Factores determinantes para la implementación del esquema de metadatos para repositorios de datos de investigación de la Política de Ciencia Abierta en México

Miguel Adolfo Guajardo Mendoza
CONACYT

Introducción

De acuerdo al Capítulo X, "Del Acceso Abierto, Acceso a la Información Científica, Tecnológica y de Innovación y del Repositorio Nacional" de la Ley de Ciencia y Tecnología (Capítulo adicionado DOF 20-05-2014), el Consejo Nacional de Ciencia y Tecnología (CONACYT) es el órgano instruido para el diseño e impulso de la estrategia nacional para la democratización de la información científica, tecnológica y de innovación nacional e internacional a texto completo, en formatos digitales, a través de repositorios construidos por las instituciones de educación superior y centros de investigación.

La regulación de esta estrategia parte de los lineamientos emitidos por el CONACYT, conformados actualmente por los Lineamientos Generales de Ciencia Abierta (LGCA), los Lineamientos Jurídicos de Ciencia Abierta (LJCA) y los Lineamientos

Específicos para Repositorios (LER), que sumaron los programas de Revistas, Conricyt, SIICYT, Comunicación Pública de la Ciencia y Conectividad al programa de Repositorios para la conformación de una Política de Ciencia Abierta en México.[1]

Los lineamientos específicos para repositorios tienen por objeto coordinar las políticas, los recursos, los programas y las acciones realizadas por el CONACYT relacionadas con los repositorios Nacional e institucionales de ciencia abierta. Están conformados por veintitrés lineamientos específicos, tres transitorios y tres apéndices con la descripción de la estructura de los metadatos de literatura y datos para la interoperabilidad de los repositorios Nacional e institucionales.

La conformación de esta estructura se encuentra basada en el esquema *OpenAIRE* para el manejo de repositorios de literatura 3.0 y para el manejo de repositorios de datos 2.0, como se indica en los Apéndices 1 y 2, así como en los Lineamientos Décimo Quinto al Décimo Séptimo de los Lineamientos Específicos para Repositorios.

El crecimiento exponencial en la generación de los datos y su proliferación ha generado nuevas perspectivas para el uso, análisis e interpretación de la información que contienen, lo que ha aumentado la atención en el papel que tienen los metadatos en la organización de la información en herramientas como los repositorios de datos, lo que permite su localización y recuperación (Oliphant 2017).

Es necesario diferenciar que cuando se habla de datos, se pueden referir a diversos tipos de datos, como los administrativos, de gestión, los financieros o los de investigación.

---

1 Repositorio Nacional. Política de Ciencia Abierta. Disponible en https://repositorionacionalcti.mx/documentos.

Para el caso de la Política de Ciencia Abierta en México y los repositorios que la integran, esta última tipología de los datos es la de particular interés, pues en su objetivo está la máxima diseminación del conocimiento científico, tecnológico y de innovación, resultado de las investigaciones en nuestro país.

La política de la biblioteca de la Universidad de Melbourne (The University of Melbourne s.f.) describe los datos de investigación como hechos, observaciones o experiencias sobre un argumento, teoría o prueba. Estos pueden ser numéricos, descriptivos o visuales y pueden presentarse de manera cruda, analizada, experimental u observacional.

Para Greenberg, los datos son "la esencia de la ciencia", los portadores de la verdad científica cuando se revisan los hallazgos (Greenberg *et al.* 2009). Ella también retoma el comentario de Davis y Vickery, quienes mencionan que "los conjuntos de datos pueden abarcar desde los sistemas de información geográficos o geoespaciales (GIS), hasta datos genómicos o cualquier conjunto de datos que respalde a una publicación académica, como los datos de un censo".

Desde el nacimiento de la estrategia de Acceso Abierto y su posterior evolución a la Política de Ciencia Abierta en México, se les ha denominado datos primarios de las investigaciones a aquellos conjuntos de información recolectada y utilizada para la investigación académica, científica y de innovación, y ha culminado con la diferenciación entre los repositorios de datos y los de literatura, con la implementación de un esquema de metadatos particular para cada tipo, con la intención de describir de la mejor forma la información de los recursos que exponen y son cosechados por el Repositorio Nacional, los cuales son descritos en los lineamientos mencionados anteriormente.

De la estrategia de acceso abierto a la política de ciencia abierta en México

El Estado es el principal financiador de la investigación, que a través de los recursos públicos apoya el desarrollo de los países. En el caso de México, el gobierno federal a través del Conacyt distribuye dichos recursos públicos para la producción científica. Sin embargo, los productos de la investigación generalmente son publicados en editoriales con alto prestigio académico, pero con un elevado costo de suscripción. Esta situación en la que las investigaciones financiadas con recursos públicos y que para acceder a la misma se requiere invertir más recursos afecta el ciclo de producción científica no sólo en México, sino en el mundo.

Desde 2002, se han apoyado iniciativas a nivel global en beneficio del acceso a la información científica, entre las que se destaca la Declaración de Bethesda (2002), la Declaración de Berlín (2003) y la Iniciativa de Budapest (2003), desde las que nace el movimiento internacional denominado "Acceso Abierto" (*Open Access*), que es definido como

> la disponibilidad gratuita en Internet público, para que cualquier usuario la pueda leer, descargar, copiar, distribuir, imprimir, buscar o usarla con cualquier propósito legal, sin alguna barrera financiera, legal o técnica, fuera de las que son inseparables de las que implica acceder al Internet mismo (Iniciativa de Budapest 2003).

Considerando esto, el 20 de mayo de 2014 se publicó en el *Diario Oficial de la Federación* el Decreto por el cual se modifica la Ley de Ciencia y Tecnología, en el cual se incluye el Capítulo X titulado Del Acceso Abierto, Acceso a la información Científica, Tecnológica y de Innovación

y del Repositorio Nacional. Este decreto establece que el CONACYT es el órgano encargado de la creación y operación del Repositorio Nacional, al cual define como "[...] la plataforma digital centralizada que, siguiendo estándares internacionales, almacena, mantiene y preserva la información científica, tecnológica y de innovación, la cual deriva de las investigaciones, productos educativos y académicos" (CONACYT Informe general…2017a).

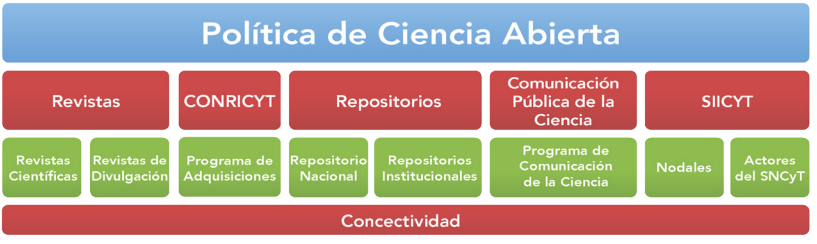
Desde ese momento, el CONACYT adquiere la obligación de presentar los lineamientos y las disposiciones correspondientes para el funcionamiento de los Repositorios Nacional e Institucionales, así como para capacitar, convocar, organizar y coordinar a las instituciones e instancias en materia de acceso abierto y el funcionamiento de dichas plataformas a través de lo que se denominó la Estrategia de Acceso Abierto a la Información Científica, Tecnológica y de Innovación. Esto llevó a la publicación de los Lineamientos Generales para el Repositorio Nacional y los Repositorios Institucionales en noviembre de 2014 y los Lineamientos Técnicos para el Repositorio Nacional y los Repositorios Institucionales en noviembre del siguiente año.

Durante los siguientes dos años, se identificó una nueva tendencia a nivel internacional que se orientaba a hacer más transparente y colaborativo el proceso de generación del conocimiento científico (principalmente financiado con recursos públicos) denominado Ciencia Abierta (*Open Science*), paradigma que busca que cualquier interesado pueda acceder libre y gratuitamente a los materiales y recursos de información que resultan del proceso de investigación, en cualquiera de sus etapas, con la posibilidad de usarlos, reusarlos, modificarlos, compartirlos y difundirlos mediante la utilización de las Tecnologías de la Información y Comunicación (TIC).

En sincronía con este paradigma, el CONACYT revisó los lineamientos vigentes y la viabilidad de transición; dentro de su espectro de acción incluiría además del Repositorio Nacional y los repositorios institucionales a los programas de Comunicación Pública de la Ciencia, Revistas, Consorcio para la Adquisición y Diseminación de la información, Conectividad y el Sistema Integrado de Información Científica y Tecnológica. Esto permitiría diluir las barreras para compartir cualquier producto, recurso o herramienta generada en cualquiera de las etapas del proceso de investigación, desde la generación de datos, la publicación de los resultados en documentos arbitrados, hasta la divulgación de las investigaciones al público en general, ampliando el alcance originalmente planeado.

En consecuencia, el 9 de junio de 2017 se emitieron los Lineamientos Generales de Ciencia Abierta, que sustituían a los Lineamientos Generales para el Repositorio Nacional y los Repositorios Institucionales; posteriormente, cada uno de los seis programas que componen la política ha realizado las adecuaciones necesarias en sus instrumentos con la intención

Imagen 1



Fuente: Elaboración interna para la presentación de la política en diferentes instancias.

de alinearse con lo establecido en los nuevos lineamientos, y han quedado articulados de la siguiente manera:

El CONACYT, a través de la política de Ciencia Abierta, busca asegurar la máxima diseminación del conocimiento científico, tecnológico y de innovación entre la población en general a través de cualquier medio, incluidos los digitales. La Ciencia Abierta será democrática y universal, y se regirá por los siguientes principios:

I. Máxima apertura.
II. Máxima captación y colaboración.
III. Máxima facilidad de acceso.
IV. Costos mínimos o gratuidad.
V. Respeto a otros regímenes de Derecho, como la seguridad nacional, propiedad Intelectual, confidencialidad y reserva de datos, secretos protegidos, entre otros aplicables.[2]

El repositorio nacional y los repositorios institucionales

El programa de repositorios es el que da origen a la Política de Ciencia Abierta. Tiene por objetivo acopiar, preservar y asegurar el acceso abierto a los recursos de información científica, tecnológica y de innovación generados principalmente con recursos públicos. Para lograrlo, el programa se desagrega en dos componentes: 1) Repositorio Nacional y 2) Repositorios institucionales. Cada uno de estos componentes tiene un objetivo específico, que son:

2 De acuerdo con lo establecido en el Lineamiento Cuarto. Componentes y principios de la Política de Ciencia Abierta en los Lineamientos Generales de Ciencia Abierta (2017).

- Repositorio Nacional: diseminar los recursos de información (publicaciones científicas, productos del desarrollo tecnológico y la innovación y los datos primarios de la investigación) provenientes de los Repositorios Institucionales para fomentar su utilización, reúso y acelerar la colaboración científica.

- Repositorios Institucionales: apoyar mediante una convocatoria pública a aquellas instituciones públicas o privadas que realicen investigación científica y tecnológica el desarrollo de repositorios institucionales que agreguen valor al Repositorio Nacional.

Los repositorios mencionados contendrán tres tipos de información sin perjuicio de las disposiciones en materia de patentes, protección de la propiedad intelectual o industrial, seguridad nacional y derechos de autor: I) Publicaciones científicas, II) Productos del desarrollo tecnológico y la innovación y III) Datos primarios de las investigaciones.[3]

Si bien muchos de los agregadores de repositorios tradicionales se concentran especialmente en los denominados repositorios de literatura, una de las particularidades del Repositorio Nacional es su capacidad para agregar también datos primarios de las investigaciones, que se definen en los LER como aquella información recolectada y utilizada para la investigación académica, científica y de innovación. Esta

3 De acuerdo con lo establecido en el Lineamiento Cuarto. Componentes y principios de la Política de Ciencia Abierta en los Lineamientos Generales de Ciencia Abierta (2017).

información será presentada en los formatos originales de su creación y deberá contar con licencias que permitan su libre reutilización a través de los repositorios institucionales que las alojen.[4]

Su depósito, además, está condicionado a que se cumpla con una serie de documentos que permitan al usuario conocer la información y las instrucciones relevantes sobre la recopilación de los datos, la unidad de observación o la información del ponderador, entre otros. Estos documentos adicionales son a) Metodología, b) libro de códigos, c) cuestionario (si aplica) y d) resumen de contenido.

Contar con la posibilidad de compartir este tipo de recursos de información permiten a los investigadores y estudiantes de las instituciones depositar en sus repositorios estos conjuntos de datos que fomentan el aceleramiento del proceso de la investigación científica, acortando los tiempos de la liberación de los datos posteriores a los procesos editoriales y compartiéndolo libre y gratuitamente.

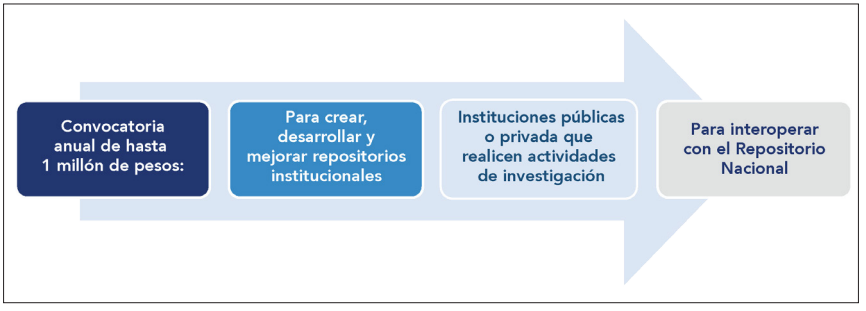
Aunado a la publicación de los lineamientos para los repositorios, desde diciembre de 2015 el CONACYT ha publicado tres convocatorias anuales con la finalidad de apoyar a las instituciones públicas y privadas que realicen actividades de investigación científica, tecnológica y de innovación y que busquen construir, mejorar o adecuar un repositorio institucional interoperable con el Repositorio Nacional, de acuerdo con lo establecido en los Lineamientos Específicos —antes Técnicos— para Repositorios.[5] Dichas convocatorias han otorgado hasta $1 000 000.00 (un millón de pesos) a diversas

4 Véase el Lineamiento Séptimo. Datos primarios de las investigaciones, de los Lineamientos Específicos para Repositorios.

5 De acuerdo con el objetivo de la Convocatoria 2017 para Desarrollar Repositorios Institucionales de Ciencia Abierta.

instituciones que poco a poco se han ido integrando al Repositorio Nacional a través de la interoperabilidad entre las plataformas, gracias al protocolo Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH por sus siglas en inglés), diseñado para la cosecha o recolección de metadatos de recursos de información, que permite redireccionar al usuario por medio de enlaces al recurso u objeto del repositorio de origen.

Imagen 2



Fuente: Elaboración interna para la presentación del proceso de las convocatorias en diversas instancias.

Al 26 de junio de 2018, se cuenta con más de cincuenta Repositorios Institucionales interoperando con el Repositorio Nacional, que aportan más de 22mil recursos de información y rondando el millón de visitas desde el lanzamiento de la versión 1.0 del RN a mediados de 2016.[6]

La meta para este año es que se encuentren interoperando más de ochenta repositorios institucionales y que se encuentren

6 Información disponible en el sitio del Repositorio Nacional https://repositorionacionalcti.mx/.

disponibles más de 50mil recursos de información, provenientes tanto de repositorios de literatura como de datos.

Factores determinantes para la implementación del esquema de metadatos para repositorios de datos

Uno de los elementos clave para lograr la interoperabilidad es el manejo de esquemas de metadatos estandarizados y adaptables para el correcto intercambio de información. Existe una amplia variedad de esquemas de metadatos compatibles con el estándar del OAI-PMH que los trabaja de una manera muy flexible.

El estándar estipula que, como mínimo, los proveedores de datos —en este caso, los repositorios— deben incorporar y exponer metadatos en algún esquema revisado, aceptado y conciliado que sea apropiado para el intercambio de la información de acuerdo con las disciplinas o el contenido de las plataformas (Breeding 2002).

Dentro de la discusión sobre la funcionalidad de los metadatos, Park incluso hace una analogía relacionado los esquemas de metadatos con los catálogos y las bases de datos de bibliotecas en línea tradicionales para encontrar, identificar, seleccionar y obtener documentos (Park 2009).

Contar con una política pública que puede otorgar estímulos a diversos tipos de instituciones del sector público o privado como universidades o centros de investigación, que a su vez concentran recursos de información de diversas áreas del conocimiento y disciplinas, requiere de implementar esquemas estandarizadas y de gran flexibilidad que permitan a las instituciones adecuar la descripción de los recursos de acuerdo a las necesidades de sus comunidades, así como mantener una

homogeneización de la información para la interoperabilidad.

Force 11 (Force 11 s.f.) incluso propone los Principios FAIR para el manejo de datos de investigación, los cuales han sido considerados como elementos clave para la adaptación de los Lineamientos Específicos para Repositorios:

1. Findable (recuperable).
2. Accesible.
3. Interoperable.
4. Re-usable.

A nivel del principio *Findable* (recuperable), se hace hincapié en que se utilicen identificadores persistentes y únicos a nivel global, que haya una enriquecida descripción de los metadatos, así como que se especifique el esquema a utilizar.

Para la accesibilidad, refieren a que los metadatos sean recuperables a través del identificador antes mencionado, con protocolos de comunicación estandarizados, libres y universalmente aplicables (como OAI-PMH), lo que permita una autenticación y autorización de las peticiones de los datos entre las plataformas.

En la interoperabilidad, se refuerza la importancia de que los metadatos sean accesibles y permitan la representación del conocimiento, que para el caso del CONACYT, se han enfocado en el uso del catálogo de áreas del conocimiento de la UNESCO.

Para que los datos sean reusables, mencionan que los metadatos deben ser compartidos de manera clara y con un lenguaje accesible, pero, sobre todo, que sean relevantes para los estándares de la comunidad que los va a utilizar.

Algunos de los elementos clave más persistentes en la

literatura y los ejercicios de evaluación sobre la calidad de los metadatos son la consistencia y la persistencia (Stivila y Gasser *apud* Park 2009). Si bien la variedad de "estándares" es amplia, para el manejo de una gama de tipología de repositorios es necesario que los esquemas a utilizar sean persistentes y además sean lo suficientemente adaptables para el uso correcto de los metadatos por parte de las instituciones, de modo que se exponga la mayor cantidad de información y se limite lo más posible el margen de información no recuperable.

Greenberg (Greenberg *et al*. 2009) menciona que no tiene sentido crear esquemas totalmente nuevos cuando otras iniciativas se han tomado el tiempo para deliberar y formalizar las suficientes propiedades de los metadatos.

En este ejercicio de exploración en la literatura y la implementación de esquemas de metadatos en redes de repositorios, algunos de los factores determinantes para la selección han sido:

1. Sencillez
2. Interoperabilidad
3. Adaptabilidad
4. Arquitectura semántica
5. Consistencia

Esto llevó a que para los repositorios de literatura se haya establecido el Apéndice 1 de los Lineamientos Específicos para Repositorios, que se encuentra basado en OpenAIRE 3.0 para el manejo de repositorios de literatura que a su vez cuenta con elementos de DublinCore, mientras que el esquema de metadatos para repositorios de datos se

7 Para mayor especificidad sobre el uso de los DataCite se puede consultar el Apéndice 1 de los Lineamientos Específicos para Repositorios del CONACYT.

encuentra en el Apéndice 2 de los mismos lineamientos, basados en OpenAIRE 2.0 para el manejo de repositorios de datos, estructurados con DataCite.[7]

OpenAIRE ha demostrado a lo largo de los años una gran persistencia y consistencia en sus esquemas, tomando en cuenta las ventajas que ofrecen diversos modelos de metadatos como DublinCore y DataCite, por citar algunos, trabajando colaborativamente no sólo con la comunidad europea, sino con la participación de expertos en materia de repositorios de América del Norte y Latinoamérica, lo que ha permitido tomar estos modelos y flexibilizar el uso de algunos metadatos para el mayor beneficio de diversos tipos de repositorios.

Estos convenios de comunicación e intercambio de información comparten la meta de proveer un esquema de metadatos independiente del dominio y proporcionar interoperabilidad mediante un pequeño número de propiedades, de la manera más sencilla posible y manteniendo las barreras técnicas para la implementación lo más bajas posible, lo que fortalece las capacidades para el desarrollo e implementación de estas plataformas bajo el paradigma de la Política de Ciencia Abierta en México, gracias a la basta documentación que la comunidad global pone en línea a través de diferentes foros y en el sitio de OpenAIRE.

Visión 2.0 del Repositorio Nacional

Los metadatos son "datos sobre los datos", lo que vuelve a los repositorios herramientas que permiten la explotación e interconexión de datos más allá del contenido que contienen y describen. El Repositorio Nacional genera una gran cantidad de información gracias al suministro de estadísticas de uso de los Repositorios Institucionales

y a sus propias herramientas de registro de consumo, que al cruzarse puede generar información de gran valor esta creciente tendencia de análisis de datos.

La visión 2.0 del Repositorio Nacional se encamina a la adaptación y explotación de esta plataforma para exponer libremente estos datos y estadísticas sobre la información que se está consultando, las instituciones que están haciendo mayor uso de las plataformas a nivel de depósito y consultas, así como la generación de perfiles de usuarios que utilicen esta plataforma centralizada para su mejoramiento y adecuación a futuro, considerando también las nuevas tendencias tecnológicas y de preservación de los documentos.

Los datos primarios de las investigaciones juegan un papel importante en la potencialización de los repositorios y del desarrollo científico, académico y de innovación en México. La muestra está en la diversidad de repositorios de datos que actualmente interoperan o están en proceso de lograrlo, como son: de observaciones geoespaciales, de ciencias del mar y limnología, de especies de maíz y trigo, de genomas, especies naturales y de sismicidad en México, entre otros.

Actualmente, se encuentran interoperando más de setecientos conjuntos de datos en el Repositorio Nacional. Si bien el porcentaje relativo con el poblamiento total aún es bajo, se estima que esta cifra se multiplicará con la interoperabilidad de los repositorios que se encuentran en proceso de ser cosechados, siendo los datos primarios de las investigaciones el recurso de información que mayores números aporta al Repositorio Nacional.

La continuidad de la publicación de convocatorias y el otorgamiento de los recursos económicos son importantes para apoyar a las instituciones para desarrollar y fortalecer no solo las plataformas, sino sus capacidades, recursos

humanos y el *know how* para generar nuevos especialistas en materia de repositorios, que seguirán innovando con la implementación de nuevas herramientas, infraestructuras o conceptos a la escena de los repositorios institucionales, como ya ha sucedido desde hace dos años, cuando se lanzó la primera versión del Repositorio Nacional.

Bibliografía

Breeding, Marshall. 2002. "Understanding the Protocol for Metadata Harvesting of the Open Archives Initiative", *Computers in Libraries*, vol. 22, núm. 8. Disponible en https://librarytechnology.org/document/9944.

CONACYT. 2017a. *Informe general del estado de la ciencia, la tecnología y la innovación 2016*. México: Consejo Nacional de Ciencia y Tecnología. Disponible en http://www.siicyt.gob.mx/index.php/estadisticas/informe-general/informe-general-2016/3835-informe-general-2016/file.

___ 2017b. *Lineamientos Generales de Ciencia Abierta*. México: Sistema Integrado de Información sobre Investigación Científica, Desarrollo Tecnológico e Innovación. Disponible en http://www.siicyt.gob.mx/index.php/normatividad/conacyt-normatividad/programas-vigentes-normatividad/lineamientos/lineamientos-generales-de-ciencia-abierta.

___ 2017c. *Lineamientos Jurídicos de Ciencia Abierta*. México: Sistema Integrado de Información sobre Investigación Científica, Desarrollo Tecnológico e Innovación. Disponible en http://www.siicyt.gob.mx/index.php/normatividad/conacyt-normatividad/programas-vigentes-normatividad/lineamientos/lineamientos-juridicos-de-ciencia-abierta.

___ 2017d. *Lineamientos Específicos para Repositorios*. México: Sistema Integrado de Información sobre Investi-

gación Científica, Desarrollo Tecnológico e Innovación. Disponible en http://www.siicyt.gob.mx/index.php/normatividad/conacyt-normatividad/programas-vigentes-normatividad/lineamientos/lineamientos-especificos-para-repositorios.

FORCE11. *The FAIR data principles*. Disponible en https://www.force11.org/group/fairgroup/fairprinciples.

Greenberg, J., H. C.White, S. Carrier y R. Scherle . "A metadata best practice for a scientific data repository", *Journal of Library Metadata*, vol. 9, núm 3–4 (2009): 194-212. https://doi.org/10.1080/19386380903405090

Iniciativa de Budapest para el Acceso Abierto. 2002. https://www.budapestopenaccessinitiative.org/translation/spanish-translation.

Ley de Ciencia y Tecnología. 2015. Capítulo X Del Acceso Abierto, Acceso a la Información Científica, Tecnológica y de Innovación y del Repositorio Nacional. Capítulo adicionado DOF 20-05-2014. México: Cámara de Diputados del H. Congreso de la Unión. Disponible en http://www.diputados.gob.mx/LeyesBiblio/pdf/242_081215.pdf.

Oliphant, Tami. 2017. "A case for critical data studies in library and information studies", *Journal of Critical Library and Information Studies*, vol. 1, núm. 1.

Park, Jung-Ran. 2009. "Metadata Quality in Digital Repositories: A Survey of the State of the Art", *Cataloging & Classification Quarterly*, vol. 47.

The University of Melbourne. *Melbourne Policy Library. Management of Research Data and Records Policy* (MPF1242). Disponible en https://policy.unimelb.edu.au/.