BIOCYT Biología, Ciencia y Tecnología, 12(47): 869-880, 2019 (julio-septiembre)
ISSN: 2007-2082

http://revistas.unam.mx/index.php/biocyt

Publicada en la Facultad de Estudios Superiores Iztacala, Universidad Nacional Autónoma de México

TENDENCIAS ACTUALES EN EL MANEJO DE DATOS DE INVESTIGACIÓN

ACTUAL TRENDS OF DATA RESEARCH MANAGEMENT

Layla Michán[1,1]✉ y Eduardo Álvarez[1,2]

[1]Laboratorio de Manejo de Información Biológica. Facultad de Ciencias, Universidad Nacional Autónoma de México, Av. Universidad No. 3000 Circuito Exterior S/N C.P. 04510, Ciudad Universitaria, Delegación Coyoacán, Ciudad de México, México.

[1,1]✉laylamichan@ciencias.unam.mx , [1,2]bioeduardo@ciencias.unam.mx

ABSTRACT

We all have, generate and use information inevitably, but scientists also use it as input and product, as a means of communication, as evidence, as an object of study and as an evaluation tool. At present, scientific information is digital, immense, diverse, complex and constantly evolving. It is generally systematized in digital collections, such as databases, repositories, indexes and catalogs; thus, managing information in an electronic environment is an indispensable skill for 21st century scientists. The generation, systematization, analysis and use of scientific information is ubiquitous, indispensable, strategic, and a current trend in biological sciences. The objective is to investigate which are the most relevant and innovative topics, services and software, entities and collections, thesauri and ontologies, which should be known to apply to manage the data product of research, especially in the biological area.

Keywords: databases, data curation, linked data, ontologies, semantic, thesaurus.

Manuscrito recibido el 15 de enero de 2019, aceptado el 24 de mayo de 2019.

RESUMEN
Todos tenemos, generamos y utilizamos información de manera inevitable, pero los científicos además la utilizan como insumo y producto, como medio de comunicación, como evidencia, como objeto de estudio y como herramienta de evaluación. En la actualidad, la información científica es digital, inmensa, diversa, compleja y evoluciona constantemente. Generalmente se encuentra sistematizada en colecciones digitales, tales como bases de datos, repositorios, índices y catálogos. De tal manera que, manejar la información en un entorno electrónico es una habilidad indispensable para los científicos del siglo XXI. La generación, sistematización, análisis y aprovechamiento de la información científica es ubicua, indispensable, estratégica y una tendencia actual en las ciencias biológicas. El objetivo en este artículo es investigar cuáles son los temas, servicios y software, entidades y colecciones, tesauros y ontologías más relevantes e innovadoras, que se deben conocer aplicar para manejar los datos producto de investigación, en especial en las áreas biológicas.
Palabras clave: bases de datos, curación de datos, datos ligados, ontologías, semántica, tesauros.

INTRODUCCIÓN
Todos generamos, utilizamos y transformamos información de manera inevitable. Manejar la información contenida en la Web se ha convertido en una habilidad indispensable principalmente para tener acceso a la información (Shorish, 2015), uno de los derechos humanos básicos. En muchos de los países desarrollados es una competencia básica que se debe fomentar llamada comúnmente alfabetización informacional (*information literacy* en inglés) (Bruce, 1997). Esta alfabetización es un conjunto de aptitudes para localizar, manejar y utilizar la información de forma eficaz para una gran variedad de finalidades (Grassian y LeMire, 2016).

La información está en diversos formatos, mucha información es digital y está disponible a través de la Web, se archiva, se sistematiza y se accede a ella a través de bases de datos u otros sistemas de consulta; se transfiere utilizando aplicaciones de la Web 2.0 (web social), se procesa usando avances sofisticados como los que ofrece la Web 3.0 (semántica y ligada) y en varios casos puede consultarse en acceso abierto, incluso con una licencia *Creative Commons* (Hall y Tiropanis, 2012).

Hay una producción acelerada de una gran variedad de programas, aplicaciones, herramientas, utilidades, colecciones, bases de datos, repositorios, en fin, diversos recursos y servicios electrónicos que permiten agrupar, clasificar y visualizar documentos y objetos digitales de manera inmediata y sistematizada, lo que ha reducido la energía, el costo y el tiempo requeridos para el procesamiento de esta información.

Los científicos, tienen como tarea principal, generar nuevo conocimiento a partir de la estructuración de información que producen y del procesamiento de los datos que generan por medio de experimentación, observación y comparación los cuales constituyen, una vez articulados, expuestos, probados y evaluados, la evidencia teórica y/o experimental que es sometida a escrutinio y publicada por medio de artículos científicos, los cuales constituyen el producto, insumo, medio de comunicación, sustento, objeto de estudio y fungen como herramienta de evaluación e innovación, a través de un proceso iterativo dirigido a la creación de **nuevos** productos, procesos, **conocimientos** o servicios (Kusiak, 2009) (Fig. 1).

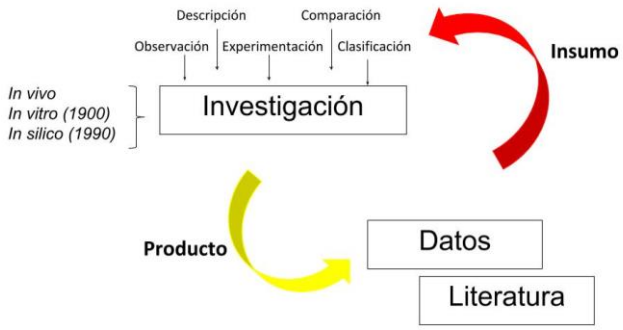

Fig. 1. Generación de datos de investigación en biología.

Una de las innovaciones, producto de la revolución informática, es la que han experimentado los datos científicos, como objeto, como proceso y como producto. El ciclo, formatos, procesos y la dinámica de los datos científicos se han transformado radicalmente, esto modifica la forma en que los académicos producen, comunican y acceden a los resultados de una investigación. El formato digital y el acceso abierto están particularmente representados en este cambio (Bezuidenhout et al., 2017). La innovación basada en datos implica la explotación de cualquier tipo de proceso de innovación para crear valor (Stone y Wang, 2014).

La tendencia emergente de la gran innovación impulsada por los datos está llevando al desarrollo de productos basados en datos y servicios, que permite la planificación impulsada por datos, la comercialización basada en datos y la generación de datos, en todos los sectores y dominios industriales. Pero en plena era de la información y los grandes datos, cerca del 80% de los datos científicos son oscuros (*dark data*) esto es, no son útiles, porque muchos de ellos no están disponibles o son heterogéneos, imprecisos, incompletos, o incorrectos, porque no se sigue un plan adecuado para generarlos, no son cuidadosamente registrados, son casi invisibles para los especialistas y otros usuarios potenciales, son subutilizados y con el tiempo se pierden (Heidorn, 2008).

¿Quién no tiene sus datos en un documento de texto o en una hoja de cálculo en su computadora, desestructurados, los usa para una sola investigación, no se integran a otros conjuntos de datos, no se normalizan, esto es no se utilizan estándares o especificaciones para unificarlos y que sean interoperables, no se publican, no se comparten, ni se re usan? ¿Cuántos científicos agregan los créditos, la forma de citar y la licencia correspondiente a su conjunto de datos, Quiénes lo autoarchivan en un repositorio institucional o los ponen disponibles en acceso abierto?

Por todo esto, la cantidad de tiempo y recursos que se pierden es incalculable, se invierten grandes cantidades de recursos para subvencionar macroproyectos innovadores, interdisciplinarios, en el que participan decenas o cientos de investigadores de diferentes instituciones y diversos países, pero en muy pocos casos se tiene la ciberinfraestructura necesaria para manejarlos, que incluya la tecnología, recursos, personal, proyectos, procesos, documentación y protocolos que facilitan el manejo, codificación, captura, validación y metadatos pertinentes.

El manejo de información científica es la especialidad encargada de investigar, difundir, enseñar y aplicar las teorías, los métodos, los conceptos, los enfoques, las herramientas y las

aplicaciones computacionales relacionadas con gestionar los datos insumo y producto de la práctica científica para obtención de nuevo conocimiento, aplicaciones tecnológicas y de innovación como un recurso valioso (Detlor, 2010). Implica la planeación, el desarrollo y la ejecución de diseños, arquitecturas, políticas, prácticas y procedimientos para procesar apropiadamente todo el ciclo de la información a nivel individual, grupal o institucional (Ou y Zhou, 2016).

Por lo tanto, esta es la disciplina encargada de caracterizar, estudiar, recolectar, sistematizar, estructurar, preservar, curar, recuperar, clasificar, evaluar, procesar, compartir, usar, publicar, visualizar y analizar de manera innovadora y óptima la información y los conocimientos asociados a ésta para obtención de nuevo conocimiento, resolver problemas y tomar decisiones por procesos informáticos (Venkatakrishnan et al., 2016).

En la figura 2 se representan todos estos niveles, que siempre inician con el manejo de información adecuada a nivel individual.

Fig. 2. Niveles de manejo de los datos científicos.

Por todas estas razones es importante investigar, difundir, enseñar y aplicar las teorías, métodos, conceptos, enfoques, herramientas y aplicaciones computacionales más novedosas y útiles relacionadas con la gestión de los datos científicos, como un recurso valioso e innovador, acorde a las tendencias mundiales de punta.

El objetivo de este artículo fue investigar cuáles son los temas, servicios y software, entidades y colecciones, tesauros y ontologías más relevantes e innovadoras que un científico debe conocer y aplicar para manejar los datos producto de investigación de manera eficiente y óptima, en especial en las áreas biológicas, se debe conocer y aplicar el manejo de datos producto de investigación de manera pertinente, eficiente y óptima.

MATERIALES Y MÉTODOS

Se realizó una recuperación de información de recursos y bibliografía en la Web, con lo que se generaron colecciones en línea.

El manejo de datos científicos

Uno de los retos para los científicos del siglo XXI es generar colecciones de información robustas, pertinentes e interoperables y cuyos datos sean sistematizados con la finalidad de estructurarlos, normalizarlos (estandarizarlos), analizarlos y utilizarlos. Además deben cumplir con ciertas características para que sean útiles, pertinentes y de calidad. No solo eso, se ha implementado el uso de estándares, lenguajes y vocabularios controlados que permiten la interoperabilidad y la facilidad de compartir los datos digitales, lo que propicia el aumento de la colaboración entre diferentes disciplinas y permite que investigadores, estudiantes, profesores e instituciones localizadas en diversas áreas geográficas, se interrelacionen, compartan y acuerden formas de manejar sus datos (Gallagher et al., 2015).

Un conjunto de datos científicos debe cumplir con los siguientes propósitos: 1) manejar los datos, producto de investigación, con base en un plan y protocolos establecidos; 2) sistematizar los datos en bases de datos en línea con metadatos adecuados y compartirlos, esta es característica indispensable para la colaboración y difusión de los resultados, así como para la reproductibilidad; 3) para realizar análisis *in silico* de la información, mediante aproximaciones informáticas y computacionales, tales como la bioinformática, la informática médica, la informática biológica o la quimioinformática, por ejemplo; 4) y finalmente privilegiar los datos abiertos. De hecho en la actualidad una de las novedades es anexar al artículo científico publicado, el conjunto de datos utilizado en el análisis realizado ya sea como anexo o depositarlo en un repositorio para dicho fin. Esto para cumplir con uno de los propósitos básicos de la ciencia que es la reproducibilidad del experimento. Nature (2019), presenta una de los repositorios más recomendados para este fin https://www.nature.com/sdata/policies/repositories.

Para generar un conjunto de datos que pueda explotarse lo más posible, pueda contestar muchas preguntas, reutilizarse y ser interoperable con otro conjunto de datos debe hacer un buen plan y procedimiento para la curación de los datos, usar metadatos adecuados, hacerlo con calidad y aplicar buenas prácticas (Veiga et al., 2017), esto se expone a continuación con más detalles.

Curación de datos

La curación de datos se define como la actividad de organizar, representar, y hacer que la información sea accesible para los seres humanos y las computadoras mediante la traducción e integración de metadatos de una colección, cuyos objetivos se centran en formar conjuntos de información de calidad, precisa, pertinente, estructurada, interoperable, accesible y reutilizable (Johnston, 2017). En la colección 1, se presentan recursos disponibles en la Web para manejar datos producto de investigación de manera pertinente, estandarizada e interoperable.

La generación de una colección es un proceso detallado y meticuloso, existen varias propuestas que establecen las acciones, condiciones y procesos que deberán seguirse para producir una colección de datos digital consistente, muchas de ellas diseñadas por las distintas entidades encargadas de realizarlas o estandarizarlas, resalta por ser una de las más usadas la establecida por el Centro de Curación Digital de la Universidad de Edimburgo (DCC, 2019), un centro líder a nivel mundial en el que trabajan expertos en información digital enfocados a construir capacidades, capacitaciones y habilidades para la investigación en el manejo de datos digitales.

Esta instancia establece un modelo denominado "Ciclo de vida de la curación digital" el cual representa en un gráfico las etapas necesarias para la curación y la preservación de los datos desde la conceptualización inicial hasta su publicación. Este modelo es muy útil para planear las actividades de producción de una colección digital. El modelo establece 11 pasos y 16 acciones que establecen, según las necesidades, todas las etapas necesarias y en la secuencia correcta para definir

las funciones y responsabilidades específicas de todos los involucrados en el proyecto y construir un marco de normas y tecnología idóneos que se deben implementar, tanto en casos generales, como en determinadas situaciones o disciplinas, para garantizar que los procesos y las políticas estén adecuadamente elegidas, implementadas y documentadas.

Otras instancia que promueven las buenas prácticas del manejo de datos producto de investigación y que son imprescindibles para este tema como DRYAD (https://datadryad.org/) (DRYAD, 2019), DataCite (https://datacite.org/) (DataCite, 2019), FAIR Principles (www.go-fair.org/fair-principles/) (FAIR Principles, 2019), OpenAIRE (www.openaire.eu/) (Open AIRE, 2019), Figshare (https://figshare.com/) (Figshare, 2019) y Foster Open Science (www.fosteropenscience.eu/) (Para más ejemplos ver la colección 1 de recursos).

Los metadatos

Los metadatos constituyen información estructurada que describe, explica, localiza y hace que sea más fácil de recuperar, utilizar o manejar la información (NISO, 2019) y pueden ser de tres tipos: estructurales, descriptivos y administrativos. El uso de metadatos adecuados permite una correcta estructuración y descripción de los datos, puesto que detallan las características de la información que se genera como la fecha de creación, la licencia de uso, el autor, el tipo de objeto digital (texto, imagen, sonido, video) o la dirección electrónica en la que se encuentra alojado. Además, para generar metadatos pertinentes existen esquemas de metadatos como Dublin Core (DCMI, 2019) y Darwin Core (Wieczorek et al., 2012; BIS, 2019), estándares como los ISO y NISO, identificadores como PMID (NCBI, 2019), ORCID (ORCID, 2019) y DOI (Crossref, 2019).

Vale la pena resaltar que para generar metadatos descriptivos se pueden utilizar herramientas de organización del conocimiento, especialmente vocabularios controlados, que permiten describir la información utilizando tecnologías semánticas, tal es el caso de los tesauros que consisten en una lista controlada y estructurada de términos para describir contenidos temáticos como el Tesauro de la UNESCO (http://vocabularies.unesco.org/browser/thesaurus/es/) (UNESCO, 2019) y las ontologías, que son una representación formal de un conjunto de conceptos y las relaciones lógicas entre esos conceptos dentro de un dominio de conocimiento, por ejemplo la muy usada Gene Ontology (GO) (http://geneontology.org/) que se han convertido en una de las herramientas más sofisticadas e innovadoras para recuperar, organizar y analizar la información (Ashburner et al., 2000), su desarrollo y uso más sofisticado se ha logrado en la biomedicina (Lapatas et al., 2015).

En la colección 2 se registran algunos tesauros y en la colección 3 algunas de las ontologías más utilizadas para describir y clasificar datos científicos, en especial biológicos.

Acceso, créditos y licencias

Por supuesto, que para que estos datos sean útiles no basta que sean sistematizados es necesario que sean interoperables, es deseable que estén disponibles en acceso abierto (*open access*) y datos abiertos (*open data*) y que cuenten con una licencia *Creative Commons*®, que es el sistema del derecho de autor que promueve la libertad creativa que plantea un esquema en el que se otorga permiso para usar las obras (Creative Commons, 2019). Los datos abiertos se definen como los que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona (OKI, 2019).

La interoperabilidad es la capacidad que tiene un producto o un sistema, cuyas interfaces son totalmente conocidas, para funcionar con otros productos o sistemas existentes o futuros y eso sin restricción de acceso o de implementación (I´AFUL, 2019). Sin embargo, no es tan fácil tener interoperabilidad entre sistemas, debido a que no siguen con normas de estilo, etiquetas XML,

estructuración de datos ni el uso de estándares específicos, como los descritos en la sección de metadatos.

Cuando los datos son estructurados bajo estos criterios, es como las plataformas comienzan a tener interoperabilidad. Se estima que para el año 2020, 15 millones de usuarios consumirán datos interoperables, aplicándolo a cualquier cosa y todo lo que podemos pensar: automóviles, aviones, trenes, marcapasos, bombillas, monitores para bebés, hogares, oficinas, fábricas, centrales nucleares, rejillas eléctricas, e incluso juguetes.

De tal forma que se han desarrollado nuevas técnicas analíticas, tecnologías de acceso y modelos de organización proporcionados por disciplinas como el cómputo, la bioinformática y la ciencia de los datos, para explotar los datos con disciplinas y análisis complejos como por ejemplo, los datos ligados, la altimetría, la minería, la semántica y los grandes datos.

Debido a la creciente información digital, ha surgido la necesidad de interpretar las métricas y datos cualitativos de la información proveniente de la Web de sitios académicos, blogs científicos, citas de wikipedia, gestores de referencias como Mendeley, revisión por pares de Faculty of 1000 o citas compartidas por twitter. La altmetría (*altmetric* en inglés) es un método usado recientemente para procesar estos datos generados en redes sociales que se ha implementado en diversos sitios web académicos (Altmetric, 2019).

Cuando los datos son estructurados con base en normas y estándares, se pueden organizar y jerarquizar de tal manera que se puede estructurar el lenguaje natural por medio de la semántica, agregando metadatos que representan el conocimiento asociado a una ontología (Gardner, 2005) para que tenga significado y que además sea interoperable. Este es el caso de Gene ontology (GO) mencionado anteriormente, cuya tecnología semántica actual permite incorporar conocimiento biológico para asociar y clasificar diferentes genes con sus funciones, procesos y organelos correspondientes. Esto facilita una mejor comprensión de los fenómenos biológicos subyacentes al experimento correspondiente, permite la identificación de procesos pertinentes a diferentes condiciones biológicas y auxilia a los usuarios para tomar decisiones apropiadas en un biológicos determinado y asegurar un conocimiento efectivo basado en las anotaciones entre Gene Ontology y las bases de datos biológicas (Mazandu et al., 2017).

Para consultar vocabularios controlados como taxonomías, tesauros y ontologías de interés biológico se pueden consultar directorios como BARTOC (www.bartoc.org/es) (BARTOC, 2019), NCBO BioPortal (https://bioportal.bioontology.org/) (NCBO, 2019) y The OBO Foundry (www.obofoundry.org/) (Smith et al., 2007).

Actualmente existen colecciones inmensas de información de todo tipo: financiera, educativa, de investigación, médica, política, social, ambiental, música, arte, físico-matemática, entre muchos otros. A estas colecciones gigantes se les llaman grandes datos (*Big data* en inglés) y pueden contener terabytes ($1x10^{12}$), petabytes ($1x10^{15}$), exabytes ($1x10^{18}$) o zetabytes ($1x10^{21}$). Los análisis de esta información se realizan mediante un procedimiento llamado minería de datos (*data mining* en inglés) que actualmente es una subdisciplina especializada en la literatura científica, ya que comprende un desafío intelectual creciente (Cheadle et al., 2017).

Desde el punto de vista económico, los datos como bienes comunes podrían ser explotados simultáneamente por muchos usuarios para diferentes metas, los grandes datos ofrecen potencialmente retornos significativos a escala y alcance (Marx, 2013). La Web semántica se utiliza para dar significado a los datos en un contexto específico; que además permite procesar la

información. El portal de Europe PMC (https://europepmc.org/) contiene una de las colecciones de literatura biomédica más importante, despliega para cada artículo un análisis del texto y permite identificar con diferentes colores términos de distintas clases como enfermedades, organismos, productos químicos, nombres de genes o proteínas y términos de Gene Ontology) (Europe PMC, 2019), las relaciones biológicas funcionales (por ejemplo las asociaciones gen-enfermedad, interacciones proteína-proteína), los eventos biológicos (por ejemplo, la fosforilación), así como las funciones biológicas (por ejemplo función de un gen o una proteína); estas anotaciones se resaltan en los resumes y el texto completo y son una herramienta muy útil de visualización y análisis (Fig. 3).

Fig. 3. Procesos y características de los datos científicos.

El uso pertinente de los datos de investigación permite identificar áreas de oportunidad, optimizar los recursos y aprovechar el tiempo al máximo, sin duda se ha reducido la energía, el costo y el tiempo requeridos para el análisis de los datos, se diseñan constantemente nuevas herramientas para realizar búsquedas más eficientes y precisas, así como para hacer análisis mejores y más extensos. Pero la innovación solo tiene sentido en un contexto en el que existe la cultura necesaria para conocerlos e implementarlos.

No importa que tan sofisticados o innovadores sean las soluciones, implementaciones o las herramientas, si no se difunden, adoptan e institucionalizan, para ello es necesaria la aplicación de procesos, protocolos y estándares que permiten a los usuarios seguir los pasos y formatos adecuados, sea esta una invitación para que optimicen, aprovechan, quieran y consientan sus datos de investigación.

Con todo lo expuesto anteriormente, se concluye que el manejo pertinente de los datos científicos en la era digital implica: 1. Diseñar un plan de curación de datos, 2. Plantear buenos datos y metadatos, 3. Publicar los datos, 4. Compartir los datos, 5. Reutilizar los datos, 6. Ligar los datos, 7. Generar datos interoperables, 8. Usar un repositorio, 9. Asignar identificadores permanentes a los datos, 10. Agregar licencia de uso a los datos, 11. Dar crédito a los datos, 12. Citar los datos correctamente. 13. Promover buenas prácticas para los datos, 14. Invitar a los alumnos y colaboradores a utilizar protocolos y herramientas para optimizar el uso de datos y 15. Colaborar con especialistas en cómputo e información para generar datos de calidad.

Con la información recopilada en éste trabajo, se han generado las siguientes colecciones en línea:

Colección 1. Algunos recursos que promueven el manejo adecuado de datos.
https://hypothes.is/groups/MbpJgqKj/bioinfovestigacion-edu?q=manejo_datos_cientificos

Colección 2. Algunos tesauros de interés.
(https://hypothes.is/groups/MbpJgqKj/bioinfovestigacion-edu?q=spar%2Ffabio%2FThesaurus)

Colección 3. Algunas ontologías de interés más utilizados para describir y clasificar información científica.
(https://hypothes.is/groups/MbpJgqKj/bioinfovestigacion-edu?q=spar%2Ffabio%2FOntology)

REFERENCIAS

1. Altmetric, 2019. Altmetric LLP. https://www.altmetric.com/ (accesado en mayo 27, 2019).

2. Ashburner M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin y G. Sherlock, 2000. Gene Ontology: tool for the unification of biology. Nature Genetics, 25: 25.29. DOI: 10.1038/75556

3. BARTOC (Basel Register of Thesauri, Ontologies & Classifications), 2019. https://www.bartoc.org/es (accesado en mayo 27, 2019).

4. Bezuidenhout L.M., S. Leonelli, A.H. Kelly y B. Rappert, 2017. Beyond the digital divide: towards a situated approach to open data. Science and Public Policy, 44 (4): 464–475.
DOI: 10.1093/scipol/scw036

5. BIS (Biodiversity Information Standards TDWG), 2019. Darwin Core http://rs.tdwg.org/dwc/ (accesado en mayo 27, 2019).

6. Bruce C., 1997. Las siete caras de la alfabetización en información en la enseñanza superior. Anales de Documentación, 6: 289-294.

7. Cheadle C., Cao H., Kalinin A. y J. Hodgkinson, 2017. Advanced literature analysis in a Big Data world. Annals of the New York Academy of Sciences, 1387(1): 25–33.
DOI: https://doi.org/10.1111/nyas.13270

8. Creative Commons, 2019. https://creativecommons.org/licenses/?lang=es
(accesado en mayo 27, 2019).

9. Crossref, 2019. https://www.crossref.org/ (accesado en mayo 27, 2019).

10. DataCite, 2019. Datacite.org. https://datacite.org/(accesado en mayo 27, 2019).

11. Detlor B., 2010. Information management. International Journal of Information Management, 30(2): 103-108. DOI: 10.1016/j.ijinfomgt.2009.12.001

12. DCMI (The Dublin Core Metadata Initiative), 2019. Dublin Core. http://dublincore.org/ (accesado en mayo 27, 2019).

13. DCC (Digital Curation Centre), 2019. http://www.dcc.ac.uk/ (accesado en mayo 27, 2019).

14. DRYAD (Dryad Digital Repository), 2019. https://datadryad.org/ (accesado en mayo 27, 2019).

15. Europe PMC, 2019. Europe PMC. https://europepmc.org/ (accesado en mayo 27, 2019).

16. FAIR Principles (GO FAIR), 2019. GO FAIR. https://www.go-fair.org/fair-principles/ (accesado en mayo 27, 2019).

17. Figshare, 2019. Figshare-credit for all your research. https://figshare.com/ (accesado en mayo 27, 2019).

18. FOSTER, 2019. Foster Open Science https://www.fosteropenscience.eu/ (accesado en mayo 27, 2019).

17. Gallagher J., J. Orcutt, P. Simpson, D. Wright, J. Pearlman y L. Raymond, 2015. Facilitating open exchange of data and information. Earth Science Informatics, 8(4): 721-739. DOI: 10.1007/s12145-014-0202-2

18. Gardner S.P., 2005. Ontologies and semantic data integration. Drug Discovery Today, 10(14): 1001-1007. DOI: https://doi.org/10.1016/S1359-6446(05)03504

19. Grassian E. y S. LeMire, 2016. Information literacy and instruction: how can this column help You? Reference and User Services Quarterly, 56(2): 75-76 https://journals.ala.org/index.php/rusq/article/view/6182 (accesado en mayo 27, 2019)

20. Hall W. y T. Tiropanis, 2012. Web evolution and Web Science. Computer Networks, 56: 3859-3865. DOI: 10.1016/j.comnet.2012.10.004

21. Heidorn P.B., 2008. Shedding light on the dark data in the long tail of science. Library Trends, 57(2): 280-299. DOI: https://doi.org/10.1353/lib.0.0036

22. I´AFUL, 2019. Interopérabilité-Liberté-Pérennité. https://aful.org/gdt/interop (accesado en mayo 27, 2019).

23. Johnston L.R. (Ed), 2017. Curating research data volume one: practical strategies for your digital repository. Association of College & Research Libraries, Chicago, Illinois.

24. Kusiak A., 2009. Innovation: a data-driven approach. International Journal of Production Economics, 122 (1): 440–448. DOI: 10.1016/j.ijpe.2009.06.025

25. Lapatas V., M. Stefanidakis, R.C. Jiménez, A. Via y M.V. Schneider, 2015. Data integration in biological research: an overview. Journal of Biological Research-Thessaloniki, 22(9).1-16. DOI: https://doi.org/10.1186/s40709-015-0032-5

26. Marx V., 2013. Biology: The big challenges of big data. Nature, 498(7453): 255–260. DOI: https://doi.org/10.1038/498255a

27. Mazandu G.K. Chimusa E.R. y N.J. Mulder, 2017. Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Briefings in Bioinformatics, 18(5): DOI: 886-901. https://doi.org/10.1093/bib/bbw067

28. Nature, 2019. https://www.nature.com/sdata/policies/repositories (accesado en mayo 27, 2019).

29. NCBI (National Center for Biotechnology Information), 2019 https://www.ncbi.nlm.nih.gov/pmc/pmctopmid/ (accesado en mayo 27, 2019).

30. NCBO (NCBO BioPortal), 2019. Bioportal.bioontology.org https://bioportal.bioontology.org/ (accesado en mayo 27, 2019).

31. NISO (National Information Standards Organization), 2019. http://www.niso.org/home/ (accesado en mayo 27, 2019).

32. OKI (Open Knowledge International), 2019. Open data handbook. http://opendatahandbook.org/guide/es/ (accesado en mayo 27, 2019).

33. OpenAIRE. (Openaire.eu.) https://www.openaire.eu/ (accesado en mayo 27, 2019).

34. ORCID (Open Researcher and Contribution ID), 2019. https://orcid.org/(accesado en mayo 27, 2019).

35. Ou S. y Y. Zhou, 2016. Current status of scientific data curation research and practices in Mainland China. LIBRES: Library and Information Science Research Electronic Journal, 26(1): 73-88.

36. Shorish Y., 2015. Data information literacy and undergraduates: a critical competency. College & Undergraduate Libraries, 22(1): 97-106. DOI: https://doi.org/10.1080/10691316.2015.1001246

37. Smith B., M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, The OBI Consortium, Neocles Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel y S. Lewis, 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11): 1251-1255. DOI: https://doi.org/10.1038/nbt1346

38. Stone D. y R. Wang, 2014. Deciding with data–How data-driven innovation is fuelling Australia's economic growth. PricewaterhouseCoopers (PwC). https://www.pwc.com.au/consulting/assets/publications/data-drive-innovation-sep14.pdf

39. UNESCO (United Nations Educational, Scientific and Cultural Organization), 2019. Tesauro de la UNESCO http://vocabularies.unesco.org/browser/thesaurus/es/
SKOS Tesauro UNESCO. https://skos.um.es/unescothes/?l=es (accesado en mayo 27, 2019).

40. Veiga A.K., A.M. Saraiva, A.D. Chapman, P.J. Morris, C. Gendreau, D. Schigel y T.J. Robertson, 2017. A conceptual framework for quality assessment and management of biodiversity data. PLoS ONE, 12(6): e0178731. DOI: 10.1371/journal.pone.0178731

41. Venkatakrishnan S.V, K.A. Mohan, K. Beattie, J. Correa, E. Dart, J.R. Deslippe, A. Hexemer, H. Krishnan, A.A. MacDowell, S. Marchesini, S.J. Patton, T. Perciano, J.A. Sethian, R. Stromsness, B.L. Tierney, C.E. Tull, D. Ushizima, D.Y. Parkinson, 2016. Making advanced scientific algorithms and big scientific data management more accesible. Electronic Imaging, 2016(19): 1-7.
DOI: 10.2352/ISSN.2470-1173.2016.19.COIMG-155

42. Wieczorek J., B. David, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson y D. Vieglais, 2012. Darwin Core: an evolving community-developed biodiversity data standard. PLoS ONE 7(1): e29715. DOI: https://doi.org/10.1371/journal.pone.0029715

BIOCYT Biología, Ciencia y Tecnología, se encuentra actualmente indexada en

alojada en los repositorios

y en bases electrónicas de bibliotecas