-Serrano, 2018), que es paradójico cuando se trata de investigaciones financiadas con fondos institucionales. La competitividad y el deseo de reservarse futuras oportunidades de publicación con los datos también han sido señaladas como causas sociales para no compartir datos (Couture *et al.,* 2018). A estas se suman, la ausencia de presión institucional, la

carencia de recursos y la poca experiencia en la gestión de datos (Michener, 2015a).

La comunidad científica de las ciencias ambientales muchas veces no domina el hecho de que compartir los datos públicamente no implica la pérdida de derechos, sino que se acoge a las nuevas licencias *Creative commons* (con las cuales todo investigador actual debe estar familiarizado) por lo que cualquier uso de ellos debe reconocer públicamente su fuente y eso redunda en más reconocimiento para el propio investigador. Para ello, más que todo, se necesita de un cambio fundamental en el modo de pensar de los investigadores desde el "yo soy el dueño de los datos" a "yo recopilo los datos en nombre de la ciencia y a favor de la sociedad" (Hampton *et al.,* 2015). Sin embargo, este cambio de mentalidad tiene que estar acompañado de un mayor reconocimiento en el mundo académico e institucional al hecho de compartir los datos. Es decir, según Michener (2015a), se necesita un cambio en la cultura de la ciencia hacia una en la que se valoren por igual datos y publicaciones, para lograr que las reutilizaciones de los datos tengan también un im pacto positivo que pueda ser percibido por quienes los generaron (Pierce *et al.,* 2019). Así, por ejemplo, los indicadores que se emplean para evaluar el trabajo de un científico podrían considerar, además de los artículos primarios producidos en un año, los datos que han sido aportados a la ciencia global. Además del factor de impacto de la revista, el número de citas recibidas o el volumen de resultados, las publicaciones pueden evaluarse también por el nivel de apertura o transparencia y por la frecuencia de citación o uso de los datos. Martone (2014) encabeza un amplio grupo de científicos que promovieron y firmaron la Declaración Conjunta de Principios para la Citación de Datos, estableciendo los principios esenciales para poder universalizar esta práctica.

CONSIDERACIONES FINALES

La evolución hacia una ciencia abierta y transparente, potenciada por un mayor intercambio de datos, debe asociarse tanto a cambios prácticos en la forma en que se desarrollan las investigaciones como a cambios conceptuales en relación al papel que desempeñan los datos en la produc ción de conocimiento científico. Estos cambios sociales requieren del compromiso de todos los actores implicados, cuatro de los cuales pueden ser claves: las revistas y editoriales, las universidades, las entidades financiadoras y los propios investigadores. Se ha mencionado que, sobre todo, los financistas y las direcciones científicas institucionales tienen más poder para exigir el intercambio de datos a los investigadores puesto que, en cierto modo, comparten esa "propiedad" sobre ellos (Couture *et al.,* 2018). Pero las revistas también tienen un alto protagonismo en la evaluación y difusión de los resultados científicos (Caldera-Serrano, 2018) y pueden apoyar adoptando y haciendo cumplir políticas de datos que promuevan la publicación abierta de los datos e, incluso, fomentando la creación de artículos de datos (Michener, 2015), hacia lo cual debían moverse las revistas cubanas. Las políticas editoriales e instrucciones para autores deberían incluir directrices claras que promuevan y recompensen la publicación de los datos en repositorios internacionales referidos desde la información suplementaria, así como asignar responsabilidades específicas en el cuerpo editorial para la revisión y aseguramiento de la calidad de los datos asociados a las publicaciones (Sholler *et al.,* 2019).

Pero otro aspecto imprescindible es la formación, ya que muchos investiga dores pueden estar dispuestos y animados a compartir sus datos, pero los frenan la falta de experiencia o conocimiento de las vías para hacerlo. Las universidades deben crear espacios formativos, tanto de postgrado como de pregrado que llenen las necesidades relativas a las habilidades técnicas y teóricas necesarias para el manejo y puesta en valor de los datos, de forma paralela a los métodos para hacer uso de ellos.

El mayor reto está en la mano de los investigadores más jóvenes que tienen que familiarizarse con nuevos conceptos (como los metadatos) o herramientas no tradicionales para el ecólogo (como las bases de datos, la programación en R, los sistemas de control de versiones) para lo cual existen numerosos recursos en internet que hacen posible la preparación autónoma. Soranno (2019) menciona cómo los investigadores pueden incluir seis sencillos pasos en su flujo de trabajo para com partir los datos asociados a los artículos: 1) decidir de forma separada la autoría de los datos y la del artículo de investigación, 2) utilizar formatos simples para organizar sus datos primarios, 3) escribir los metadatos, 4) de positar ambos, datos y metadatos, en repositorios, 5) incluir una declaración de la disponibilidad de datos en sus artículos y 6) citar sus propios datos de los repositorios en la sección de Métodos e incluir la referencia en la Bibliografía. Además, los investigadores tienen que apostar también por productos como los *Data Papers*.

Todos estos esfuerzos llevan una inversión considerable, que pudiera parecer una pérdida de tiempo al principio, pero tanto adoptar flujos reproducibles como aprender a compartir los datos, puede ahorrar muchos problemas en el futuro y nos pueden poner a la altura de cualquier investigador de primer nivel del mundo desarrollado. Cuba representa el mayor núcleo de tierras emergidas del área del Caribe, uno de los más importantes núcleos de biodiversidad (*hot spots*) de biodiversidad a nivel mundial. Los datos primarios procedentes de las investigaciones en nuestro país pueden marcar una diferencia en los estudios regionales y globales hechos por muchos investigadores en cualquier lugar del mundo, beneficiando en primer lugar a la Ciencia y a la conservación de la naturaleza.

AGRADECIMIENTOS

Se agradece los comentarios realizados por Jorge Alberto Sánchez Rendón y las correcciones sugeridas por dos árbitros anónimos, que contribuyeron significativamente a la mejora en el manuscrito inicial.

LITERATURA CITADA

Alonso B, Valladares F. 2006. Bases de datos y metadatos en ecología: compartir para investigar en cambio global. *Ecosistemas*. 15(2): 83 88.

Baker M. 2016. Is there a reproducibility crisis? *Nature*. 533:452-454.

Blischak JD, Davenport ER, Wilson G. 2016. A quick introduction to version control with Git and GitHub. *PLOS Computational Biology*. 12: e1004668.

Blumenthal D, Campbell E, Anderson M, Causino N, Louis K. 1997. Withholding research results in academic life science. Evidence from a national survey of faculty. *Journal of the American Medical Association*. 277: 1224-1228.

Boettiger C. 2019. Ecological metadata as linked data. *Journal of Open Source Software*. 4: 1276.

Boettiger C, Jones MB, Maier M, Mecum B, Salmon M, Clark J. 2019. *EML: Read and Write Ecological Metadata Language Files*. Disponible en https://cloud.r-project.org/web/packages/EML/index.html. (consultado: 01 de septiembre de 2020).

Caldera-Serrano J. 2018. Repositorios públicos frente a la mercantilización de la ciencia: apostando por la ciencia abierta y la evaluación cualitativa. *Métodos de Información*. 9: 74-101.

Chavan V, Penev L. 2011. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*. 12 Suppl 15: S2.

Claerbout J, Karrenbach M. 1992. Electronic documents give reproducible research a new meaning. En: Proceedings of the 62nd annual international meeting of the society of exploration geophysics, New Orleans (Octubre 1992).

Costello MJ, Michener WK, Gahegan M, Zhang Z-Q, Bourne PE. 2013. Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology and Evolution*. 28: 454-461.

Couture JL, Blake RE, McDonald G, Ward CL. 2018. A funder-imposed data publication requirement seldom inspired data sharing. *PLOS ONE*. 13: e0199789.

Crowther M, Lim W, y Crowther MA. 2010. Systematic review and meta-analysis methodology. *Blood, The Journal of the American Society of Hematology*. 116: 3140-3146.

Culina A., Baglioni M., Crowther TW, Visser ME, Woutersen-Windhouwer S, Manghi P 2018. Navigating the unfolding open data land scape in ecology and evolution. *Nature Ecology and Evolution*. 2: 420-426.

Díaz-Delgado R. 2016. La investigación y seguimiento ecológico a largo plazo (LTER). *Ecosistemas*. 25: 1-3.

Editorial Nature. 2018. Data sharing and the future of science. *Nature Communications*. 9: 2817.

Fegraus EH, Andelman S, Jones MB, Schildhauer M. 2005. Maximiz ing the value of ecological data with structured metadata: An introduction to ecological metadata language (EML) and principles for metadata creation. *The Bulletin of the Ecological Society of America*. 86: 158-168.

Fiedler K y Schwarz N. 2016. Questionable research practices revisited. *Social Psychological and Personality Science*. 7: 45-52.

Fraser LH. 2020. TRY-A plant trait database of databases. *Global Change Biology*. 26:189-190.

Fraser, H., T. Parker, S. Nakagawa, A. Barnett y F. Fidler. 2018. Questionable research practices in ecology and evolution. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0200303.

Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, Di Stefano R, Gil Y, Groth P, Hedstrom M, Hogg DW, Kashyap V, Mahabal A, Siemiginowska A, Slavkovic A. 2014. Ten simple rules for the care and feeding of scientific data. *PLOS Computational Biology*. 10: 1-5.

Hampton SE, Anderson SS, Bagby SC, Gries C, Han X, Hart EM, Jones MB, Lenhardt WC, Macdonald A, Michener WK, Mudge J, Pourmokhtarian A, Schildhauer MP, Woo KH, Zimmerman N. 2015. The Tao of open science for ecology. *Ecosphere*. 6: art120.

Hart E, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, Poisot T, Woo KH, Zimmerman NB, Hollister JW. 2016. Ten simple rules for digital data storage. *Peer J preprints*. 4: e1448v2.

Kervin K, Michener W, Cook R. 2013. Common errors in ecological data sharing. *Journal of Science Librarianship*. 2: e1024.

Kornfeld DS, Titus SL. 2016. Stop ignoring misconduct. *Nature*. 537: 29

Kuebbing SE, Reimer AP, Rosenthal SA, Feinberg G, Leiserowitz A, Lau JA, Bradford MA. 2018. Long-term research in ecology and evolution: A survey of challenges and opportunities. *Ecological Mono graphs*. 88: 245-258.

Martone M. (ed.) 2014. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. San Diego CA: FORCE11. (https://www.force11.org/group/joint-declaration-data-citation-principles-final).

McNutt M. 2014. Reproducibility. *Science* 343, 229.

Michener WK. 2015a. Ecological data sharing. *Ecological Informatics*. 29: 33-44.

Michener WK. 2015b. Ten simple rules for creating a good data management plan. *PLoS Computational Biology*. 11(10): e1004525. doi:10.1371/journal.pcbi.1004525.

Michener WK, Brunt JW, Helly JJ, Kirchner TB, Stafford SG. 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications*. 7: 330-342.

Michener WK, Jones MB. 2012. Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*. 27: 85-93.

Munafo MR, Smith GD. 2018. Repeating experiments is not enough. *Nature* . 553: 399

Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences*. 115: 2600-2606.

Peng RD. 2011. Reproducible Research in Computational Science. *Science*. 334: 1226-1227.

Pierce HH, Dev A, Statham E, Bierer BE. 2019. Credit data generators for data reuse. *Nature* . 570: 30-32.

Pigliucci M, Boudry M. 2013. *Philosophy of pseudoscience: reconsidering the demarcation problem*. University of Chicago Press, Chicago.

Piwowar HA, Day RS, Fridsma DB. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE*. 2: e308.

Piwowar HA. 2011. Who shares? Who doesn't? Factors associated with openly archiving raw research data. *PLoS ONE*. 6(7): e18657.

Reichman OJ, Jones MB, Schildhauer MP. 2011. Challenges and op portunities of open data in ecology. *Science* . 331: 703-705.

Roche DG, Kruuk LEB, Lanfear R, Binning SA. 2015. Public data archiving in ecology and evolution: How well are we doing? *PLOS Biology*.13: e1002295.

Rodríguez-Sánchez F, Pérez-Luque AJ, Bartomeus I, Varela S. 2016. Ciencia reproducible: qué, por qué, cómo? *Ecosistemas*. 25: 83-92.

Rüegg J, Gries C, Bond-Lamberty B, Bowen GJ, Felzer BS, McIntyre NE, Soranno PA, Vanderbilt KL, Weathers KC. 2014. Completing the data life cycle: using information management in macrosystems ecology research. *Frontiers in Ecology and the Environment*. 12: 24-30.

Santos C, Blake J, States DJ. 2005. Supplementary data need to be kept in public repositories. *Nature*. 438: 738.

Schnitzer SA, Carson WP. 2016. Would Ecology Fail the repeatability test? *Bioscience*. 66: 98-99.

Sholler D, Ram K, Boettiger C, Katz DS. 2019. Enforcing public data archiving policies in academic publishing: A study of ecology journals. *Big Data and Society*. 6: 1-18.

Smith C. 2019. *EMLassemblyline: A workflow for creating* EML. Disponible en https://github.com/EDIorg/EMLassemblyline/blob/mas-ter/DESCRIPTION (consultado: 01 de septiembre de 2020).

Soranno PA. 2019. Six simple steps to share your data when publishing research articles. *Limnology and Oceanography Bulletin*. 28: 41-44.

Specht A, Bolton MP, Kingsford B, Specht RL, Belbin L. 2018. A story of data won, data lost and data re-found: the realities of ecological data preservation. *Biodiversity Data Journal*. 6: e28073.

Stodden V, Guo P, Ma Z. 2013. Toward reproducible computational research: An empirical analysis of data and code policy adoption by jour nals. *PLOS ONE*. 8: 1-8.

Strasser C. 2014. The 10 things every new grad student should do. Blog en UC3 - University of California Curation Center. Disponible en https://uc3.cdlib.org/2014/10/14/the-10-things-every-new-grad-student-should-do/ (consultado: 02 de junio de 2020).

Torres-Salinas D, Robinson-García N, Cabezas-Clavijo Á. 2012. Compartir los datos de investigación en ciencia: introducción al *data sharing*. *El Profesional de la Información*. 21: 173-184.

Vanderbilt KL, Lin C-C, Lu S-S, Kassim AR, He H, Guo X, San Gil I, Blankman D, Porter JH. 2015. Fostering ecological data sharing: collaborations in the International Long Term Ecological Research Network. *Ecosphere*. 6(10):204. DOI.org/10.1890/ES14-00281.1

Vandewalle P. 2012. Code sharing is associated with research impact in image processing. *Computing in Science and Engineering*.14: 42-47.

Vines TH, Albert AY, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore JS, Renaut S, Rennison DJ. 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology*. 24: 94-97.

White E, Baldridge E, Brym Z, Locey K, McGlinn D, Supp S. 2013. Nine simple ways to make it easier to (re)use your data. *Ideas in Ecology and Evolution*. 6: 1-10.

Whitlock MC. 2011. Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution* . 26: 61 -65.

Wickham H. 2014. Tidy data. *Journal of Statistical Software*. 59: 1-23

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D. 2012. Darwin core: An evolving community- developed biodiversity data standard. *PLoS ONE*. 7: e29715.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, Van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, Van Der Lei J, Van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*. 3: 1-9.