ANDREA SIXTO-COSTOYA, RAFAEL ALEIXANDRE-BENAVENT, ANTONIO VIDAL-INFER, RUT LUCAS-DOMÍNGUEZ, LOURDES CASTELLÓ-COGOLLOS

Data sharing:
qué son y cómo se pueden compartir los datos de investigación. Manual de recomendación para gestores de la información

7

DOCUMENTOS de TRABAJO nº7

Diciembre 2019 · 1ª Revisión

SEDIC

DATA SHARING: QUÉ SON Y CÓMO SE PUEDEN COMPARTIR LOS DATOS DE INVESTIGACIÓN. MANUAL DE RECOMENDACIÓN PARA GESTORES DE LA INFORMACIÓN.

ANDREA SIXTO-COSTOYA, RAFAEL ALEIXANDRE-BENAVENT, ANTONIO VIDAL-INFER, RUT LUCAS-DOMÍNGUEZ, LOURDES CASTELLÓ-COGOLLOS

COMITÉ EDITORIAL
BLANCA SAN JOSÉ MONTANO
CARMEN MORALES SANABRIA

EDITORIAL
Sedic. Sociedad Española de Documentación e Información Científica
www.sedic.es

AUTORES
ANDREA SIXTO-COSTOYA[1], RAFAEL ALEIXANDRE-BENAVENT[2], ANTONIO VIDAL-INFER[3], RUT LUCAS-DOMÍNGUEZ[4], LOURDES CASTELLÓ-COGOLLOS[5]

DISEÑO DE PORTADA
JULIO IGUALADOR

DISEÑO
MARTA PONS

PATROCINADO POR
GOBIERNO DE ESPAÑA
MINISTERIO DE CULTURA Y DEPORTE

ISBN 978-84-09-17458-4

1. Departamento de Historia de la Ciencia y Documentación, Facultad de Medicina y Odontología, Universitat de València, Unidad de Investigación Social y Sanitaria (UISYS- uisys.es). ORCID: orcid.org/0000-0001-9162-8992 / Correo: andrea.sixto@uv.es

2. Ingenio (CSIC-Universitat Politècnica de València). Unidad de Investigación Social y Sanitaria (UISYS- uisys.es) ORCID: orcid.org/0000-0002-6678-8844 / Correo: rafael.aleixandre@uv.es

3. Departamento de Historia de la Ciencia y Documentación, Facultad de Medicina y Odontología, Universitat de València. Unidad de Investigación Social y Sanitaria (UISYS- uisys.es). ORCID: orcid.org/0000-0002-4697-7832 / Correo: rut.lucas@uv.es

4. Departamento de Historia de la Ciencia y Documentación, Facultad de Medicina y Odontología, Universitat de València. Unidad de Investigación Social y Sanitaria (UISYS- uisys.es). ORCID: orcid.org/0000-0002-7860-8652 / Correo: antonio.vidal-infer@uv.es

5. Departamento de Sociología y Antropología Social, Facultad de Ciencias Sociales, Universitat de València. Unidad de Investigación Social y Sanitaria (UISYS- uisys.es). ORCID: orcid.org/0000-0002-0305-3154 / Correo: lourdes.castello@uv.es

ÍNDICE

1. DEFINICIONES BÁSICAS

1.1 USO COMPARTIDO DE DATOS

El uso compartido de datos, data sharing en inglés, es la acción de compartir con el resto de la comunidad científica el material sin procesar generado durante el curso de la investigación que sirve para extraer y validar resultados (Torres-Salinas, 2010). Actualmente, el data sharing se engloba dentro de la filosofía de acceso abierto, open access en inglés (en adelante, OA), entendiéndose el compartir datos como una práctica que favorece que la ciencia sea más abierta y accesible (European Commission, 2016). Desde esta perspectiva de OA, el datasharing promueve que los datos sin procesar puedan tener una "segunda vida" y se puedan utilizar más allá del fin para el que fueron generados en un principio. Estos usos pueden ir desde la reutilización de los datos para producir nuevos estudios, a servir como verificadores de resultados de investigación.

Desde una perspectiva histórica, el uso compartido de datos es una práctica entre investigadores que tiene mucho recorrido, tal y como se puede comprobar en el trabajo de Stanley y Stanley (1988), donde hace referencia a que compartir datos en una relación de colaboración tiene una larga trayectoria en ciencia. En una línea parecida, un artículo publicado por la revista Nature Communications (Nature, 2018) refiere que hay áreas de la ciencia con más tradición de compartir datos que otras. Por ejemplo, mientras que en ciencias sociales y biológicas es una tendencia creciente, hay disciplinas como la economía o la meteorología donde compartir datos ha sido la norma desde hace más de un siglo. Sin embargo, actualmente el data sharing es una práctica que tiene entre sus principales objetivos ir más allá del intercambio entre colegas que ya se conocen y extender la práctica a todas las disciplinas. El cambio importante en la concepción de la práctica de compartir datos como la describían Stanley y Stanley en sus inicios viene dado, como se adelantaba antes, por la irrupción del movimiento por el OA. Con este movimiento, la práctica de compartir datos adquiere una nueva dimensión cada vez más consolidada a medida que la comunidad de investigadores y relevantes instituciones científicas la reconocen como útil. Este hecho quedó cristalizado en importantes declaraciones como la de Berlín (Sociedad Max Planck, 2003), en definiciones como la que sigue:

> *"Para establecer el acceso abierto como un procedimiento meritorio, se requiere idealmente el compromiso activo de todos y cada uno de quienes producen conocimiento científico y mantienen el patrimonio cultural. Las contribuciones del acceso abierto incluyen los resultados de la investigación científica original, datos primarios y metadatos, materiales, fuentes, representaciones digitales de materiales gráficos y pictóricos, y materiales eruditos en multimedia."*

En el documento de una institución de gran envergadura como es la Organización para la Cooperación y el Desarrollo Económicos (OCDE) (Melero, 2005), que en el año 2004 promueve el compromiso firmado por 34 países, donde se relatan una serie de afirmaciones sobre las ventajas del acceso abierto a los datos de investigación y el acuerdo de tenerlas en cuenta para fomentar lo máximo posible la transparencia, la interoperatibidad y la eficiencia (entre otras) en la ciencia. A modo de ejemplo, el primer reconocimiento de una lista de 8, es el siguiente:

> *"Recognising that an optimum international exchange of data, information and knowledge contributes decisively to the advancement of scientific research and innovation."*

Además de la Declaración de Berlín y de documentos como el de la OCDE, otras instituciones de alto nivel, como la US Academy of Sciences, Engineering and Medicine, también se suman a la petición de una ciencia más abierta, incluyendo expresamente a los datos en esta concepción. En general, son muchas las agencias financiadas por gobiernos, por ejemplo, de EEUU, Australia o de países europeos, que requieren a los autores que realicen planes de gestión de datos, y, siempre que sea posible, que los pongan a disposición del público en general.

Esta perspectiva del uso compartido de datos dentro del espectro del el OA y del Open Science (OS), que entiende que los datos deben ver la luz para ser utilizados más allá del fin para el que fueron concebidos y estaría ligado a la reutilización y a la transparencia, recibe el nombre de datos abiertos e investigación, open data en inglés (Peset et al., 2017). En este documento se hablará del data sharing en este contexto de OA y de open data, que se irá desgranando en sus múltiples aspectos.

1.2 DATOS BRUTOS DE INVESTIGACIÓN

Una vez conceptualizado el data sharing, en este apartado se proporcionan una serie de definiciones sobre los datos brutos de investigación aportadas por organismos internacionales. Aunque existen varias definiciones de qué son datos de investigación, seleccionamos tres que son ilustrativas. La primera, ofrecida por la Office of Management and Budget (Office Management and Budget, 1999) de EEUU, representa el punto de vista gubernamental, ya que se estableció con el propósito de uniformizar los requerimientos administrativos para acuerdos y subvenciones con instituciones universitarias, hospitales y organizaciones sin ánimo de lucro. Esta definición de los datos dice lo siguiente:

> *"Los datos de investigación se definen como aquel material registrado comúnmente aceptado por la comunidad científica como necesario para validar resultados de investigación. No serían datos de investigación ni los análisis preliminares, borradores de la elaboración de artículos, planes de investigaciones futuras, revisiones por pares ni comunicaciones entre colegas."*

Otra definición, dada por Borgman (Borgman, 2008) desde una perspectiva más centrada en la investigación, se refiere al concepto de dato bruto como:

> *"A reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing. Examples of data include a sequence of bits, a table of numbers, the characters on a page, the recording of sounds made by a person speaking, or a moon rock specimen."*

La tercera, facilitada desde el ámbito de la Unión Europea (en adelante, UE) y más orientada a los proyectos que financia (European Commission, 2019), define los datos de la siguiente forma:

> *"In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images. The focus is on research data that is available in digital form."*

De estas tres definiciones, se deduce que los datos de investigación son lo que se considera el "bruto", los primeros datos que se extraen de las investigaciones y que todavía no han sido tratados. De hecho, es muy frecuente que la comunidad científica se refiera a esos datos con la denominación de "datos brutos", por su traducción del inglés de "raw data".

Los datos brutos de investigación son, de hecho, un tipo de datos que existen desde que se empieza a concebir la investigación y la comunicación científica tal y como se entiende hoy. Se necesita generar datos en la medida en que desea obtener resultados que ofrezcan respuestas a las preguntas de investigación. La cuestión es que, la mayoría de las veces, no se utilizan el cien por cien de los datos brutos que se genera, sino que se seleccionan y analizan los que más interesan para los fines de determinado estudio y obtener así resultados. De ahí la utilidad del data sharing, práctica que permite dar a los datos una segunda vida. No obstante, desde que el impulso para el uso compartido de datos comienza a tomar fuerza a mediados de los 2000, se ha ido desarrollando toda una estructura que engloba desde los protocolos de acción hasta plataformas digitales diseñadas para publicación, almacenamiento y preservación de los datos. En definitiva, un conjunto de mecanismos que hacen de esta práctica algo realmente útil. Los siguientes apartados se centrarán en explicar los aspectos más relevantes de esta estructura que se encuentra, además, en continuo crecimiento y perfeccionamiento.

1.3 PRINCIPIOS FAIR

Una vez encuadrados los conceptos de "datos brutos" y "data sharing", es importante hacer una introducción en lo que se denominan los principios FAIR. Estos cuatro principios, cada uno representado en las siglas FAIR (Findable, Accessible, Interoperable, Reusable), son los que rigen cómo deben de ser los datos una vez se decide que sean compartidos. La realidad es que, si alguno de estos principios no se cumple correctamente, el hecho de compartir los datos va a perder su potencial. El que las siglas estén ordenadas de esta forma concreta no parece casual, ya que responde a una secuencia lógica. Lo primero, que los datos deben ser fáciles de encontrar (findable), si no, poco sentido tiene compartirlos; lo segundo, que además de encontrarlos, se puedan obtener (accessible); lo tercero, que una vez los se encuentran y obtienen, sea posible comprenderlos y trabajar con ellos (interoperable); y cuarto y último, que sean reutilizables (reusable), para que una vez

encontrados, obtenidos, comprendidos y trabajado con ellos, también se puedan reutilizar para otros fines estando clara su procedencia y las condiciones de reúso.

En el artículo de Wilkinson et al., (2016) definen las características de cada principio de la siguiente forma:

Para ser "Findable":

1. Los datos y metadatos deben tener asignado un DOI (Digital Object Identifier).
2. Los datos tienen que estar descritos con suficientes metadatos.
3. Los metadatos claramente tienen que poder identificar a los datos que describen.
4. Los datos y metadatos deben estar registrados o indexados en un recurso que sea fácil de localizar.

Para ser "Accessible":

1. Los datos y metadatos pueden ser recuperables con su identificador, utilizando un protocolo estandarizado de comunicación.
    1.1 El protocolo debe ser abierto, gratis y universalmente implementable.
    1.2 El protocolo debe permitir, cuando sea necesario, un proceso de autentificación y autorización.
2. Los metadatos deben ser accesibles, incluso cuando los datos ya no están a disposición.

Para ser "Interoperable":

1. Los datos y metadatos deben utilizar un lenguaje formal, accesible, compartible y ampliamente aplicable que sea útil para la representación del conocimiento.
2. Los datos y metadatos deben utilizar vocabularios que sigan los principios FAIR.
3. Los datos y metadatos pueden incluir referencias a otros datos o metadatos.

Para ser "Reusable":

1. Los datos y metadatos deben ser descritos de manera rica, con una pluralidad de atributos precisos y relevantes.
    1.1 Los datos y metadatos deben publicarse con una licencia sobre su uso y reutilización clara y accesible.
    1.2 Los datos y metadatos tienen que estar asociados con una información sobre su procedencia detallada.
    1.3 Los datos y metadatos deben seguir los estándares que utiliza la comunidad del dominio concreto que se esté usando.

Para finalizar este apartado, es interesante recalcar que los principios FAIR, pensados para establecer un marco para el uso compartido de los datos en abierto, no implican que absolutamente todos los datos deban de tener el mismo nivel de apertura ni que todos los datos de todas las disciplinas sigan las mismas reglas. Se trata más bien de un cambio de paradigma, que persigue el objetivo de que los datos estén abiertos por defecto en vez de cerrados, como venía sucediendo hasta ahora. Esto se ve muy claramente con el lema de los principios FAIR "*tan abiertos como sea posible, tan cerrados como sea necesario*" (Henning et al., 2019).

2. POLÍTICAS SOBRE DATOS ABIERTOS: INTERNACIONAL, UE Y ESPAÑA

En el siguiente apartado se realiza una aproximación a las políticas sobre datos abiertos vigentes en la actualidad en tres contextos: el internacional, para obtener una vista aérea de las iniciativas que se están llevado a cabo a nivel mundial; el de la Unión Europea, por ser el marco que define la investigación que se lleva a cabo en los Estados miembros; y España, ya que es el contexto más cercano.

2.1 INTERNACIONAL

A nivel internacional, se desatacan en este trabajo cuatro contextos que sirven como referencia sobre las tendencias de uso compartido de datos de investigación en diferentes lugares. Se tomarán como referencia: Australia, EEUU, América Latina y Canadá.

En primer lugar, el caso de Australia es el de un país con una trayectoria bastante amplia en los datos abiertos de investigación y su uso compartido. Prueba de ello es la Australian National Data Service (ANDS), una asociación liderada por la Monash University en colaboración con la Australian National University y la Commonwealth Scientific and Industrial Research Organisation y financiada por el National Collaborative Research Infrastructure Strategy, que está vigente desde el año 2008. El propósito principal de la ANDS es hacer que los datos de investigación en Australia adquieran más valor para los científicos, las instituciones de investigación y la nación en general. Según refieren, desde el año 2008 han apoyado numerosos proyectos de investigación en Australia y han jugado también un rol importante a nivel internacional. Entre otros servicios, destacan como estrella el Research Data Australia, un repositorio donde se puede encontrar, acceder y reutilizar datos de investigación provenientes de instituciones australianas de investigación, agencias gubernamentales y asociaciones culturales (Australian National Data Service, s.f).

Por otra parte, en el contexto de América Latina, el proyecto LEARN, del que la Comisión Económica para América Latina y el Caribe (CEPAL) es una de las instituciones socias, realiza un trabajo para observar de qué manera los países e instituciones de esta región están participando en la gestión los datos de investigación. Dentro de las iniciativas que se han detectado están las relacionadas con nueva legislación, el desarrollo de políticas en las agencias de financiación y la implementación de repositorios. Por ejemplo, en el plano legal países como Argentina y Perú han desarrollado leyes según las que los investigadores deben cumplir una serie de requisitos con respecto a los datos cuando se trate de investigaciones financiadas con fondos públicos; en esta línea, los cambios legislativos trajeron consigo el desarrollo de infraestructuras como el Sistema Nacional de Repositorios Digitales en el caso de Argentina, y el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación en Perú. Por otro lado, el Consejo Nacional de Ciencia y tecnología (CONICYT) de Chile, país que todavía no tiene una legislación en este sentido, tiene en marcha una propuesta para crear una política de acceso y preservación de la información y datos de investigación también financiada con fondos públicos. Otra iniciativa interesante es la que lleva a cabo el servicio de preservación digital Rede Cariniana de Brasil, cuyas instituciones miembros tienen hacen uso de Dataverse, un repositorio multidisciplinar de datos que se explicará más adelante (Andaur, 2016).

En el caso de EEUU, coexisten diversas iniciativas relacionadas con el uso compartido de datos en investigación. A modo de ejemplo, se explica el caso de los National Institute of Health (NIH), por ser la mayor institución financiadora de investigación en Ciencias de la Salud a nivel mundial. Aunque todavía no tiene un mandato para que las investigaciones financiadas con sus fondos deban poner los datos en abierto de manera obligatoria, como sucede con las publicaciones finales, ha trazado un camino en este sentido. Prueba de ello es el plan que lanza en el año 2015 titulado *"National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research"*, en el que se preparan las bases para que los datos de investigación sean compartidos. De este modo, pone en valor tanto los datos en sí como su potencial cuando son compartidos (National Institute of Health, 2015). Este plan, también conocido como NIH Data Commons Pilot Phase, deja entrever que una política de acceso abierto a los datos por parte de esta entidad puede estar cerca de ver la luz (National Institute of Health, 2018).

Por último, también relacionado con las Ciencias de la Salud, se encuentra el caso de Canadá. Este país, que va un paso más allá del NIH al tener ya aprobada una política de depósito y publicación de los datos. Esta institución requiere a sus investigadores que, después de que hayan publicado sus resultados, los datos sean depositados, siempre que sea posible, en repositorios abiertos de datos y que, además, se deben de conservar durante un mínimo de cinco años, tanto si han podido ser publicados como si no (Canadian Institutes of Health Research, 2015).

2.2 UNIÓN EUROPEA

En lo que se refiere a la UE y su postura sobre el uso compartido de datos, la tendencia es clara: se defiende y se promueve, aunque de momento no es un mandato obligatorio. Para examinar esta postura se toma como referencia las bases que estipula la Comisión para los proyectos Horizon 2020 (H2020), ya que es en ellos donde la UE engloba la mayoría de su actividad investigadora. Los proyectos H2020 comprenden el período que va del año 2014 al 2020 y se centran en tres pilares: la excelencia científica en Europa, el liderazgo industrial y los retos sociales a los que se enfrenta la Unión.

Básicamente, lo que desde la UE se indica sobre uso compartido de datos sigue la misma tendencia que sobre el OA en general: debe hacerse por sistema y por defecto. En el caso de las publicaciones finales de los artículos, esto significa que debe hacerse siempre por alguna de las dos opciones disponibles, que son la vía verde o dorada[1], y solo en el caso de los estudios que implican patentes pueden guardarse el derecho de no publicar en abierto. Los razonamientos que defienden esta apertura son varios, sobre todo relacionados con la acelerar la ciencia y fomentar la colaboración. Pero uno de los argumentos más destacados por su lógica y contundencia es el siguiente: no debería ser necesario pagar por la información financiada con fondos públicos cada vez que se accede o se utiliza la misma (European Commission, 2019).

Como ya se mencionaba, la línea seguida en cuanto al fomento de las publicaciones en OA y la apertura de los datos es similar. Lo interesante, en este caso, está en los matices que las diferencian. La definición ofrecida por el European Research Council sobre los datos de investigación (2017), dice que:

> *"The European Research Council supports the basic principle of Open Access to research data. It therefore recommends to all its funded researchers that they follow best practice by retaining files of all the research data they have produced and used during the course of their work, and that they be prepared to share these data with other researchers whenever they are not bound by copyright restrictions, confidentiality requirements, or contractual clauses."*

En esta definición se observa que el aspecto diferenciador entre el acceso abierto a las publicaciones, obligatorio en vía verde o dorada a excepción de las investigaciones que impliquen patentes, y el acceso abierto a los datos está en la siguiente frase: *"it therefore recommends to all its funded (…)"*. Recomienda, pero no obliga. Para los beneficiarios que deseen optar por poner sus datos en abierto, la Comisión ofrece una serie de directrices recogidas en el "Open Research Data Pilot" (ORD Pilot), que se empieza a poner en marcha en el año 2017, tres años después del período que abarcan los proyectos H2020. Según estas directrices, los trabajos financiados hasta el 2016 podrán acogerse al ORD Pilot voluntariamente, mientras que los aprobados a partir del 2017 lo tendrán incluido por defecto. A pesar de esto, con respecto a los proyectos a partir del 2017 también se indica que, aunque compartir los datos es la opción por defecto, se entiende que no todos los datos pueden compartirse siempre. Comprobamos, pues, que el ORD Pilot gira en torno al lema de los principios FAIR de "tan abierto como sea posible, tan cerrado como sea necesario". Los casos en los que los investigadores puede optar por no compartir, que son bastante amplios, son los siguientes:

1. Cuando es incompatible con la obligación de proteger resultados que pueden ser potencialmente utilizados para su explotación comercial o industrial.
2. Cuando es incompatible con la necesidad de confidencialidad y seguridad.
3. Cuando es incompatible con la protección de datos personales.
4. Cuando el hecho de compartir datos pueda suponer que el principal objetivo del proyecto podría no ser conseguido.
5. Cuando el proyecto no genera ni recoge datos.
6. Cuando hay otras razones legítimas, que pueden ser descritas en un espacio pensado para este fin.

Como conclusión de este apartado, se destaca que la UE efectivamente muestra una predisposición hacia la apertura de datos que podría estar mostrando una tendencia de la que habrá que estar pendiente de cara a la siguiente remesa de proyectos, que se llamarán Horizon Europe, y que serán lanzados en el año 2021. En el siguiente apartado se verá qué sucede, a grandes rasgos, en los países europeos de modo más individual y, más concretamente, qué sucede en España.

1. La vía dorada es aquella por la cual los autores optan porque su artículo tenga un acceso público, inmediato, gratuito y permanente, normalmente pagando una tasa a la editorial conocida como APC (Article Publishing Charge). La vía verde es la opción que pueden escoger los autores de que su artículo sea depositado normalmente en un repositorio digital una vez sido aceptado por una revista, de manera que, pasado un posible periodo de embargo, el artículo tenga un acceso público (Elsevier Connect, 2019).

2.3 POLÍTICAS A NIVEL DE PAÍSES MIEMBROS DE LA UE

Para hablar de las políticas a nivel de países de la UE, tomaremos como referencia el análisis publicado por SPARC Europe y el Digital Curation Center (DCC), ya que es muy reciente (se publica en agosto de 2019) e ilustra muy bien cuál es el panorama actual en cuanto a políticas relacionadas con el uso compartido de datos en los distintos Estados Miembros. Además, tiene la ventaja de ser un documento dinámico, es decir, que se va actualizando cada cierto tiempo, por lo que es una buena herramienta para estar al día (SPARC Europe y Digital Curation Center, 2019).

Según este reciente análisis, 14 de 28 países tienen políticas vigentes relacionadas con los datos abiertos de investigación. ORD Pilot, comentado en el apartado anterior, es mencionado en muchas de las políticas. Pero además de hacer mención a estas directrices, el análisis menciona una importante ley que se ha aprobado recientemente en el marco de la UE y que se prevé que empezará a tomar fuerza a partir del 2019 (Diario Oficina de la Unión Europea, 2019). Se trata de medidas encaminadas a promover el acceso abierto a los datos, no solo de investigación, sino también los datos generados por sectores empresariales o gubernamentales. Es una ley que afecta a todos los estados miembros, por lo que se espera que en los próximos dos años haya acciones en todos ellos en este sentido. Para entender las implicaciones de esta ley, es interesante un fragmento del artículo 4 de la misma:

> *"Los cambios de fondo introducidos en el texto legislativo con el fin de explotar plenamente el potencial de la información del sector público para la economía y la sociedad europeas se centran en los siguientes aspectos: la prestación de acceso en tiempo real a los datos dinámicos a través de medios técnicos adecuados, aumentando el suministro de datos públicos valiosos para la reutilización, incluidos los de las empresas públicas, organizaciones que financian la investigación y organizaciones que realizan actividades de investigación (…)".*

O este otro fragmento del artículo 27:

> *"(…) El acceso abierto mejora la calidad, reduce la necesidad de duplicaciones innecesarias en la investigación, acelera el progreso científico, combate el fraude científico y puede favorecer de manera general el crecimiento económico y la innovación. Además del acceso abierto, es encomiable que se esté procurando garantizar que la planificación de la gestión de datos se convierta en una práctica científica estándar y apoyar la divulgación de datos de investigación que sean fáciles de encontrar, accesibles, interoperables y reutilizables (principios FAIR)."*

A día de hoy, siguiendo también el análisis de SPARC Europe y el DDC, a nivel nacional el período de años en los que estas políticas entran en vigor va desde 2009 a 2019, pero mostrando una clara tendencia hacia los años más recientes. En el análisis detectan varios aspectos de importancia e interés. Uno de ellos es que hay dos variables que marcan la diferencia entre países. Una variable es que, mientras hay países en los que la política sobre los datos abiertos está relacionada con el Open Access en general, hay otros que la trabajan de manera paralela. La otra variable es sobre la dureza con la que estas políticas se quieren implantar, que va desde una implantación "suave" (a base de promover y alentar) a otra "dura" (es un imperativo, una obligación). Otro aspecto de interés es que, tanto si promueven como si obligan, los mecanismos para monitorear el cumplimiento y para ofrecer refuerzos en forma de recompensas es en general escaso en todos los países.

2.4 MARCO LEGAL EN ESPAÑA

En lo que se refiere a españa, se encuentra entre los catorce de los veintiocho que tiene algún tipo de plan o normativa referente a los datos de investigación. Si se siguen las variables que se mencionaban en el apartado anterior, se podría decir que la normativa que recoge los datos abiertos está incluida en la temática OA, y que sería de las que se refieren a la aplicación del uso compartido de datos con poca dureza, optando más por recomendar y siempre de manera opcional. El documento estatal en el que se refiere a todo este tema es el Plan Estatal de Investigación Científica y Tecnológica y de Innovación, que comprende desde el año 2017 al 2020 (Ministerio de Industria, Economía y Competividad, 2017).

En el documento se mencionan los datos de investigación varias veces para reseñar la importancia de que sean compartidos. Sin embargo, la parte en la que realmente podemos entender qué tratamiento se espera que los investigadores hagan de ellos está en el punto 5.2, apartado 2 (página 30), cuando dice que:

> *"Con el fin de impulsar el acceso a datos de investigación, los proyectos de I+D+i financiados podrán incluir,* ***con carácter optativo****, un plan de gestión de los datos de investigación que se depositarán en repositorios institucionales, nacionales y/o internacionales tras la finalización del proyecto y trascurrido el plazo establecido en las correspondientes convocatorias. No obstante, se respetarán todas las situaciones en las que los mismos han de protegerse por razones de confidencialidad, seguridad, protección, etc. o cuando los mismos sean necesarios para la explotación comercial de los resultados obtenidos. Finalmente, en la evaluación curricular de los investigadores, así como en la evaluación ex post de las actuaciones financiadas* ***se tendrán en cuenta los trabajos publicados en abierto en repositorios*** *institucionales y temáticos, nacionales y/o internacionales,* ***y la puesta de los datos de su investigación en abierto****, de modo que puedan ser utilizados para replicar y reproducir los análisis y resultados de investigación."*

De este párrafo se extraen dos ideas principales que afectan a la investigación financiada con fondos públicos en España con respecto a los datos de investigación: incluir un plan de gestión de los datos es optativo. Además, aun siendo explícitamente optativo, mencionan el derecho a abstenerse en las situaciones en las que los datos abiertos puedan causar algún problema relacionado con confidencialidad o temas comerciales. Teniendo en cuenta que el depósito de datos no se incluye por defecto, como sucede a partir del año 2017 en los proyectos H2020 con el ODR Pilot, resulta redundante la anotación de "*se respetarán todas las situaciones en las que deban protegerse los datos (..)*". Es redundante porque ya es algo optativo. Por otro lado, es interesante resaltar la mención que se hace a que cuando los datos sean compartidos, se tendrá en cuenta para la evaluación curricular de los investigadores. Aunque se echa de menos saber en qué aspectos concretamente se tendrán en cuenta, lo importante es que se haga alguna referencia a posibles incentivos para compartir.

Finalmente, a modo de conclusión de este apartado, mencionar que, aunque el acercamiento de España con respecto al uso compartido de datos en ciencia es todavía tímido y algo ambiguo, la línea está trazada. El hecho de que al menos exista una mención a su importancia, aunque sea totalmente opcional, abre la puerta a que en un futuro se pueda llegar al "se comparte por defecto". Este plan caduca en el año 2020, por lo que pronto se podrá comprobar por donde irá la tendencia.

3. TECNOLOGÍAS PARA COMPARTIR DATOS

Una vez contextualizado el uso compartido de datos y las políticas que se han establecido con respecto a ellos a nivel de Europa y de España, este apartado se centra en las tecnologías existentes a día de hoy que hacen que el uso compartido de datos pueda pasar de la teoría a la práctica.

Cuando se habla de tecnologías para compartir datos, se refiere a la infraestructura existente que permite a los usuarios subir, almacenar, preservar, buscar y descargar datos de investigación para darles vida más allá del fin para el que fueron creados. Estas infraestructuras digitales siempre implican el acceso a internet, ya que, como se mencionaba al principio de este documento, el uso compartido de datos como hoy se concibe es posible porque los medios digitales y la red lo permiten.

Dentro de las infraestructuras, los repositorios de datos y las plataformas de las editoriales son actualmente la máxima referencia en cuanto a datos de investigación compartidos. Con diversos mecanismos, ambas proporcionan la posibilidad de que tanto los usuarios que desean depositar, como los que quieran buscar y obtener datos, puedan hacerlo.

Dada su relevancia, en este documento nos focalizaremos en el papel de los repositorios de datos y de las editoriales y revistas como principales infraestructuras para compartir datos brutos de investigación.

3.1 LOS REPOSITORIOS DE DATOS

En realidad, los repositorios de datos son la respuesta "natural" a las demandas o peticiones de diversos sectores, como las instituciones que financian proyectos nacionales o internacionales (como ya se ha visto), o las exigencias de las revistas y las editoriales (que se verá más adelante), a los investigadores para que pongan a disposición de la comunidad científica los datos que sustentan sus trabajos. En su gran mayoría, estos repositorios están pensados para ser libres y gratuitos, permitiendo su acceso y su uso para establecer una especie de intercambio en el que quien hoy deposita, mañana puede buscar y descargar datos de otros y viceversa (Kim y Burns, 2016).

Actualmente, hay un importarte número de repositorios que se fueron creando como respuesta a las necesidades que iban surgiendo. Financiados y mantenidos por diversos organismos, intentan cubrir huecos para que investigadores de cualquier ámbito puedan hacer uso de ellos. Por este motivo, en el momento de escoger un repositorio, es preciso tener en cuenta cuáles son sus características y qué servicios ofrecen. Por ejemplo, es importante saber si son específicos de un área en concreto (repositorios temáticos), si son de tipo generalista y sirven para cualquier área (repositorios multidisciplinares), o si pertenecen a una institución en concreto y sirven para que los miembros de esa institución hagan uso de ellos (repositorios institucionales). Además, es importante conocer otros detalles como la capacidad de almacenamiento que ofrecen, si permiten establecer un periodo de embargo, o si ofrecen la posibilidad de ponerle un DOI, entre otros (Peset y González, 2017).

Aunque normalmente a los investigadores se les indica o aconseja, por ejemplo, en la normativa de los proyectos o las normas de las revistas, en qué repositorio pueden publicar sus datos, es interesante que sepan qué es lo que necesitan de un repositorio y qué es lo que el repositorio necesita de ellos, para que el trabajo sea lo más sencillo y rápido posible. En la Tabla 1 se presenta un resumen de los aspectos y condiciones más relevantes que un investigador debe tener en cuenta a la hora de escoger un repositorio. Si bien a día de hoy hay tantos repositorios de datos en el mundo que es imposible hacer referencia a cada uno de manera individual, se podría decir que estas condiciones son comunes a todos.

Lo que el investigador necesita de un repositorio	Lo que el repositorio necesita del investigador (o del usuario que vaya a depositar los datos)
· Saber si acepta datos de su disciplina. · Que sea intuitivo. · Que ponga a disposición de usuarios guías y protocolos sencillos de seguir. · Que, si el investigador lo necesita, ofrezca diferentes grados de apertura y periodo de embargo. · Que garantice la preservación de los datos. · Que garantice mecanismos de seguridad de los datos. · Saber cómo gestiona la propiedad intelectual y los derechos de autor. · Saber si ofrece un identificador persistente.	· Que el investigador se informe previamente de si ese repositorio es adecuado para su trabajo. · Que los datos estén en un formato reutilizable, es decir, libre y abierto para que, dado el caso, otros puedan usarlos y se asegure el cumplimiento de los principios FAIR . · Que los datos vayan acompañados de metadatos, es decir, de un contexto que les dote de significado y los haga comprensibles.

Tabla 1. Aspectos que un investigador o usuario de un repositorio deben tener en cuenta a la hora de decidir la idoneidad del mismo.

A pesar de que, como se decía, es imposible hablar de cada repositorio individualmente, lo que sí se puede hacer es una descripción de los distintos tipos de repositorios que mencionábamos antes: temáticos, multidisciplinares e institucionales. De esto se tratará en el siguiente apartado.

3.2 TIPOS DE REPOSITORIOS DE DATOS

3.2.1 REPOSITORIOS TEMÁTICOS

Los repositorios temáticos son aquellos en los que su contenido está pensado para una temática específica. Entrarían dentro de este grupo tanto un repositorio de una disciplina en concreto (por ejemplo, un repositorio de datos de Sociología), como los que abarcan un grupo de disciplinas de una misma área temática (por ejemplo, un repositorio de datos de Ciencias Sociales, o un repositorio de datos genéticos). También se incluirían dentro de este grupo los que se centran una tipología concreta de datos, como son los repositorios de datos cualitativos, o los repositorios de imágenes obtenidas mediante la técnica de neuroimagen.

A continuación, se pone algún ejemplo de repositorios temáticos de datos. Como ya se ha dicho, hay muchos repositorios temáticos y es imposible citar todos, por lo que se seleccionan tres de ellos que son bastante diferentes entre sí por lo que pueden servir de ejemplo: Gene Expression Omnibus (GEO), Qualitative Data Repository (QDR), y OpenTopography.

· **Gene Expression Omnibus** (https://www.ncbi.nlm.nih.gov/geo/)

Este repositorio, que pertenece al ámbito de las Ciencias de la Salud, es uno de los más utilizados en un campo donde el uso compartido de datos está demostrando ser más efectivo: la Genética. Además de ser efectivo, en este campo también se comprueba que es una práctica frecuente, tal y como refieren Knoppers et al., (2014) cuando dicen que *"the culture of sharing and the development of policies to enable research collaboration are clearly pervasive in genomics research"*.

Gene Expression Omnibus (GEO), es un repositorio creado y financiado por los NIH de EEUU. Se define como un repositorio de acceso público que almacena datos procedentes de la genómica funcional. Además, provee a los usuarios de herramientas de ayuda tanto para hacer consultas como para descargar experimentos y buscar perfiles de expresión génica. Es decir, que ofrece toda la gama, desde el propio servicio de subida y descarga de datos, a explicaciones y guías para poder realizarlo.

En la captura de pantalla (Figura 1), se puede visualizar la página de inicio de GEO. Como se puede comprobar, están detalladas las opciones que nos permiten obtener información sobre cómo funciona el recurso y cómo se puede tanto subir datos como consultar y descargar datos de otros, las herramientas de las que dispone o la opción de navegar por el contenido disponible. En la caja de búsqueda de la derecha se puede buscar un conjunto de datos en concreto si se conoce previamente el número de referencia. En la Figura 2, se observa un ejemplo del depósito de datos de un artículo que ya ha sido publicado y que fue buscado a través del mencionado número de referencia. Tal y como se puede comprobar en la parte de debajo de la captura, se trata de un dataset antecedido de una serie de metadatos que explican su contenido. Los datos han sido compartidos en formato .txt.

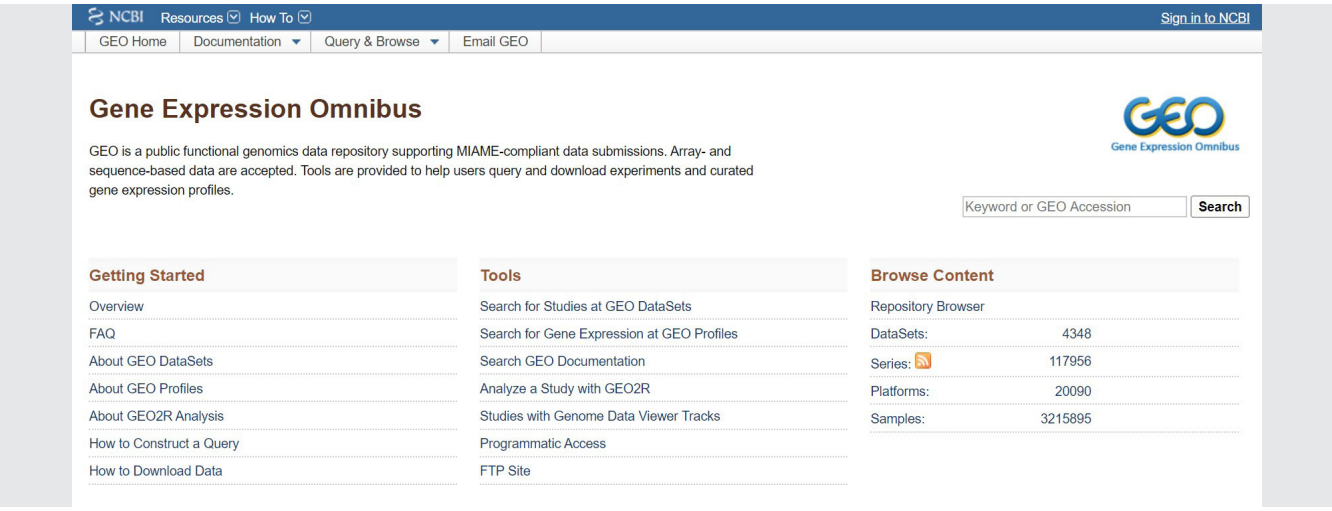

Figura 1. Captura de pantalla de la página principal de GEO.

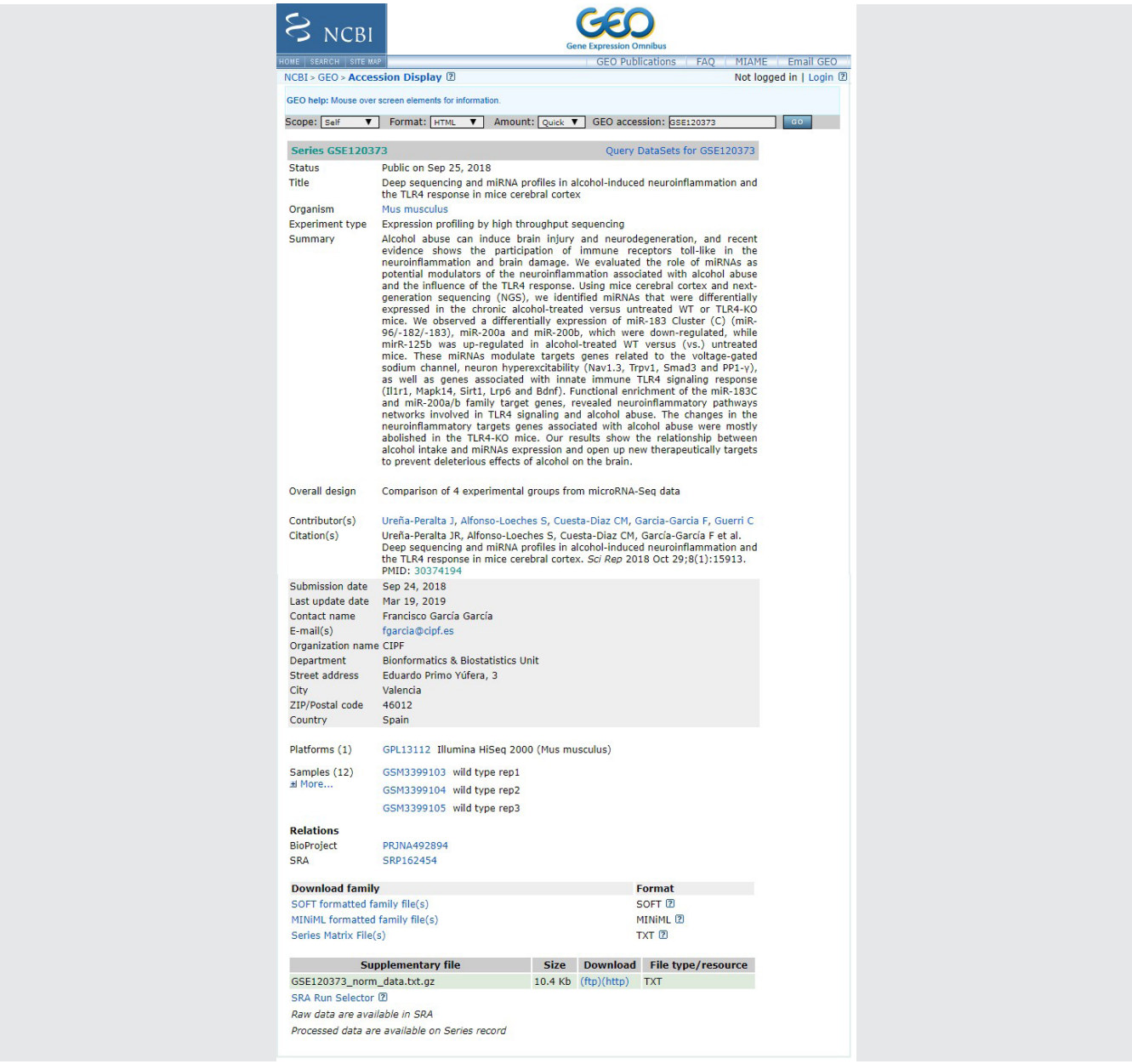

Figura 2. Captura de pantalla de un conjunto de datos depositados en GEO.

· **Qualitative Data Repository** (QDR) (https://qdr.syr.edu/about)

El QDR, como sus siglas indican, se centra en una tipología de datos específica, los datos cualitativos y, además, en un ámbito concreto de la ciencia, las Ciencias Sociales. Por este motivo, se podría decir que tiene una doble concreción por temática, al centrarse en un solo ámbito y en un solo tipo de datos. Se trata de un repositorio financiado por la National Science Foundation y hospedado en el Center for Qualitative and Multi-Method Inquiry, de la Syracuse University (EEUU). Básicamente, permite el almacenamiento y el uso compartido de datos y metadatos generados mediante la metodología cualitativa en disciplinas de las Ciencias Sociales, tales como los derivados de entrevistas semi estructuradas o no estructuradas, grupos focales o notas de campo. Igual que GEO, también ofrece herramientas de búsqueda para facilitar la localización de los datos. Aunque en principio era libre y gratuito, actualmente están cambiando hacia un modelo donde la persona que deposita tendrá que soportar algún coste, aunque todavía no especifica más. En la Figura 3, se observa la pantalla de inicio de la web de QRD. La barra de opciones de la parte superior da información sobre cómo depositar, las normas del repositorio, guías y otras cuestiones relacionadas. La caja de búsqueda de la parte superior derecha da la opción de hacer una búsqueda de términos sobre el tema que se desee y filtrar por "Search Site" o "Search Data". Si se selecciona "Search Data" y se introduce un término, por ejemplo "Young", se observa que existen varios dataset relacionados con investigaciones sobre este tema. En la Figura 4 hay un ejemplo de un conjunto de datos en formato .tsv, .txt y .xls. Relacionados con una publicación en concreto que se encuentran de libre acceso (el QDR da la opción a los autores de restringir o liberar el acceso).

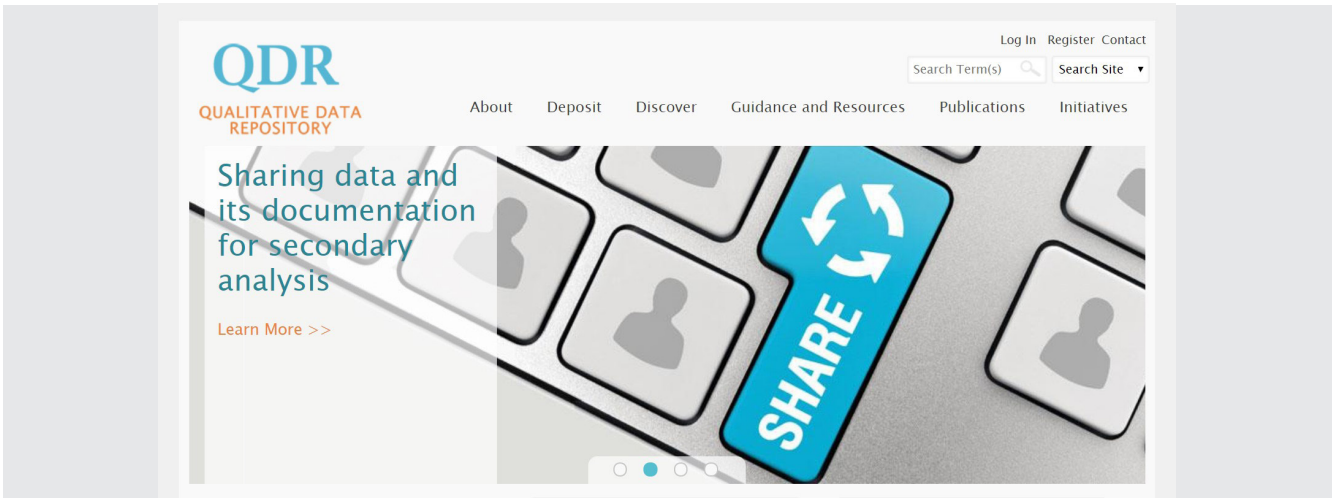

Figura 3. Captura de pantalla de la página principal del repositorio QRD.

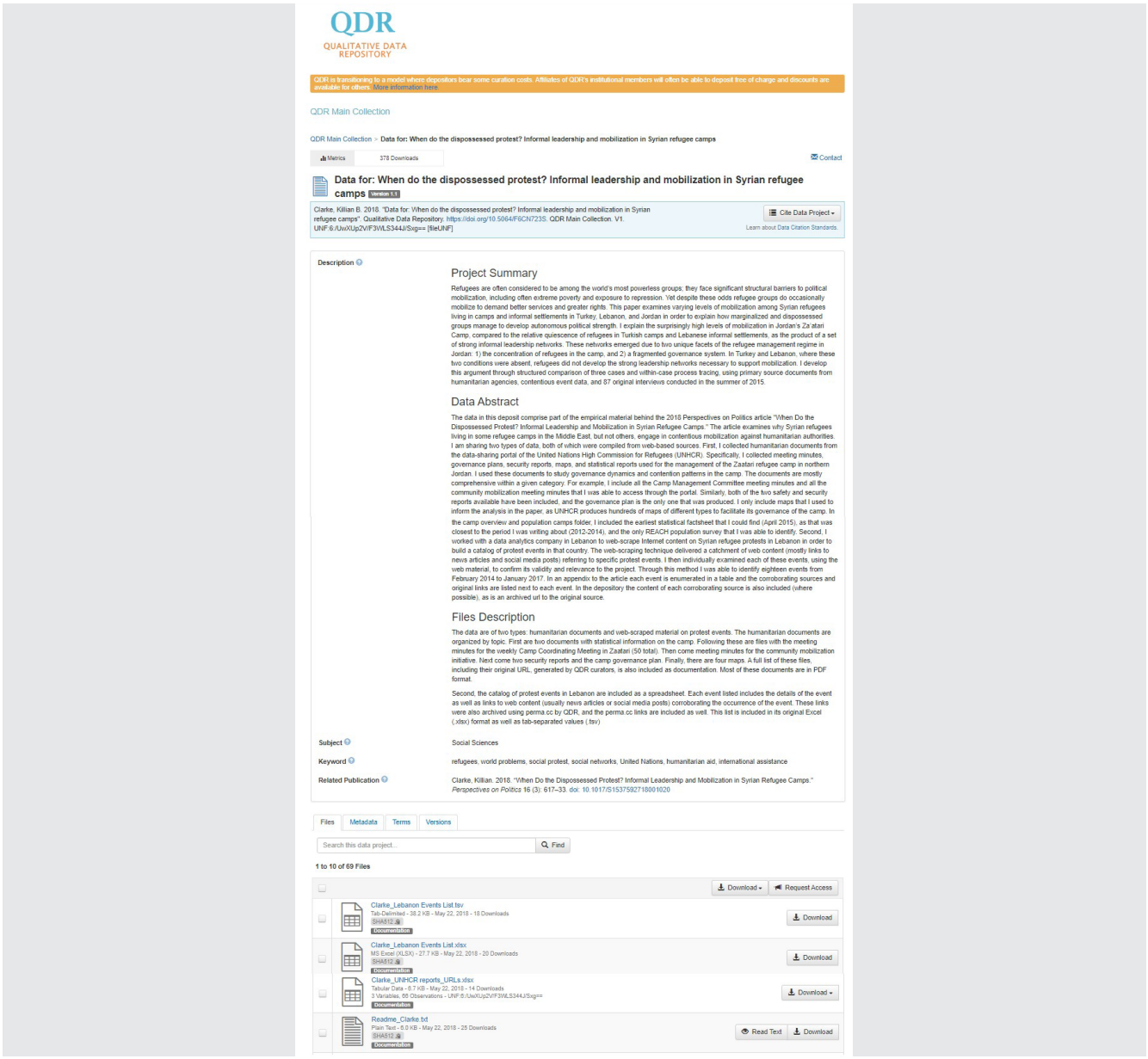

Figura 4. Captura de pantalla del depósito de un conjunto de datos en el repositorio QDR.

· **OpenTopography** (https://opentopography.org/)

En este repositorio de datos, OpenTopography, se ofrece un recurso financiado por la National Science Foundation Earth Sciences, que nace como un proyecto de ciberinfraestructura en Ciencias de la Tierra. El objetivo de este repositorio, según refieren, es democratizar el acceso online a los datos topográficos en alta resolución orientados en temáticas de Ciencias de la Tierra, adquiridos mediante LIDAR u otras tecnologías. Igual que sucede con los otros dos repositorios mencionados, cuenta con una barra de opciones en la parte de arriba que nos indica las herramientas de las que dispone, así como los datos que se pueden compartir. En la esquina superior derecha se encuentra una ventana de búsqueda, donde se puede indagar acerca de algún tema de interés (Figura 5). En la Figura 6, se ve un conjunto de datos sobre una encuesta hecha con la técnica LIDAR en la isla griega de Santorini.

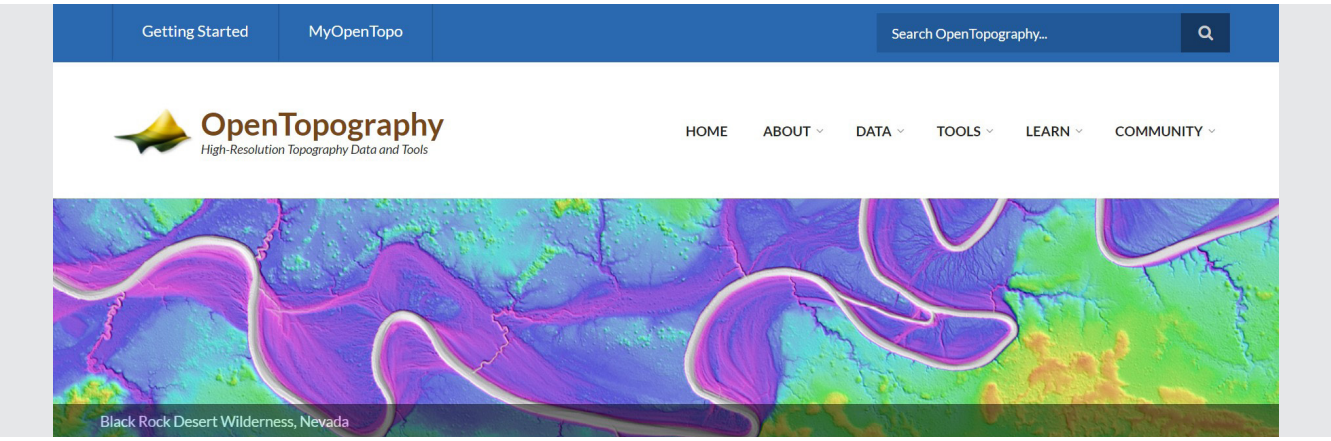

Figura 5. Captura de pantalla de la página de inicio del repositorio OpenTopography.

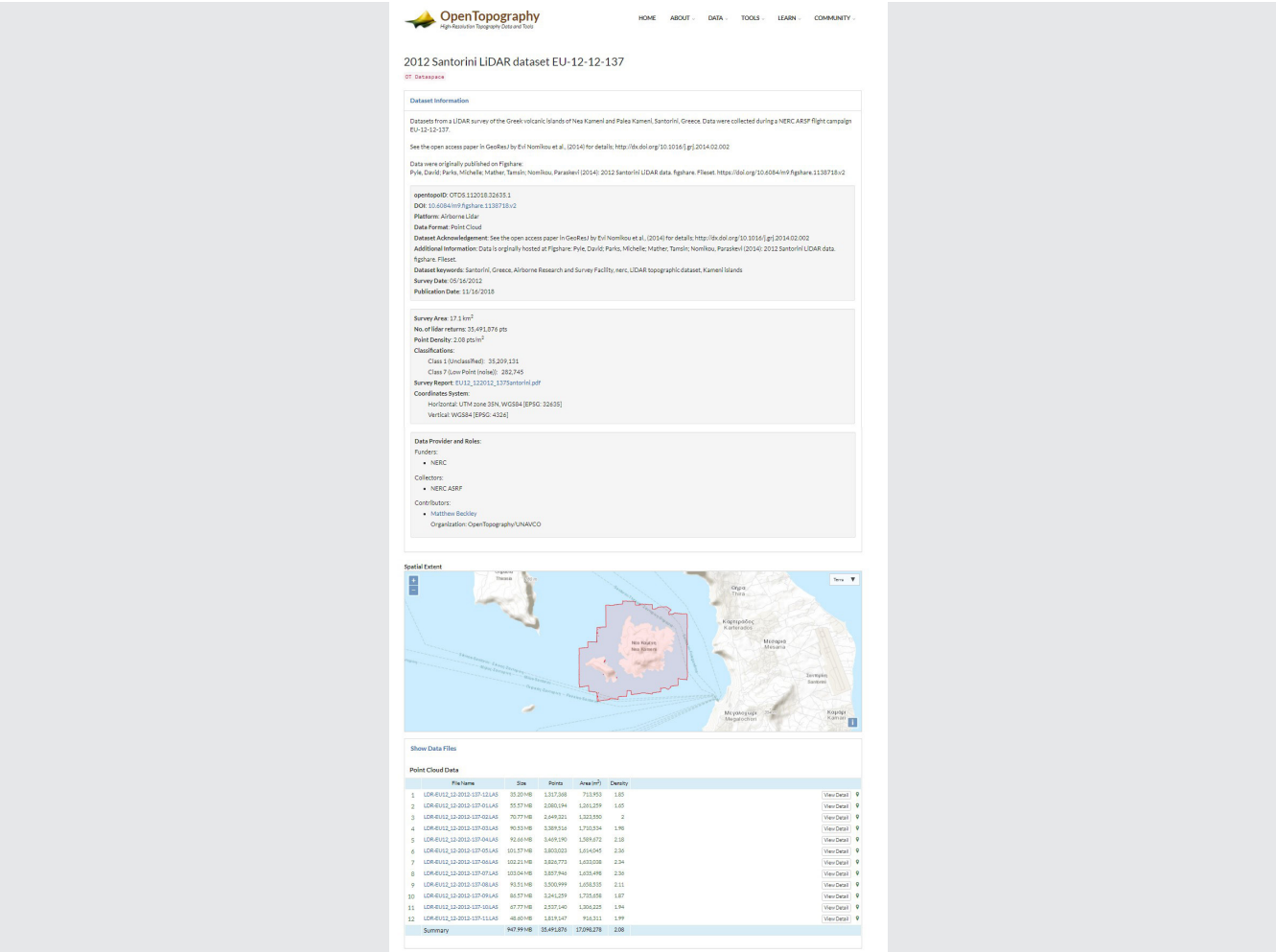

Figura 6. Captura de pantalla de un conjunto de datos obtenido en el repositorio OpenTopography.

A modo de conclusión sobre el apartado de los repositorios temáticos, hay algunos puntos que es interesante recalcar. El primero, aunque se ha visto que estos tres repositorios son tan diferentes en cuanto a contenido como las disciplinas a las que van dirigidos, su estructura es siempre parecida. Los tres están compuestos por una serie de guías y explicaciones sobre su funcionamiento, sobre cómo subir y descargar archivos y sobre cómo deben ser los datos que se admiten. Por otro lado, todos tienen una caja de búsqueda para que se puedan realizar las indagaciones pertinentes. Una vez que se encuentran los datos deseados, la organización de la información también es similar: primero metadatos que permiten contextualizar y comprender el contenido de los datos, y luego los propios datos para su descarga. Además, los tres tienen filtros que permiten perfilar la búsqueda, indicaciones sobre derechos de autor y recordatorios de cómo se han de citar los datos.

El segundo punto, recalcar que, de cara al investigador que quiere/debe depositar datos o desea buscar datos de otros, lo importante es conocer cuáles son los repositorios clave en tu campo de conocimiento, o cuáles son los que mejor se ajustan a lo que necesitas. Es una tarea cansada e innecesaria intentar conocer muchos repositorios y no llegar a entender y controlar bien ninguno.

Por último, y esto sirve de enlace para el siguiente apartado, no todas las disciplinas tienen repositorios de referencia, sea por falta de cultura de compartir datos o por cualquier otro motivo. Por eso, otra opción muy útil son los repositorios multidisciplinares, y de ellos vamos a hablar a continuación.

3.2.2 REPOSITORIOS MULTIDISCIPLINARES

Los repositorios multidisciplinares, como su nombre indica, son aquellos que están diseñados para dar cabida a todas o a la mayoría de las disciplinas, y para aceptar múltiples tipos de materiales. Son muy útiles, por ejemplo, si una disciplina no tiene repositorios de datos de referencia, o si se llevan a cabo investigaciones multidisciplinares y un repositorio temático pudiera quedarse escaso.

Al contrario de lo que sucede con los repositorios temáticos, en el caso de los multidisciplinares hay cuatro que destacan por ser los más conocidos. Es interesante conocer alguna pincelada de cada uno, ya que sus nombres suelen surgir cuando se habla de repositorios de datos. Además, aunque todos persiguen el fin de facilitar que los datos puedan ser compartidos, cada uno tiene particularidades que los hacen muy útiles a la hora de servir de ejemplo. Por ello, se dedicará un apartado a hablar de cada uno de ellos. Estos repositorios son: Zenodo, Dryad, Dataverse y Figshare.

· **Zenodo** (https://zenodo.org/)

Zenodo está pensado para ser un repositorio abierto para todos los tipos de investigaciones y de disciplinas, independientemente del tamaño o del formato, postulándose como "la solución" cuando no se cuenta con un repositorio temático ni institucional. Alojado en el CERN y financiado por la UE, es el repositorio que sirve de referencia para que lo utilicen los proyectos H2020 cuando quieran o deban depositar sus datos; por ello también es el repositorio recomendado por OpenAIRE[2] Aunque los datos son el núcleo de Zenodo y pueden colgarse sin necesidad de ir acompañados de ningún otro documento, también admiten otros archivos, como los propios artículos, actas de congresos o proyectos. Una de las posibilidades interesantes que ofrece Zenodo a los usuarios es la de subir varias versiones de su trabajo, quedando reflejada una especie de historial que permite a los propietarios de los datos seguir la evolución de su trabajo y a quienes los consultan, ver cómo evolucionan los proyectos. En la Figura 7 se puede ver la pantalla de inicio de Zenodo, donde se presentan los últimos archivos subidos y, como sucedía con los repositorios temáticos, también se observa una opción para realizar búsquedas. Entre otras opciones, también ofrece la posibilidad de ver las estadísticas de uso de tipo almétrico. Si observamos la Figura 8, tenemos un ejemplo de un conjunto de datos almacenados en Zenodo, que en este caso es un fichero .xlsx, junto con una explicación sobre los mismos. Se puede ver también que le otorga un DOI a los datos y que nos indica cómo tienen que citarse.

2. OpenAIRE es una infraestructura creada en el marco de la UE para impulsar y dar soporte a las investigaciones financiadas con fondos europeos en lo que se refiere al OA. Entre sus objetivos se encuentra el ofrecer formación, tejer vínculos y redes, facilitar la innovación y monitorear el OA (OpenAire, 2019).

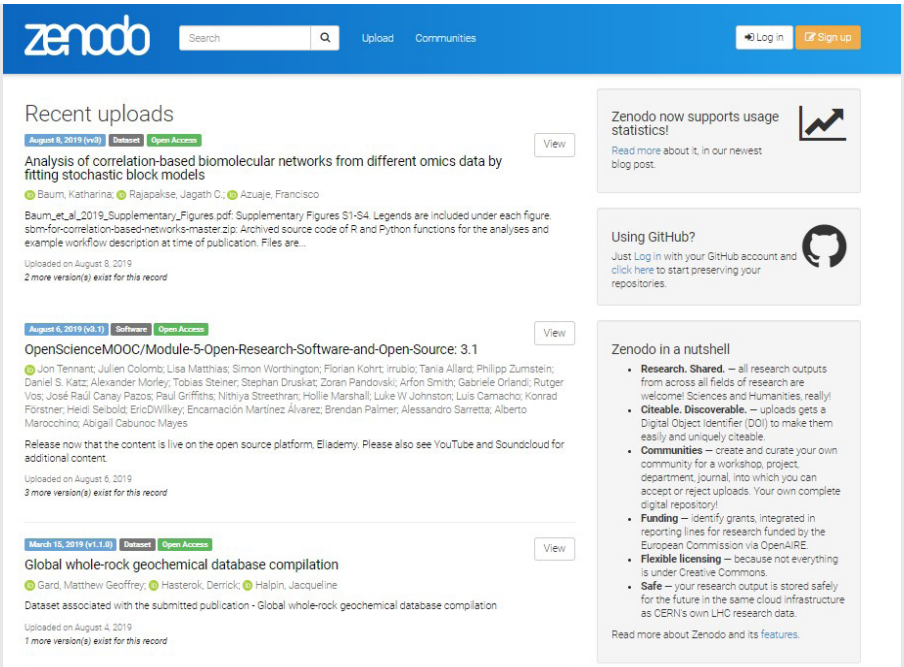

Figura 7. Captura de pantalla de la página principal del repositorio Zenodo.

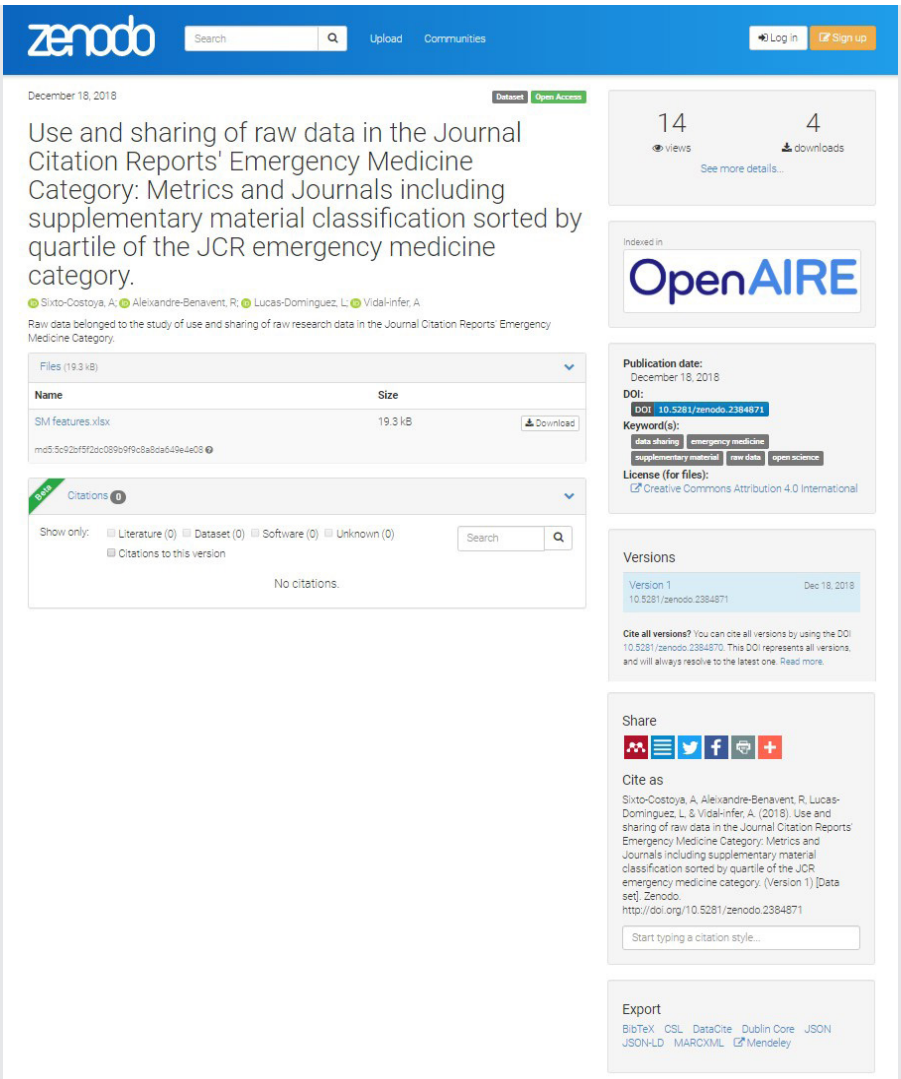

Figura 8. Captura de pantalla de un conjunto de datos depositados en el repositorio Zenodo.

· **Figshare** (https://figshare.com/browse)

Se trata de un repositorio en abierto multidisciplinar, que está apoyado por la Digital Science-MacMillan Publishers Company. Su público objetivo son los investigadores de cualquier disciplina. Igual que Zenodo, también otorga un DOI a los archivos subidos y estadísticas de uso de tipo almétrico. La gama de archivos que acepta es muy amplia, y va desde los propios artículos, a imágenes, vídeos, datasets o los pósters. Según Cho (2019), Figshare es el repositorio multidisciplinar que más datos contiene. En la Figura 9, se presenta una imagen de la pantalla de inicio de Figshare, que como se puede comprobar, tiene un aspecto mucho más minimalista y atractivo que Zenodo, casi recordando a lo que son las redes sociales. Una de las utilidades más interesantes de Figshare es que permite filtrar por materia, como se observa en la Figura 10, lo que facilita mucho la tarea a los usuarios que quieran buscar un contenido en concreto de su disciplina.

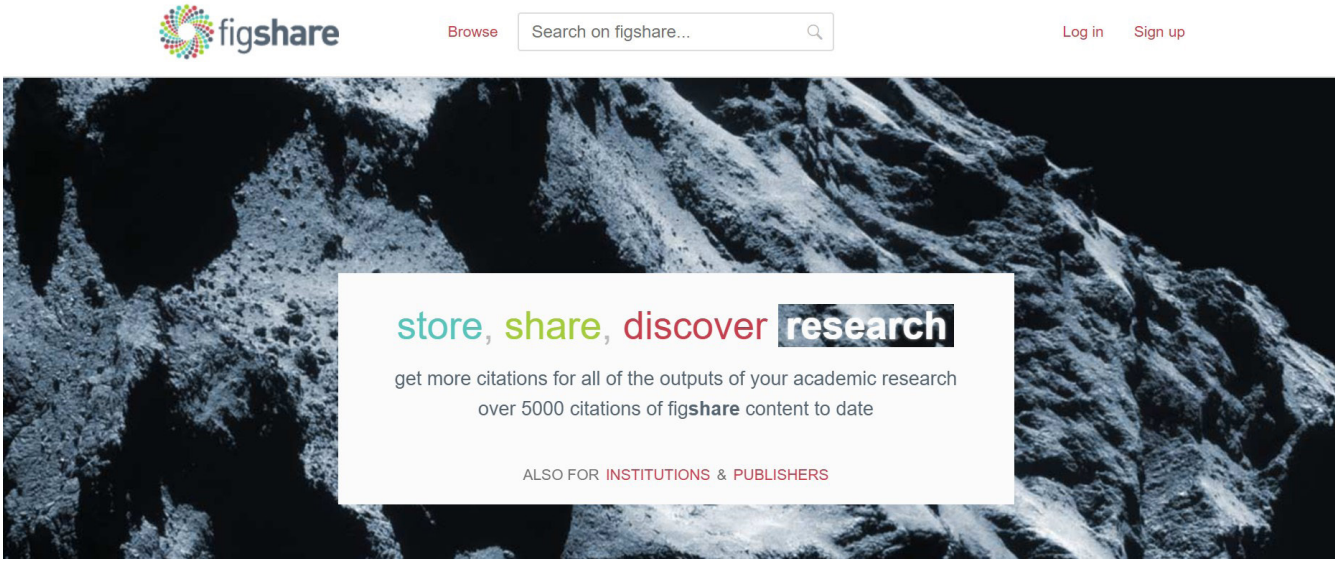

Figura 9. Captura de pantalla de la página de inicio del repositorio Figshare.

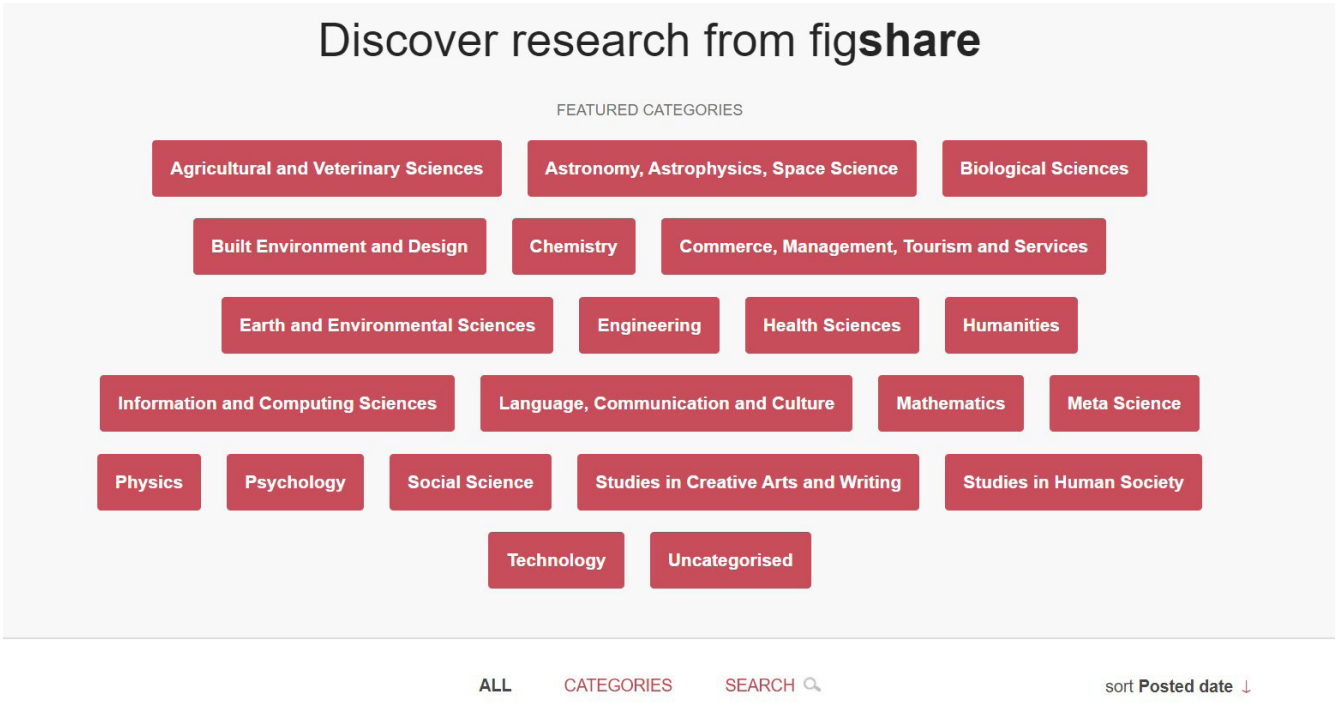

Figura 10. Captura de pantalla del filtro por disciplinas que ofrece Figshare.

· **Dataverse**

La primera particularidad de Dataverse (https://dataverse.org/) es que es uno de los repositorios más antiguos. Alojado en el Institute for Quantitative Social Science de la University of Harvard (EEUU), nace con la intención de centrarse en los datos de Ciencias Sociales. Actualmente, se ha abierto a todas las disciplinas y admite una cantidad de contenidos mucho más amplia. Un elemento diferenciador con los otros repositorios descritos es que Dataverse, además de ofrecer la opción de subir tus datos o descargar datos de otros, también da la opción a investigadores, revistas e instituciones de la instalación de un software mediante el que obtienen el servicio de Dataverse de una manera mucho más personalizada. Ofrece, por decirlo de alguna manera, un espacio a parte para que se pueda ejercer un mayor dominio de los datos, un repositorio propio para el investigador, o para su institución facilitado a través de Dataverse. En la Figura 11, en la página principal del Dataverse, se puede ver un mapa interactivo con todos los repositorios creados alrededor del mundo con esta funcionalidad. Si se pincha en el punto del mapa situado en España, se observa como el grupo español "Consorcio del Madroño" creó un repositorio a través de Dataverse (Figura 12). Para la organización de la información, Dataverse clasifica su contenido en archivos llamados "dataverses", que serían una especie de carpeta o categoría sobre un proyecto concreto o un grupo de investigación en particular, que a su vez contiene los conjuntos de datos con sus metadatos.

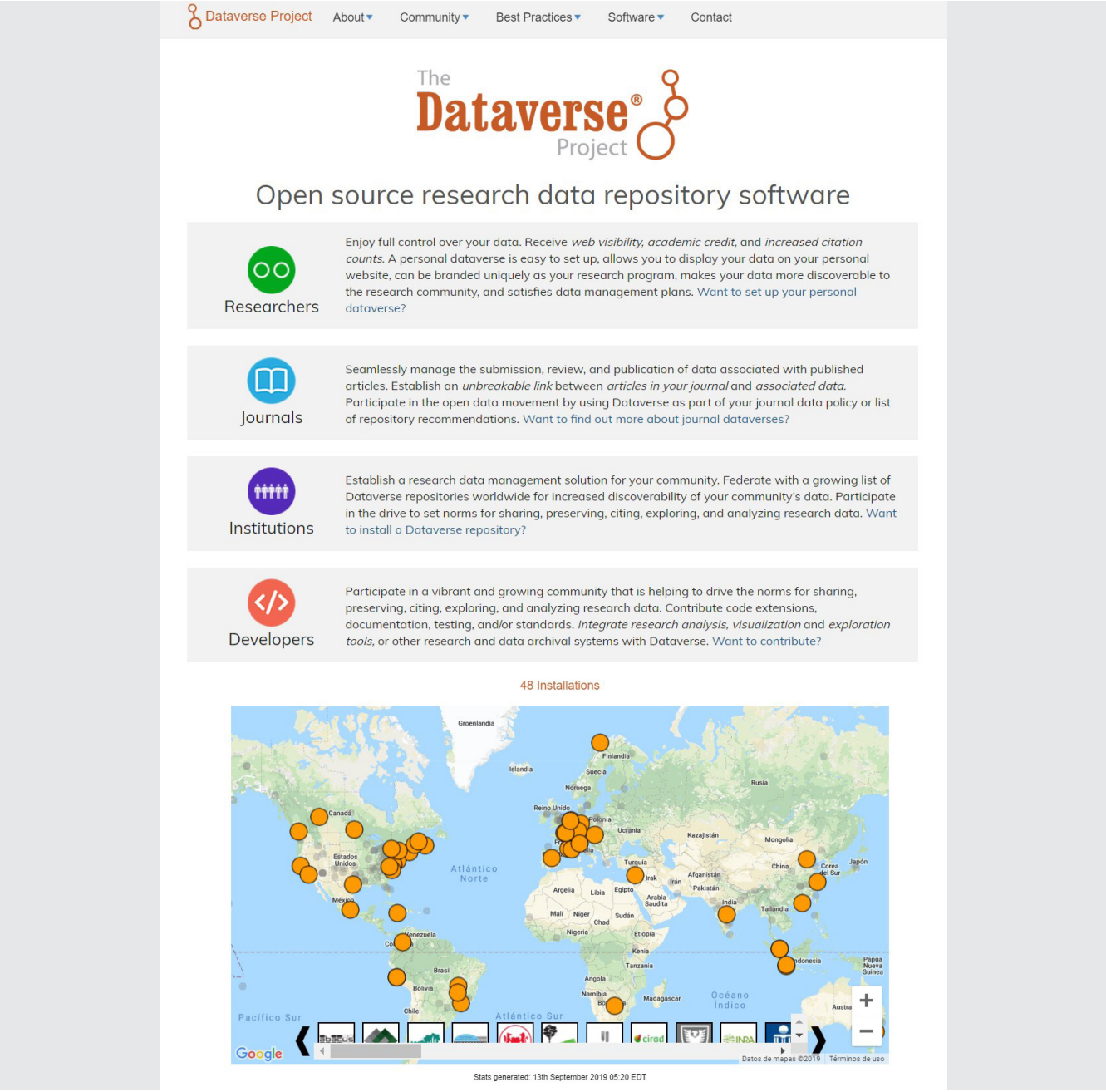

Figura 11. Captura de pantalla de la página de presentación de Dataverse.

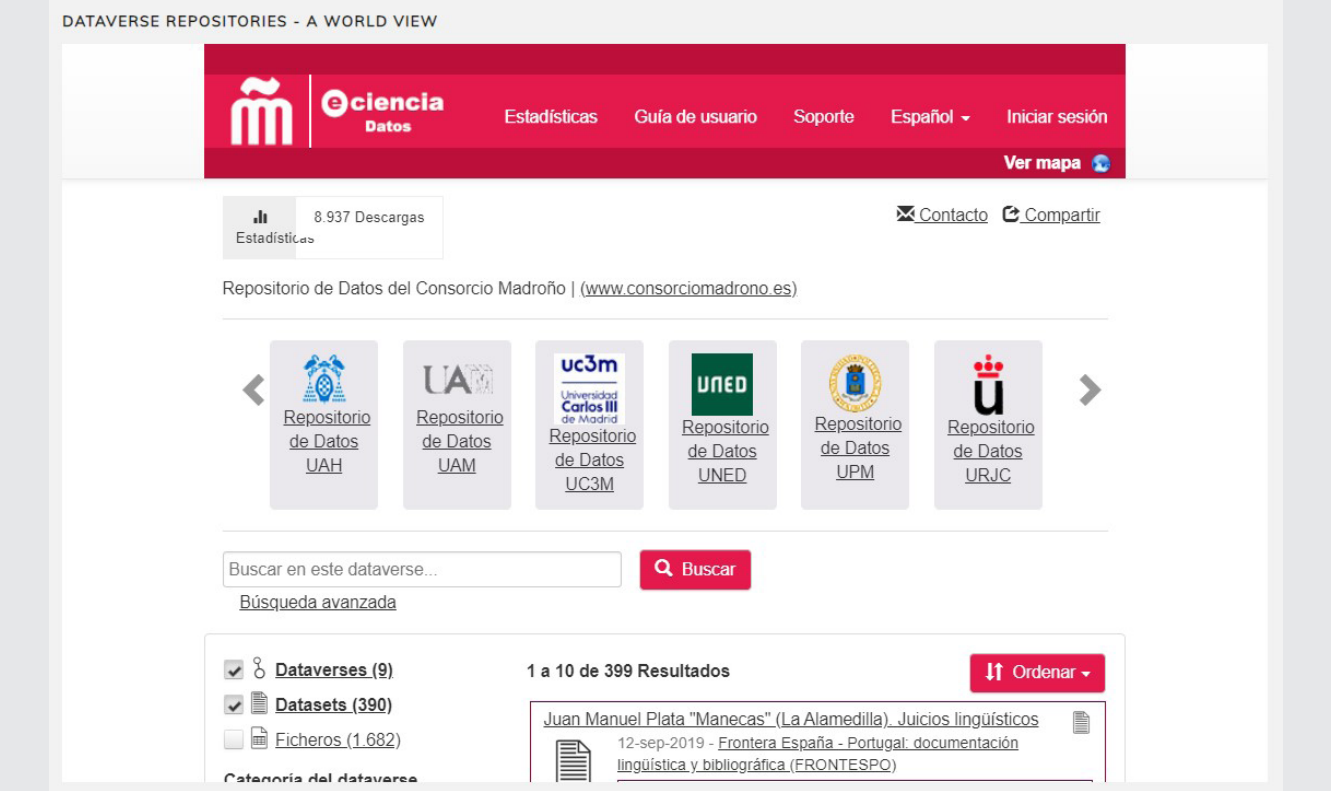

Figura 12. Captura de pantalla de repositorio creado por el Consorcio Madroño creado a través de Dataverse.

· **Dryad** (https://datadryad.org/pages/organization)

Aunque Dryad es también un repositorio considerado multidisciplinar, lo es de una manera un poco más particular, ya que las disciplinas que acoge son fundamentalmente de Ciencias de la Salud y Ciencias Naturales. Aun así, da cabida a tantas y tan diversas disciplinas dentro de estos campos que quedaría escaso considerarlo temático. Registrada como una Organización Sin Ánimo de Lucro, utiliza el software de código abierto DSpace para funcionar. Una de las particularidades de Dryad que cabe destacar es que todos los datos necesariamente deben estar ligados a una publicación, al contrario de lo que sucede, por ejemplo, en Zenodo, donde los datos se pueden subir "sueltos" solo acompañados por metadatos que los describan, pero sin enlazarse obligatoriamente a una publicación. Este punto hace que Dryad sea un recurso muy atractivo, ya que permite al usuario crearse un contexto al poder acceder a la publicación derivada de los datos brutos. En la Figura 13, se observa la página principal de Dryad, en la que, a la derecha podemos encontrar información acerca del recurso y en la de la izquierda, dos opciones para subir tus datos, o para buscar datos. Una ventaja del buscador de Dryad, que comparte con Dataverse, es que permite la búsqueda avanzada. En la Figura 14 se presenta un ejemplo de búsqueda en Dryad.

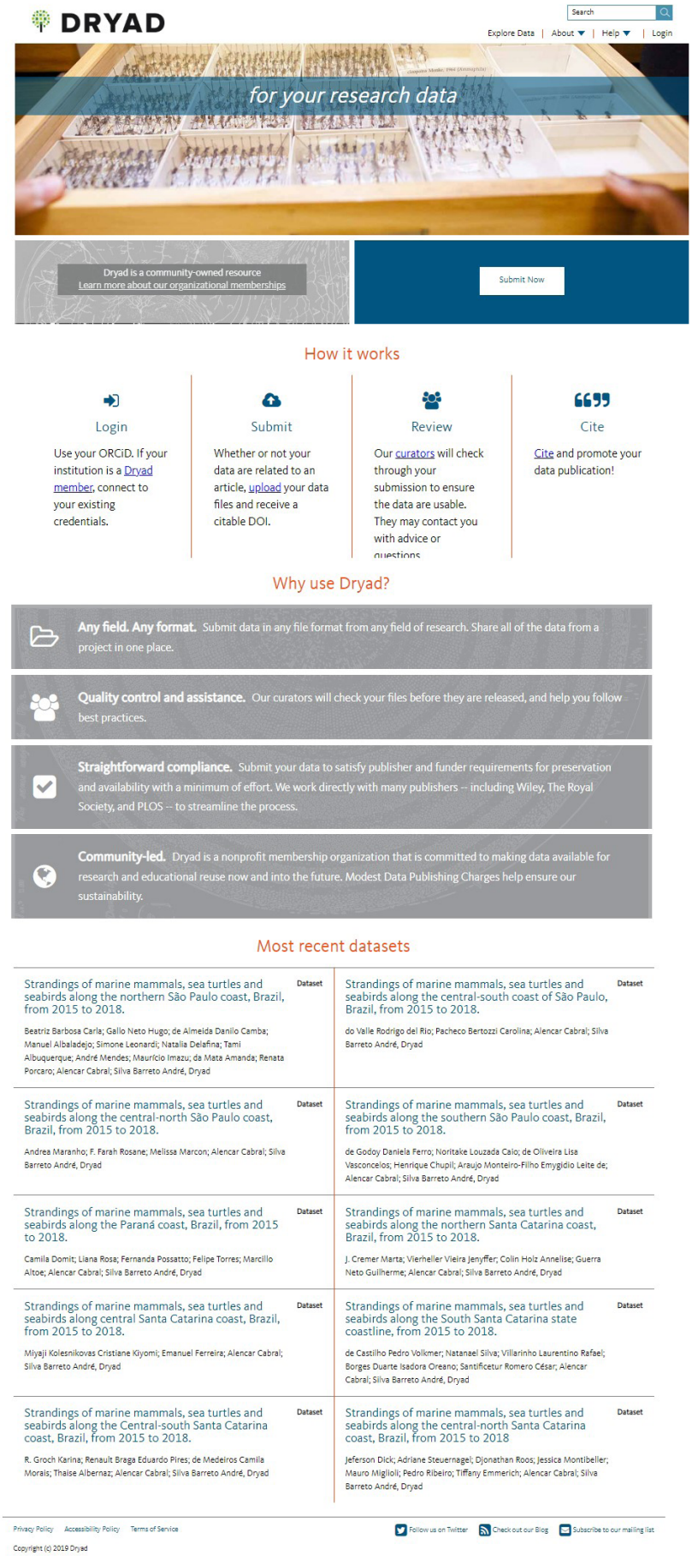

Figura 13. Captura de pantalla de la página principal del repositorio Dryad.

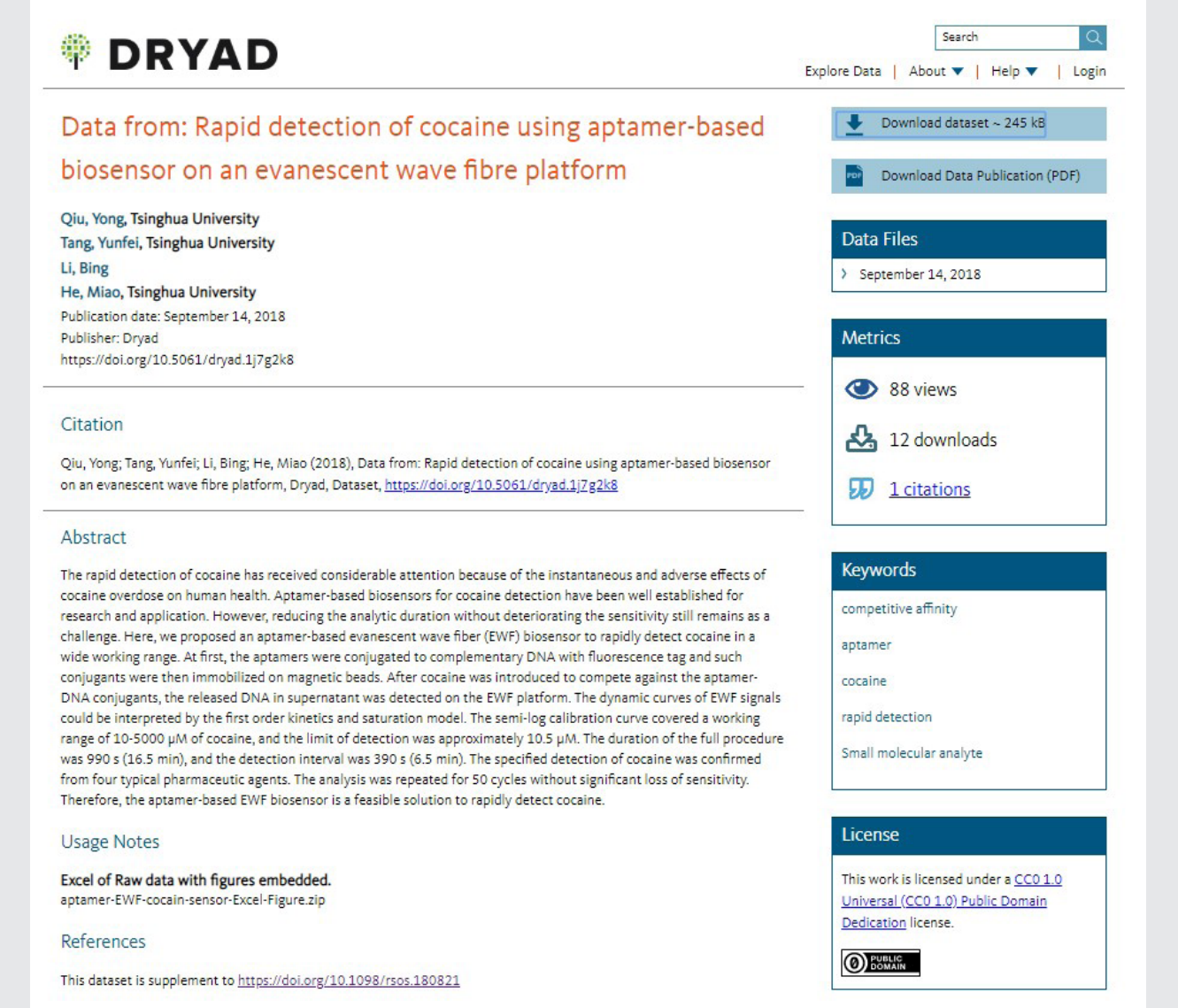

Figura 14. Captura de pantalla de una búsqueda en el repositorio Dryad.

3.2.3 REPOSITORIOS INSTITUCIONALES

En este último grupo se incluyen los repositorios de datos que fueron creados por instituciones (por ejemplo, las universidades o los centros de investigación) con el fin de que sirvan para almacenar los datos generados por sus investigadores. Como sucede con los repositorios temáticos, hay tanta variedad de instituciones en todo el mundo que crean sus propios repositorios, que es imposible mencionarlas a todas. En este trabajo pondremos un ejemplo de una a nivel europeo para que nos sirva de referencia: el CERN Open Data (**http://opendata.cern.ch/docs/about**).

El CERN Open Data, según se definen, es un repositorio que es el punto de acceso a toda la cantidad de datos producidos a través de la investigación llevada a cabo en el CERN (European Organization for Nuclear Research). Su objetivo es diseminar los ouputs de diferentes actividades científicas, incluyendo los softwares y la documentación necesarios para comprender y analizar los datos compartidos. Además, manifiesta estar en línea con los estándares globales en preservación de datos y Open Science, ya que los productos son compartidos bajo licencias abiertas. Igual que otros repositorios explicados anteriormente, también otorga un DOI para que el material pueda ser citable e identificado. En la Figura 15 se observa la página de inicio del repositorio, que cuenta con una explicación del funcionamiento del recurso y, en la parte superior, con una caja de búsqueda para realizar las averiguaciones de interés. En la Figura 16, se presenta un ejemplo de una búsqueda en el repositorio, que, como se puede comprobar, tiene una organización muy similar a los ejemplos que ya se han visto.

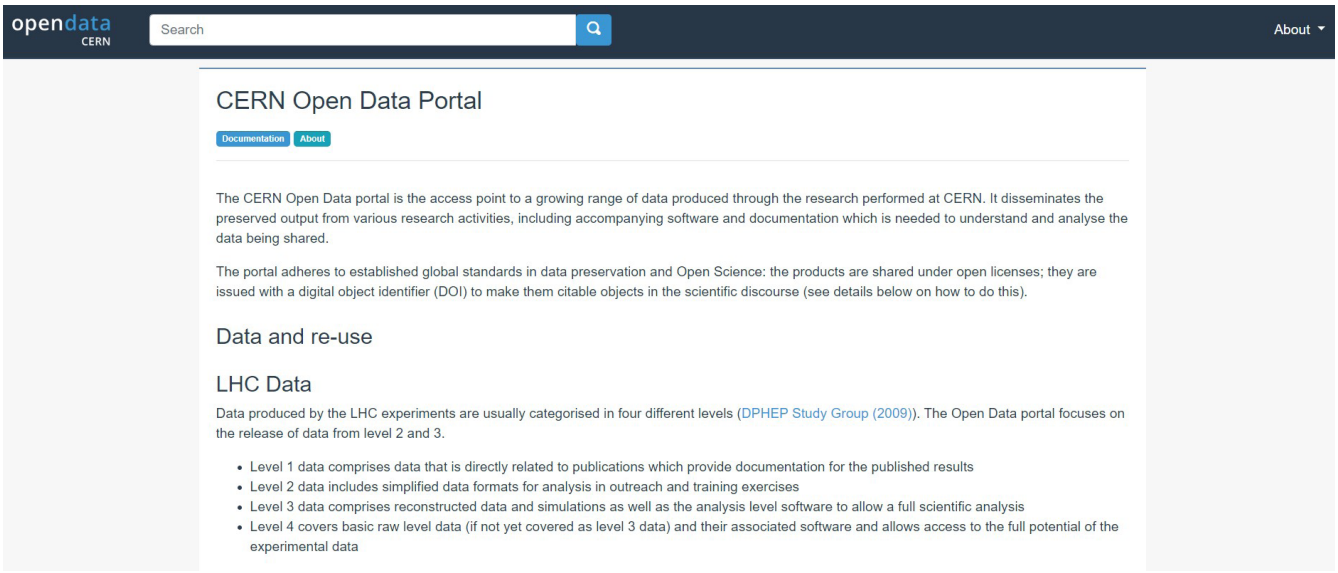

Figura 15. Captura de pantalla de la página de inicio del repositorio CERN Open Data.

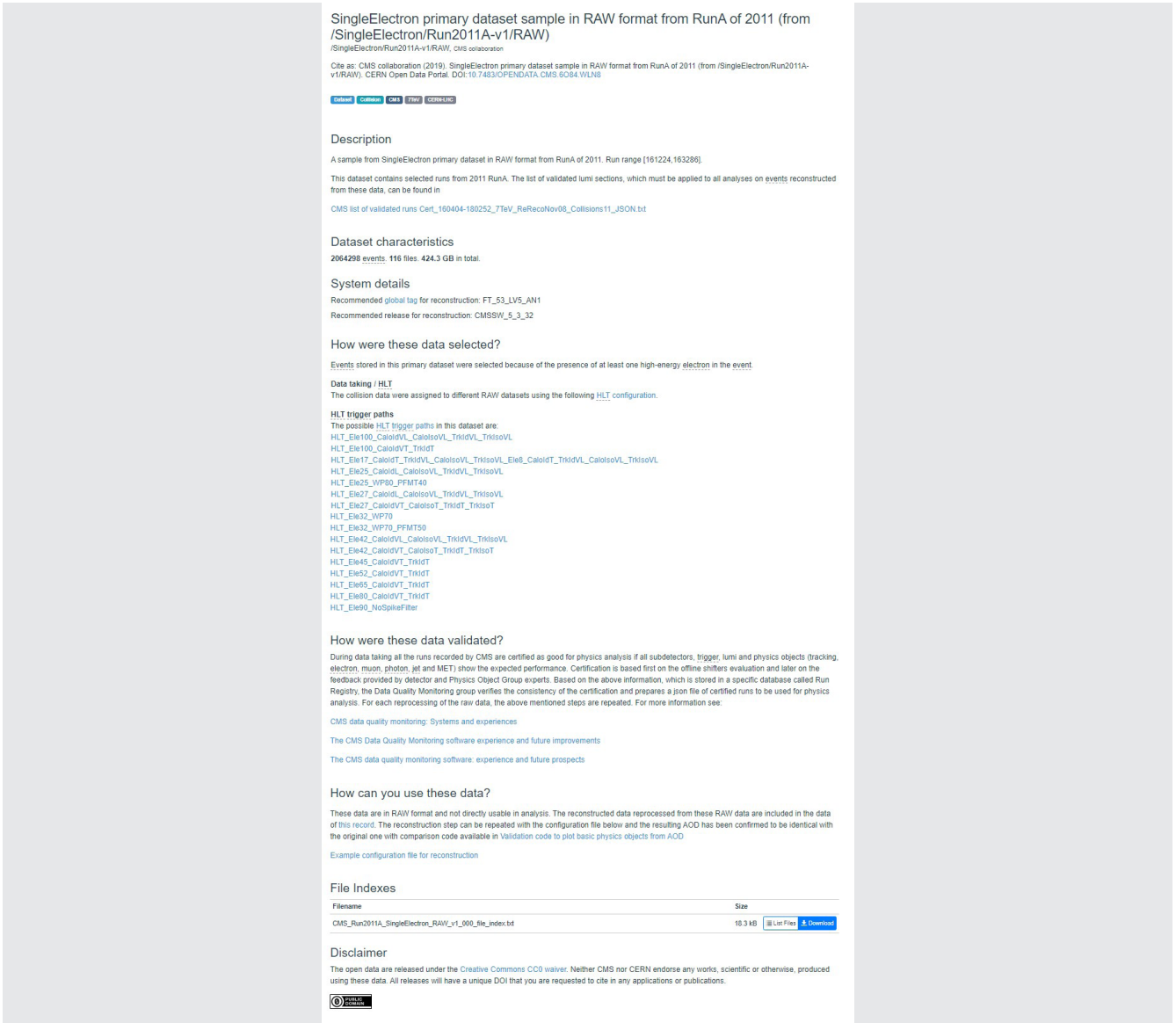

Figura 16. Captura de pantalla de la búsqueda de un conjunto de datos en el CERN Open Data.

3.2.4 BUSCADORES DE REPOSITORIOS DE DATOS

Para concluir el apartado de repositorios, es interesante mencionar el buscador de repositorios de datos "re3data" (**re3data.org**), que es de gran utilidad cuando se necesita encontrar un repositorio y no se sabe muy bien por dónde empezar.

Creado en el año 2013 por la German Research Foundation, se trata de una herramienta muy útil que tiene la intención de ofrecer a los usuarios una visión general de los repositorios de datos existentes. Por ejemplo, es muy útil para quien desconozca la gama de repositorios temáticos que hay para su disciplina y desee encontrar uno que se ajuste a sus necesidades. En opinión de Gómez, Méndez, y Hernández Pérez (2016), re3data se ha convertido en el registro por excelencia para encontrar repositorios de datos de todas las disciplinas. En las Figuras 17 y 18, se observa la pantalla de inicio de la web de re3data y un ejemplo del filtro por disciplinas que permite hacer. Además de ese filtro, también permite por "tipo de contenido" y "país".

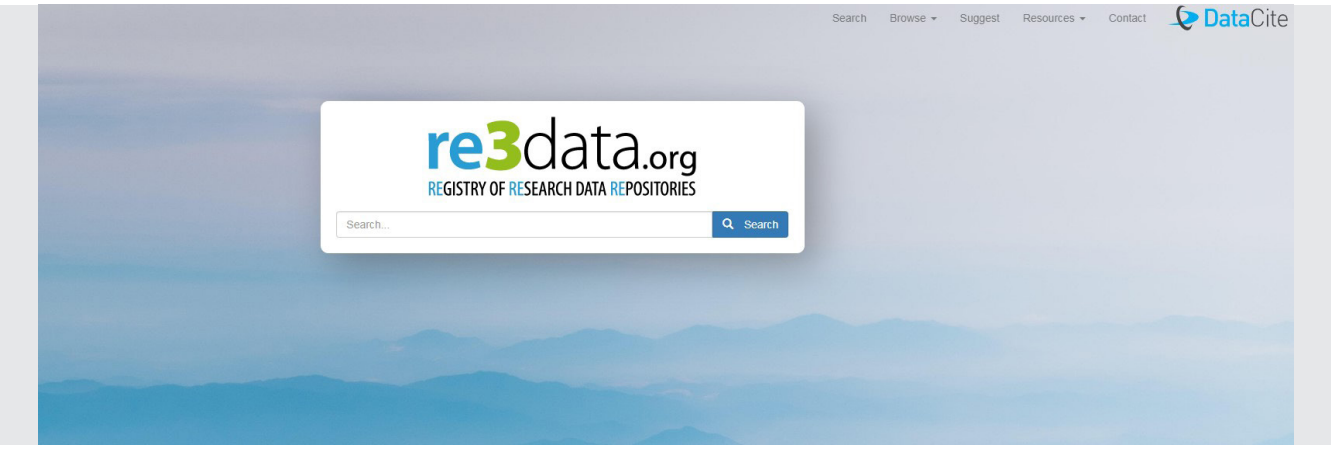

Figura 17. Captura de pantalla de la página web de re3data.

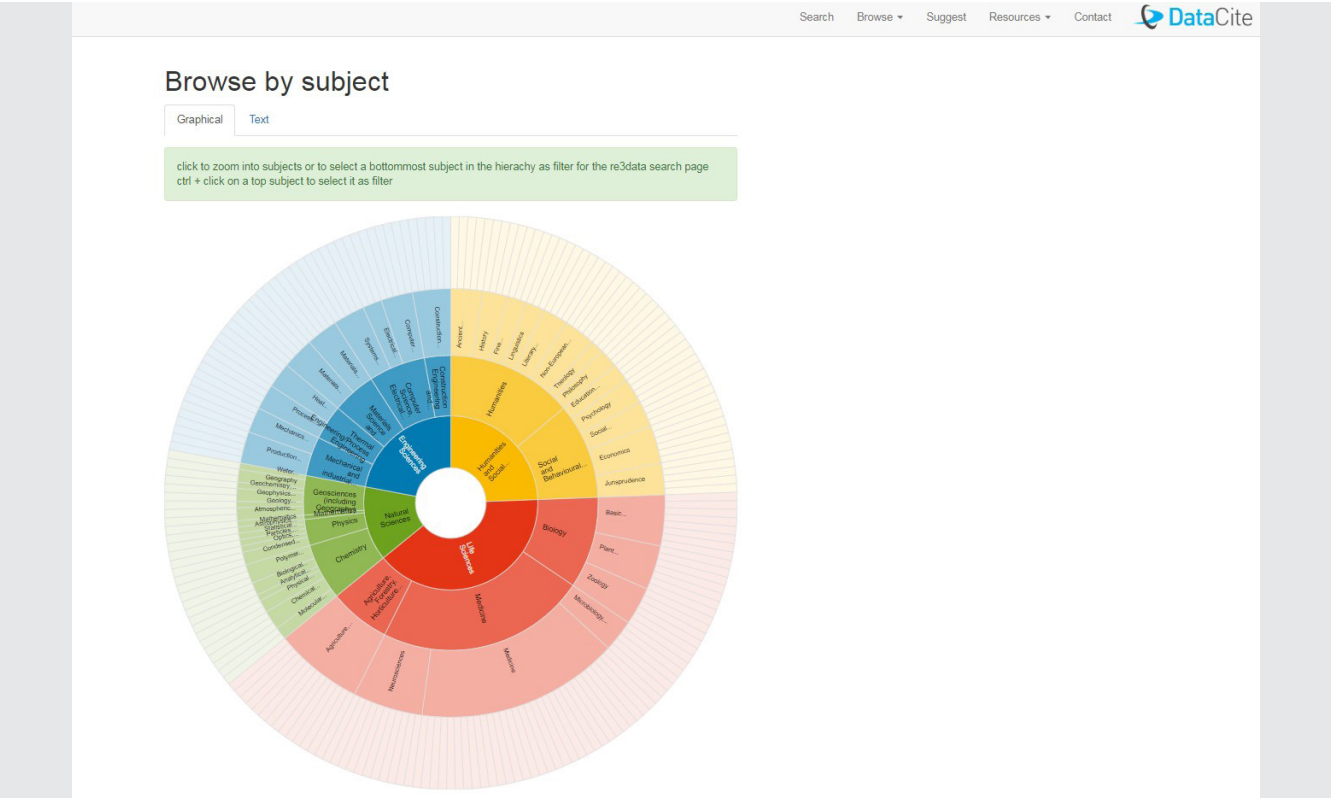

Figura 18. Captura de pantalla de la opción de filtro por disciplina que permite la herramienta re3data.

4. CARACTERÍSTICAS DE LOS DATOS DE INVESTIGACIÓN

Como ya se ha expresado al principio del documento, cuando en este contexto se habla de datos, se refiere a aquellos datos brutos que subyacen a las investigaciones y que hacen posible que se presenten resultados. Los resultados son el "cocinado" de esos datos, la información que obtenemos una vez realizados los análisis y que van en línea con los objetivos que se persigan. El objetivo del data sharing es que esos datos sean puestos a disposición de la comunidad científica para que pueda verse incrementado su valor.

Lo que se trata en este apartado es el cómo han de ser esos datos para que efectivamente puedan ser compartidos y considerados válidos, bien para reutilizar, como también para aumentar la transparencia cuando se utilizan para verificar si la información de los datos "cocinados" es cierta. Para ello, se habla del formato de los datos y de su estructura, así como de los metadatos que deben acompañarlos.

4.1 LOS DATOS

El primer apunte sobre los formatos de datos, es que se debe de asegurar que sean lo más fácilmente reutilizables posible. Buscando un ideal, los datos compartidos "perfectos" serían aquellos que están en un formato estructurado, libre, abierto y sin costes.

En este punto se hará un breve repaso de los conceptos "libre", "abierto" y "gratuito", ya que a veces se piensa que son sinónimos y se usan indistintamente. Cuando se habla de este tipo de formatos y decimos que es la forma ideal es porque se puede acceder a ellos (abiertos), se pueden obtener y modificar (libres) y se puede hacer todo ello sin costes (gratuitos). Cuando se habla del movimiento o filosofía de Open Data o Datos Abiertos, se refiere a datos que persiguen estar en este tipo de formatos.

En general, la idea principal es que han de ser formatos operables para que otros programas puedan abrirlos sin limitaciones y que el usuario pueda gestionarlos de manera gratuita. Un ejemplo de formatos de este tipo son .csv o .xlm, teniendo su opuesto en el formato .pdf (Generalitat de Catalunya, s.f). En la misma línea se manifiesta Michener (2015), cuando dice que será la mejor elección los formatos que no tengan propietario y que sean conocidos y aceptados entre la comunidad científica, y nuevamente ponen de ejemplo el formato .csv, situándolo en idoneidad por encima de Excel. La razón por la que este autor sitúa .csv por encima de .xlsx es que éste último utiliza formatos propietarios (protegido por una patente o derechos de autor, en este caso, Microsoft) y no libres (sin dueño). Sin embargo, aunque su formato es propietario, Excel tiene la gran ventaja de ser ampliamente conocido y manejado por la comunidad científica, como se puede observar en los trabajos de Aleixandre-Benavent et al., (2018) y Vidal-Infer et al., (2019), donde se puede observar que .xls es el formato interoperable más utilizado ya que, aunque no sea libre, sí que permite la posibilidad de exportar datos en otros formatos. Otra visión muy interesante es la que se ofrece a través de Portal Europeo de Datos (https://www.europeandataportal.eu/es/), recurso financiado por la UE para promover la accesibilidad de los datos abiertos y su valor, sobre el formato de los datos y cómo escoger el formato correcto. Desde su punto de vista, el formato más utilizable para otros investigadores que no sean los creadores originales, es el formato con el que el conjunto de datos se generó y gestionó por primera vez. La justificación es que ese formato original ofrecerá un contexto mucho más fácil de entender y que, además, se sabe que funciona porque ya ha sido utilizado. Aunque entienden que se puede dar el caso de que ese formato no sea el más libre y abierto, refieren que siempre existe la posibilidad de compartir los datos en dos formatos, el original y uno libre, lo que serviría también como práctica para la preservación de los datos a largo plazo en el caso de que el formato original fuera propietario y se quedara obsoleto (Portal Europeo de Datos, s.f).

Por otra parte, otras dos recomendaciones que hacen desde esta página con respecto a los datos están relacionadas con la estructura y con la forma de entrega. En cuanto a la estructura, recomiendan tener en cuenta que, dependiendo del formato de los datos, así debe ser la estructura, ya que no todos los datos se pueden representar de la misma forma. Aunque la estructura más frecuentemente utilizada es la tabular, los datos también se pueden representar de manera jerárquica (relaciones entre puntos de datos de manera vertical) o en red (relaciones entre cualquier combinación y cualquier dirección). Sobre la forma de entrega de los datos, las indicaciones que ofrecen son claras: si los datos abiertos se ofrecen en formato descargable, deben ser fácilmente descargables, es decir, pensar en detalles como comprimir el tamaño del dataset, por ejemplo en un .zip, para que no pese demasiado y no lleve mucho tiempo descargarlo, o utilizar una terminología que sea fácilmente comprensible. A continuación, se facilita una tabla (Tabla 2) donde se nombran algunos ejemplos de opciones de formatos libres utilizados actualmente dependiendo del tipo de archivo. Se trata de opciones útiles a las que se puede recurrir cuando se quiera asegurar de que los datos compartidos están totalmente accesibles para cualquier persona.

TIPO DE ARCHIVO	FORMATOS LIBRES
Imagen	· Joint Photographic Experts Group (.jpeg)
	· Portable Network Graphics (.png)
Texto	· OpenDocument Text (.odf)
Video	· XviD (.avi)
	· Ogg Theora (.ogv)
Hoja De Cálculo	· Código Seguro de Verificación (.csv)
	· OpenDocument Spreadsheet (.ods)
Audio	· Vorbis (.ogg)
	· Opus (.opus)

Tabla 2. Resumen informativo de algunos de los formatos libres existentes.

4.2 LOS PLANES DE GESTIÓN DE DATOS

Para finalizar este apartado sobre los datos, es relevante mencionar brevemente los llamados "Planes de gestión de datos" ("Data Management Plan", en inglés), por su importancia de cara a la organización de los datos en proyectos y otros trabajos, como las tesis doctorales. Siguiendo la definición que da el grupo Datasea (Datasea, s.f) para el plan de gestión de datos, éste es básicamente un breve documento sobre cómo se van a gestionar los datos en una investigación. Para elaborarlo, se debe hacer un ejercicio de reflexión antes de comenzar el trabajo que se vaya a hacer (una investigación sobre algún tema concreto, una tesis doctoral o la redacción de un proyecto) de qué datos vamos a generar y para qué, ya que el plan de gestión de gestión de datos se elaborará en función del propósito final de la investigación. En el caso de la Comisión Europea (European Commnission, s.f), se considera que los planes de gestión de datos son un elemento clave para una buena gestión de los datos, ya que describen cómo es ciclo de vida de los datos recolectados y su posterior procesado. En el ejemplo concreto de los proyectos H2020 que opten por la participación en el ORD pilot, se estipula que para los datos cumplan los principios FAIR, los planes de gestión de datos, que deben incluirse siempre, tienen que contener la siguiente información:

· Cómo será el manejo de los datos durante y después del proyecto.
· Qué datos de recopilarán, procesarán y/o generarán.
· Qué metodología y estándares se aplicarán.
· De qué modo los datos serán compartidos en acceso abierto.
· Cómo será el proceso de curación y preservación de los datos.

En el trabajo "Ten Simple Rules for Creating a Good Data Management Plan", Michener (2015) detalla diez directrices para elaborar un plan de gestión de datos de manera relativamente sencilla. Las directrices que conforman este decálogo son:

· **Regla 1**: Averiguar los requerimientos de la entidad/institución/ámbito del conocimiento en la que se enmarque la investigación.

· **Regla 2**: Identificar los datos que serán recolectados: 1) tipos: textos, tabla de datos; 2) fuentes (laboratorio, observación participante…); volumen (la cantidad de datos que se planean recoger); y formato de los datos.

· **Regla 3**: Definir cómo serán organizados los datos. Por ejemplo, explicar los programas que se van a usar para el manejo de los datos (Excel, MySQL…).

· **Regla 4**: Explicar cómo van a ser documentados los datos para que tengan contexto y sean entendibles. Básicamente, es la descripción de los metadatos (se explicarán con más atención en el siguiente apartado).

· **Regla 5**: Describir cómo será asegurada la calidad de los datos. Se trata de describir el proceso empleado para medir la calidad de los productos que se van a usar, por ejemplo, la consistencia y adecuación del software.

· **Regla 6**: Presentar una estrategia de preservación y accesibilidad, por ejemplo, explicando qué repositorios de datos se planea utilizar.

· **Regla 7**: Definir las políticas de datos que se llevarán a cabo. Se refiere a los asuntos relacionados con las licencias de uso de los datos, los períodos de embargo, o las restricciones éticas sobre los datos sensibles.

· **Regla 8**: Describir cómo lo datos serán diseminados. Por ejemplo, de qué modo los datos van a estar disponibles y quiénes van a tener acceso a los mismos.

· **Regla 9**: Asegurar los roles y responsabilidades. Consiste en la asignación de roles a cada miembro del grupo que forme parte de la investigación que se va a llevar a cabo.

· **Regla 10**:Preparar un presupuesto realista. Este último punto se refiere a establecer los costes monetarios que supondrá la gestión de los datos en términos materiales y humanos.

En el siguiente enlace: **https://library.bath.ac.uk/research-data/data-management-plans/university-dmp-templates**, se presenta un recurso ofrecido por la biblioteca de University of Bath, donde se les facilita tanto a los usuarios de su universidad como a cualquiera que accede a su web, información sobre los datos de investigación y su tratamiento. Una de las funcionalidades que aportan es información sobre qué son los planes de gestión y cómo elaborarlos. Pero lo realmente interesante es que también ofrecen una serie de plantillas para elaborar planes de gestión. Estas plantillas están dirigidas a distintos niveles de la carrera investigadora, de este modo, hay plantillas para que utilicen doctorandos para su tesis, y plantillas generales para investigadores. En la Figura 19 se observa una captura de pantalla del servicio que ofrece la biblioteca de plantillas de planes de gestión de datos.

Por otra parte, la UE, a través de los proyectos H2020, también ofrece plantillas para planes de gestión de datos, que se pueden encontrar a través de este enlace: **https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm**. En esta plantilla se requieren una serie de campos como la explicación de cuál es el propósito de generación de los datos y su relación con los objetivos del proyecto, qué formatos tendrán esos datos, si se generarán desde cero o si se hará algún tipo de reutilización o en qué medida se adaptarán los datos a los principios FAIR. Además, se matiza que no es necesario proporcionar respuestas detalladas a todas las preguntas en la primera versión del plan de gestión, sino que se trata de un documento vivo que va adaptándose a medida que avanza el proyecto.

Figura 19. Captura de pantalla de la página de la biblioteca de la University de Bath que ofrece plantillas de planes de gestión de datos.

4.3 LAS CITAS A LOS DATOS

En relación a las citas a los conjuntos de datos de investigación, la Crue Universidades Españolas y la Red de Bibliotecas (REBIUN) (2016) elaboraron un documento informativo muy útil para conceptualizar las citas a los datos e informar sobre cómo debe elaborarse una cita. En primer lugar, subrayan que los conjuntos de datos son resultados de investigación del mismo modo que lo son los artículos y las monografías, por lo que citarlos correctamente es un mecanismo que permite su identificación, su acceso, localización y reutilización. Además, las citas a los datos permiten, del mismo modo que sucede en los artículos, el reconocimiento de la autoría de sus creadores, lo que permite valorar el impacto (tanto de los datos en sí como de los investigadores que los crean) y el desarrollo de métricas sobre los mismos.

En este sentido, lanzan una serie de buenas prácticas en la citación de los datos, que se pueden encontrar con más detalle en la guía elaborada por el DCC "How to Cite Datasets and Link to Publications" (Ball y Duke, 2015): 1) facilitar la identificación, localización y acceso mediante un identificador único y persistente (normalmente, un DOI), 2) cada conjunto de datos se tiene que citar de manera independiente, 3) las citas a los datos deben estar presentes en la sección de referencias de la publicación. En relación con el último punto, la Biblioteca de la Comisión Económica para América Latina y el Caribe (2019), matiza que, además de las citas a los datos, existe otro mecanismo para mencionar a los datos dentro de una publicación científica que sería la "declaración sobre el acceso a los datos". Esta declaración debe incluir, como mínimo, información relativa a qué datos están disponibles y en qué repositorio, o, en su caso, cómo contactar para solicitar el acceso a los mismos (debe ser un correo electrónico de la institución o departamento correspondiente y nunca una dirección personal); en caso de estar en libre acceso, facilitar el identificador persistente; y, cuál es la información legal sobre las condiciones que rigen el acceso a los datos.

Sobre el cómo elaborar la cita a los datos, la Crue Universidades Españolas y la REBIUM detallan que existen unos elementos mínimos y otros recomendados que, como sucede con las citas a los artículos, se combinan para crear citas en los diferentes estilos.

Los elementos obligatorios son: autores, fecha, título, identificador persistente, tipo de recurso y versión o edición.

Los elementos optativos son: identificador autor (por ejemplo, ORCID), repositorio de datos, publicación, productor, ámbito geográfico y ámbito temporal.

A continuación, se ofrece un ejemplo de cita a los datos que confecciona el repositorio Zenodo de manera automática cuando se deposita un conjunto de datos:

Sixto-Costoya, A, Aleixandre-Benavent, R, Lucas-Dominguez, L, & Vidal-infer, A. (2018). Use and sharing of raw data in the Journal Citation Reports' Emergency Medicine Category: Metrics and Journals including supplementary material classification sorted by quartile of the JCR emergency medicine category. (Version 1) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.2384871

LOS METADATOS

Aunque ya se han mencionado en varias ocasiones en este documento, en este apartado se hace una pequeña parada para tratar con más detenimiento otro de los pilares imprescindibles en el uso compartido de datos: los metadatos. Básicamente, se trata de información que acompaña a los datos que queremos compartir para darle un contexto (Torres-Salinas et al., 2012). Tienen que ser lo suficientemente completos como para que la persona que recibe los datos tenga toda aquella información necesaria para comprender qué contienen, cómo están organizados y cuál es la procedencia del material compartido. Sin esta explicación, compartir datos puede pasar de ser una práctica muy útil, a una pérdida de tiempo tanto para la persona que comparte como para la que recibe.

En el libro "Understanding metadata. What is Metadata, and what is for?" Riley (2017) hace una clasificación muy útil de los distintos tipos de metadatos y sus propiedades. Aunque lo enfoca más en datos relacionados con museos y patrimonio cultural, su clasificación es también válida cuando hablamos de datos de investigación. En primer lugar, destaca los denominados metadatos "descriptivos", que se refieren a la información que sirve para entender el contenido del conjunto de datos. Por otra parte, estarían otros tipos de metadatos más centrados en la forma y en la estructura, como los metadatos "técnicos" o los metadatos sobre los derechos de autor. Para una comprensión más visual, a continuación, presentamos una tabla con los tipos de metadatos, ejemplos sobre su contenido y sus usos principales, inspirada en la descripción de Riley, pero adaptada a los datos de investigación. En la Tabla 3 se presenta un resumen informativo de los tipos de metadatos, con ejemplos de su contenido y sus usos principales.

TIPO DE METADATOS	EJEMPLOS DE SU CONTENIDO	USOS PRINCIPALES
Metadatos descriptivos	· Título	Dar un contexto al contenido en sí de los datos.
	· Autores	
	· Fecha de publicación del estudio	
	· Resumen de la temática del estudio	
	· Protocolos para obtención de los datos	
Metadatos técnicos	· Tipo de archivo	Detallar los tipos de archivos en los que los datos se encuentran disponibles, cuánto pesan e informar sobre la antigüedad de su creación.
	· Tipo de formato	
	· Tamaño del archivo	
	· Fecha de creación de los datos (o de las versiones)	
Metadatos sobre la preservación	· El DOI otorgado en caso de haberlo	Informar sobre qué modos de preservación existen para el conjunto de datos.
	· Indicación del repositorio de datos donde está depositado	
Metadatos sobre los derechos	· Derechos de copyright	Indicar qué derechos se han estipulado para el conjunto de datos
	· Licencias Creative Commons	

Tabla 3. Resumen informativo de los tipos de metadatos.

A continuación, se pone un ejemplo de qué metadatos se necesitarían para un dataset procedente de un estudio realizado en un laboratorio donde se hacen ensayos preclínicos con animales. En este caso, los metadatos descriptivos deberían incluir, como mínimo, qué protocolos se llevaron a cabo para la obtención de esos datos, es decir, la lista de instrucciones que se siguieron, así como el tipo de animales que se utilizaron, cuáles eran las condiciones del laboratorio, en definitiva, todo lo referente a cuál fue el método seguido y en qué consistía el estudio realizado. En cuanto a los metadatos técnicos, estaríamos hablando de qué tipo de archivo y en qué formato se encuentran esos datos; en este caso, por ejemplo, podrían ser archivos en formato .xlsx y/o .csv creados en "X" fecha. Además, deberían indicar qué derechos se otorgan para la reutilización de esos datos y, por supuesto, el DOI o el enlace que corresponda para poder consultarlos, compartirlos o citarlos.

Finalmente, mencionar que, en la actualidad, existen numerosos estándares de metadatos elaborados por organizaciones e instituciones para adaptarse a la necesidad de describir los datos de manera sistematizada y efectiva. Algunos ejemplos de iniciativas relacionadas con estándares de metadatos, explicados muy bien por los autores Kim y Burns (2016) en su trabajo centrado en el campo de la biología, son la Data Documentation Initiative (DDI), el Darwin Core, el Ecological Metadata Language (EML) o el Content Standard for Digital Geospatial Metadata (CSGDM).

5. EL PAPEL DE LAS REVISTAS CIENTÍFICAS EN EL USO COMPARTIDO DE DATOS

Cuando se habla de compartir datos en el contexto científico, es imprescindible tener en cuenta la posición de las revistas y de las editoriales, ya que suponen el vehículo más importante de comunicación científica que existe actualmente (González-Alcaide y Ferri, 2014). Debido a esto, las posiciones que adopten en relación al uso compartido de datos son concluyentes ya que van a influir, de un modo u otro, en el éxito o fracaso de estas iniciativas. En lo que respecta al movimiento por el uso compartido de datos, al igual que sucede con el acceso abierto a las publicaciones, lo cierto es que, en los últimos años, algunas revistas y editoriales han tomado nota del mismo y han actuado en consecuencia (Alsheikh-Ali et al., 2016). Esta toma de conciencia se cristaliza en políticas y directrices sobre uso compartido, que el autor se encuentra normalmente en las "Author guidelines" a la hora de querer publicar. Dentro de las revistas y de las editoriales que apuestan por el uso compartido de datos, se distinguen dos grupos: las que sugieren y las que obligan. Dependiendo de qué revista o editorial sea, se diferencian tres formas fundamentales en que se indica a los autores (bien por obligación, o por recomendación) cómo compartir sus datos:

1. El primer nivel, que no implican ni obligación ni exigencia, es la que le permite al autor poner una indicación de que los datos pueden ser solicitados al autor o autores, normalmente indicando alguna forma de contacto, y éstos los facilitarán si lo consideran oportuno. Se trata de manera más liviana, pero ha de tenerse en cuenta, ya que al menos abre la puerta a que los datos sean solicitados. En la Figura 20, se observa un ejemplo de una publicación en la revista *Frontiers in Behavioral Neuroscience*, donde los autores indican que los datos pueden ser pedidos bajo la frase "*The datasets generated for this study are available on request to the corresponding autor*".

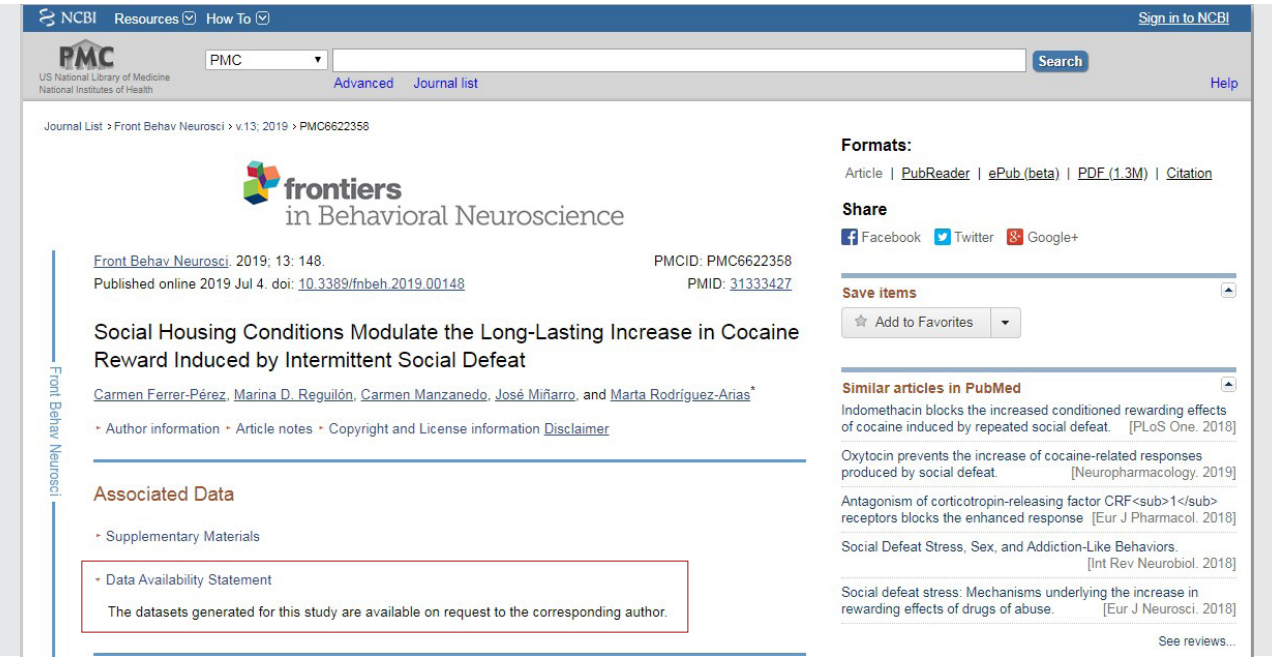

Figura 20. Captura de pantalla de un artículo de la revista *Frontiers in Behavioral Neuroscience*.

2. El segundo nivel, que implica más apertura de los datos, es el que ofrece la posibilidad de que vayan adjuntos como material suplementario al artículo. En este caso, la revista puede optar porque este sea un requisito o no. En caso de que lo sea, los autores deben subir un adjunto donde se encuentren los datos brutos que sustentan los resultados del trabajo, lo que es muy positivo ya que se permite tanto la verificación como la reutilización. Esta modalidad hace que las revistas y las editoriales puedan ejercer de plataformas digitales en las que se almacenan los datos, ya que, junto con el pdf del artículo, estarían adjuntos los datos. No obstante, tiene la desventaja de que, si el artículo no está en Open Access y solo algunas personas pueden acceder a él, solo esas personas podrán acceder a los datos, por lo que no estarían abiertos al cien por cien. En este sentido, destacan funcionalidades como la de PubMed Central (PMC), el repositorio de acceso abierto al texto completo de las publicaciones de PubMed, ya que tiene la opción de filtrar por "Associated Data", lo que permite saber qué artículos tienen asociado material suplementario. En la Figura 21, se ofrece un ejemplo de un artículo depositado en PMC que pone a disposición un adjunto con los datos brutos, en este caso, una serie de documentos en formato .doc y, sobre todo, .xlsx.

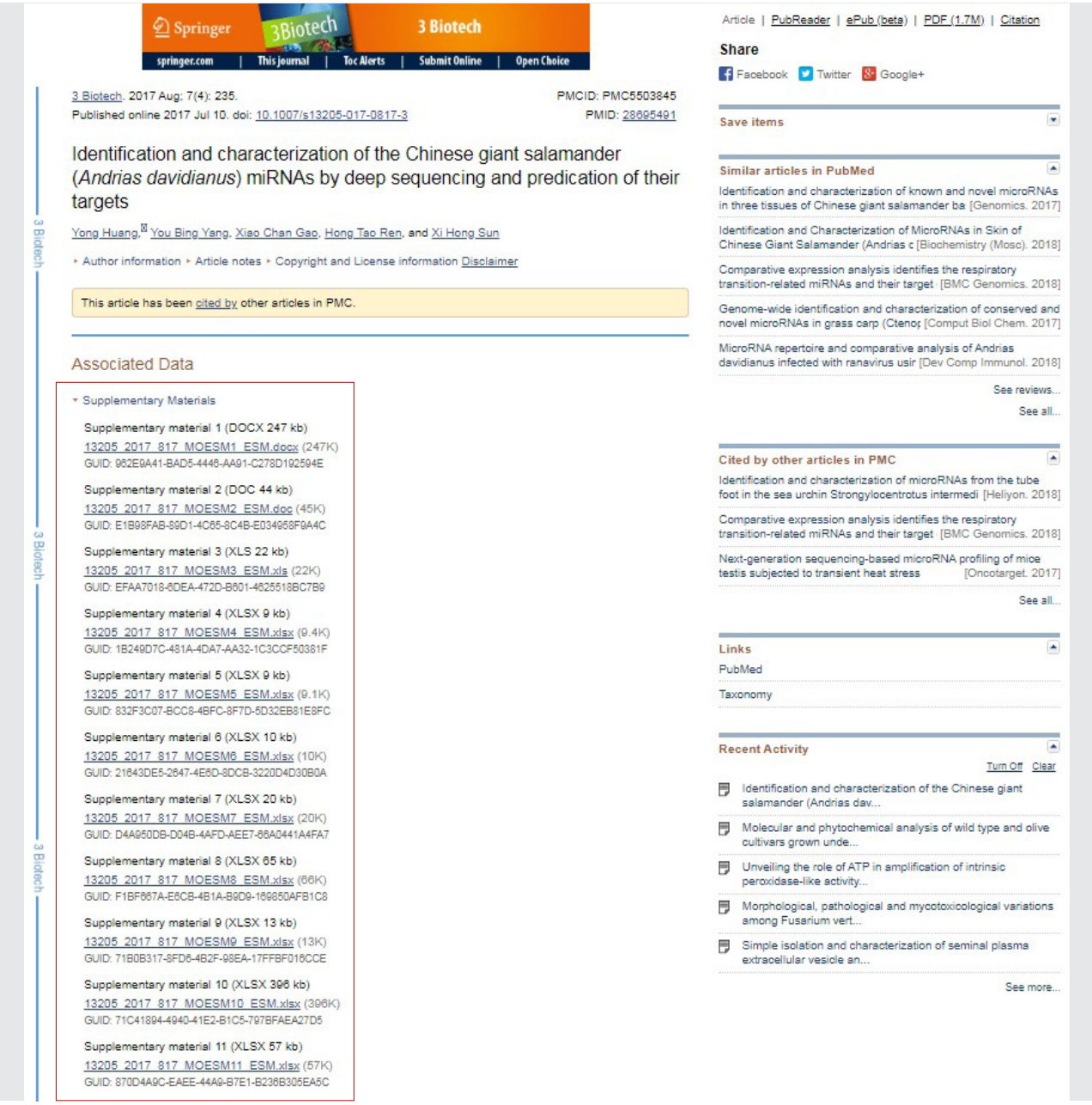

Figura 21. Captura de pantalla de un artículo de la revista 3 *Biotech*.

3. El tercer nivel, aquel en el que los datos están más abiertos, es que la revista obligue o dé la posibilidad de subir los datos a un repositorio de datos como los que se han visto en el apartado referente a repositorios. De esta forma, los datos estarían guardados con una modalidad que ofrece muchas más garantías tanto de preservación como de accesibilidad. Dependiendo de las revistas y de las editoriales a las que pertenecen, pueden variar sobre todo dos cuestiones. La primera es si se trata de una obligación o de una recomendación. La segunda, el tipo de instrucciones que se detallen, por ejemplo, si se da libertad al autor para que se escoja el repositorio o si la revista ya tiene una lista cerrada de repositorios para elegir. En la Figura 22, se presenta el ejemplo de un artículo de la revista *American Journal of Political Science*. En este caso, se trata de una revista que obliga a los autores a que faciliten los datos que sostienen sus resultados, principalmente con el fin de permitir a replicabilidad. Como se puede observar en la Figura 22, después del abstract de un artículo se facilita el DOI de los datos pertenecientes a este estudio. Si se pincha en el DOI, se redirige al repositorio de datos de referencia para esta revista: Dataverse, que se ha explicado anteriormente, y una vez aquí se observa que están todos los elementos necesarios para el uso compartido: los datos y los metadatos (Figura 23).

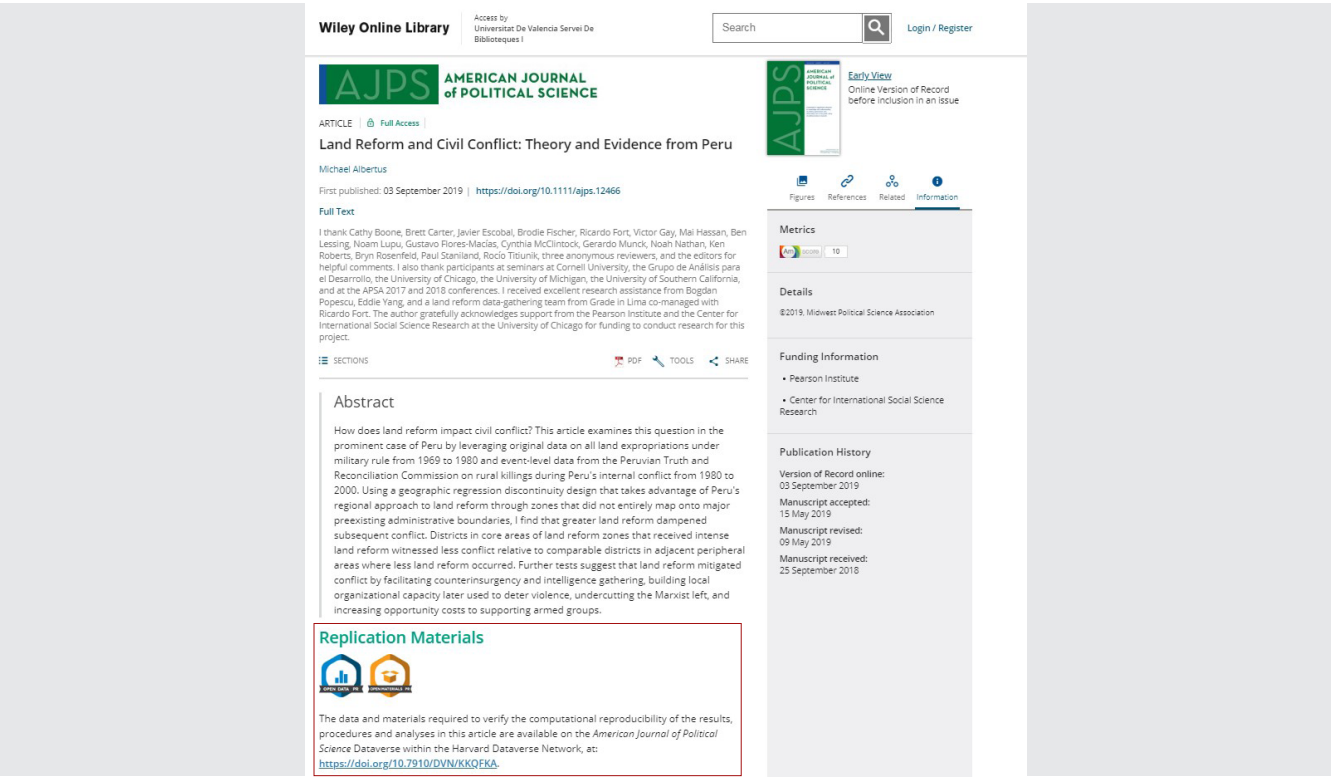

Figura 22. Captura de la pantalla de un artículo de la revista *American Journal of Political Science*.

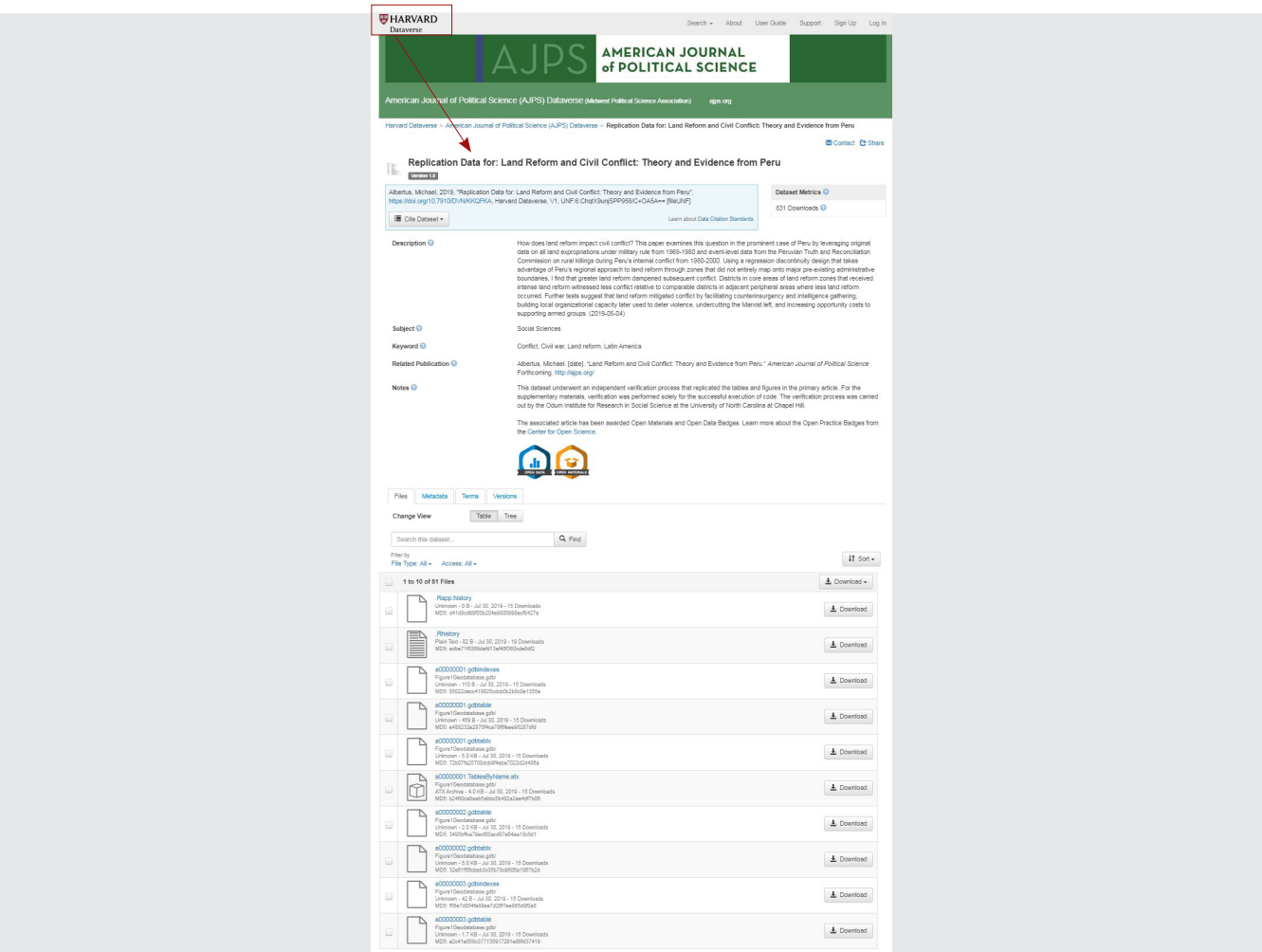

Figura 23. Captura de la pantalla de los datos de un artículo de la revista *American Journal of Political Science* depositados en el repositorio de datos Dataverse.

6. VENTAJAS Y MIEDOS

Después de exponer en los apartados anteriores aspectos relacionados con los datos de investigación y su uso compartido, tales como la perspectiva de políticas institucionales, las infraestructuras y los aspectos más técnicos relacionados con formatos y protocolos, en este apartado se hace un repaso de cuáles son las principales ventajas y miedos percibidos de esta práctica, basándonos en los trabajos de Popkin (2019), Tenopir et al., (2011) y Gewin (2016).

VENTAJAS Y OPORTUNIDADES

Tomando como referencia a Popkin (2019) y a la Red Española sobre Datos de Investigación en Abierto (2018), los principales puntos a favor de la práctica de compartir datos son los siguientes:

- Incrementa la transparencia y la credibilidad de los estudios, ya que, al poder acudir a los datos originales y brutos, se puede comprobar la validez de los resultados. De esta manera, aumenta la confianza en las investigaciones y se despejan posibles dudas que puedan surgir.

- Permite replicar y verificar de manera mucho más rápida los estudios, ya que al contar con los datos brutos y con los detalles de la metodología con la que fueron obtenidos, es relativamente sencillo volver a reproducir el estudio y poder comparar los resultados con el estudio original.

- Fomenta la participación y la colaboración. Esto se debe a que, al situar los datos al alcance de otros investigadores, también se está, de alguna manera, promocionando al investigador o al grupo que realiza esta acción, como se puede comprobar en los ejemplos de capturas de pantalla de repositorios que hemos puesto en este documento. De este modo, permite a los investigadores conocerse entre sí y establecer acuerdos para posibles colaboraciones.

- Reduce costes y tiempo por un motivo obvio, ya que, si se va a realizar un nuevo estudio reutilizando datos, se ahorra el coste económico y humano que supone la recopilación y generación de datos y todo el tiempo que se ha de emplear en ese proceso.

- Permite descubrimientos múltiples, ya que un mismo conjunto de datos se puede explotar para conseguir resultados muy diferentes que, o bien no se les habían ocurrido a los creadores originales, o que éstos no tenían el tiempo o la intención de seguir explotando.

- En el caso de investigaciones con animales, permite ahorrar el sufrimiento y vidas innecesarias al poder acceder a los datos brutos con potencial de reutilización de investigaciones que ya se han llevado a cabo. De esta manera, además, se puede cumplir más eficazmente con las tres "R" establecidas por la European Animal Research Association (2015), que son el estándar que hay que seguir en investigación con animales: Reemplazo (buscar métodos que eviten o ayuden a reemplazar el uso de animales); Reducción (localizar alternativas que ayuden a reducir el número de animales que se usan en experimentos); y Refinamiento (encontrar métodos que ayuden a minimizar cualquier dolor o angustia y mejoren el bienestar animal).

- En el caso de las investigaciones con humanos, sobre todo en el área de la biomedicina, el uso compartido de datos puede ayudar a evitar duplicidades en estudios que, en ocasiones, son muy invasivos e incómodos para los sujetos que forman parte de la muestra.

- Finalmente, incrementa la visibilidad, ya que la publicación de los datos, por ejemplo, en los distintos repositorios de acceso abierto, ofrece una ventana de publicidad para las investigaciones que se estén llevando a cabo desde una perspectiva diferente a la publicación del postprint final. Además, si un conjunto de datos es especialmente llamativo y útil y se reutiliza muchas veces, aparecerá citado cada vez que sea utilizado, lo que también le dará visibilidad e impacto al creador original.

DESVENTAJAS Y MIEDOS

Por otra parte, existen una serie de desventajas muy ligadas a los miedos y recelos de los investigadores, que también han sido destacadas y que deben tenerse en consideración, ya que es necesario comprenderlas para mejorar y pulir las deficiencias. Para hablar de ellas, tendremos en cuenta sobre todo el artículo publicado por Gewin (2016) y Tenopir (2011).

Para ello, dividiremos las preocupaciones en tres bloques: 1) la que afecta a la protección de los datos, 2) la que afecta a la carrera investigadora y 3) la que afecta a los recursos:

1. Sobre la protección de los datos. Es un tema que, en general, produce recelo a todos los investigadores, pero preocupa especialmente a los de Ciencias de la Salud y Ciencias Sociales, ya que son los que suelen manejar en mayor medida datos sensibles que implican las investigaciones con personas. A este respecto, el miedo principal es que esos datos que pueden contener información que identifiquen a los sujetos, acaben en las manos equivocadas y se haga un mal uso de la misma. Esta preocupación afecta, sobre todo, al hecho de no saber hasta qué punto estarán seguros los datos que se comparten y si pueden confiar en que los repositorios serán capaces de garantizar esta seguridad.

2. Sobre las preocupaciones y miedos relacionados con la carrera investigadora, se tocan varios aspectos. El más general es el miedo a que, al compartir los datos, otro investigador pueda hacer uso de ellos de manera más efectiva o rápida y eso le coloque por encima en cuanto a impacto, teniendo en cuenta que el mundo de la ciencia y de la publicación científica es tremendamente competitivo. Según refiere Gewin (2016), este tema afecta especialmente a los investigadores más jóvenes que están empezando a hacerse un currículum, ya que necesitan doblemente consolidar su situación. En relación con esto, una petición de los investigadores es que, si deben compartir los datos, se les de la opción de hacerlo después de haber publicado y no durante la investigación, ya que, por muy beneficioso que pueda ser para la ciencia, les coloca en la posición vulnerable de perder la delantera. Además, investigadores como Tenopir et al., (2011), han encontrado que, en el caso de que compartan los datos, una importante cantidad de investigadores desean tener la opción de hacerles un seguimiento, es decir, que puedan saber en todo momento quién está haciendo qué con los datos que generaron y compartieron. Por otra parte, otro aspecto relacionado con las implicaciones en la carrera investigadora es el aspecto disuasorio que podrían tener declaraciones como la de la influyente revista The New England Journal of Medicine (con un elevado Factor de Impacto), que se refirió a los científicos que usan datos recolectados por otros como "research parasites", alegando que el data sharing debe ser "symbiotically, not parasitically" (Popkin, 2019).

3. En cuanto a los recursos, una preocupación que alegan los investigadores es que, aunque sea una práctica útil y aún en el caso de estar de acuerdo con todas sus ventajas, el hecho es que todo el proceso de compartir datos lleva tiempo y, por consiguiente, tiene un alto coste. Con todo el proceso se refiere a lo que ya se ha explicado en los apartados anteriores de preparación de los datos para que sean inteligibles una vez compartidos (clarificación de los metadatos, búsqueda de repositorios, elección de formatos…), es decir, todo lo que implican los principios FAIR. Teniendo en cuenta el trabajo que de por sí conlleva el proceso investigador, para pedirles a los científicos que depositen sus datos y los compartan se debe tener en cuenta el esfuerzo humano y material que implica, ofreciendo recursos y herramientas con los que puedan hacerle frente a este trabajo.

7. PERSPECTIVAS DE FUTURO

En este apartado sobre las perspectivas de futuro del uso compartido de datos entre investigadores tendremos en cuenta dos perspectivas. La primera, en relación a las desventajas y a los miedos descritos en el apartado anterior y en la capacidad que existe o podría existir de hacerles frente y aportar soluciones. La otra perspectiva es desde el punto de vista de los profesionales de la documentación, ya que su enfoque sobre este tema es determinante para el futuro de esta práctica. Los temas que preocupan y que afectan a las reticencias para compartir datos, relacionados con la protección, la carrera investigadora y los recursos representan un reto, pero también una línea a seguir. En cuanto a la protección de los datos que se comparten, el rumbo ya está marcado y las continuas mejoras en los diferentes repositorios de datos lo demuestran. Por ejemplo, en el caso de Zenodo, que es el repositorio recomendado por la UE creado, mantenido y alojado por una institución del prestigio del CERN, realiza copias de seguridad cada noche de todos los datos y metadatos almacenados y crea múltiples réplicas de los mismos para asegurar su perdurabilidad. Además, ofrece a los usuarios las opciones para depositar sus archivos con distintos tipos de acceso, que puede ser abierto, cerrado o bajo embargo, con lo que demuestra estar cumpliendo con el precepto de "tan abierto como sea posible, tan cerrado como sea necesario" (Zenodo, 2019).

En relación con las preocupaciones y miedos relacionados con la carrera investigadora, es importante visibilizar y promover que el compartir datos no es un "café para todos", sino una práctica que puede y debe adaptarse a las necesidades y peculiaridades de los investigadores, grupos y disciplinas. De esta manera, es importante entender, por ejemplo, las razones por las que un investigador no quiera compartir sus datos durante la investigación, pero sí que esté dispuesto a hacerlo una vez publicada la misma. Desde el punto de vista de las entidades financiadoras y de las instituciones que se encargan de evaluar la ciencia, es relevante que asuman que tienen un papel fundamental en cuanto a la gestión de los incentivos, es decir, entender que pedir a los investigadores que compartan "solo" por favorecer un bien superior como es el avance de la ciencia, puede resultar insuficiente en un mundo que, como se mencionaba, es muy competitivo. Incentivos como puntuación positiva en procesos de acreditación o en la evaluación de proyectos, así como un sistema de conteo de citas a los datos que contribuyan a favorecer el prestigio de los investigadores y grupos es importante para ofrecer un contrapeso a las reticencias de compartir datos, que al final son fruto de mucho esfuerzo. Por ello, es muy importante que se sigan desarrollando métricas que sirvan para evaluación de la publicación y reutilización de datos y, sobre todo, incluir los datos abiertos en los indicadores de evaluación de la actividad científica (Ferrer-Sapena et al., 2016).

Sobre los recursos, también hay un amplio trabajo por delante de cara al futuro. En relación con el párrafo anterior, de la misma manera que es más eficaz pedir a los investigadores que compartan sus datos si la petición viene acompañada de incentivos, también lo es ofrecer medios materiales, económicos y humanos para hacerlo. Estos medios abarcan tanto a las necesidades de formación de los investigadores sobre este tema, como a poner medios a su alcance para hacerlo lo más sencillo posible. En cuanto a recursos económicos, por ejemplo, en el ORD Pilot (European Commission, 2016), que hemos comentado en el apartado de políticas de datos en la UE, se manifiesta que: *"Costs associated with data management, including the creation of a data management plan, can be claimed as eligible costs in any Horizon 2020 grant"*, lo que viene a significar que los grupos de investigación que se acojan a estas directrices tendrán la opción de solicitar financiación para ello.

En cuanto a las perspectivas de futuro desde el punto de vista de los y las profesionales de información y documentación, lo explica muy bien Torres-Salinas en su trabajo "Compartir datos (data sharing) en ciencia: contexto de una oportunidad" (Torres-Salinas, 2010). Aunque es un artículo de hace 9 años, son muy interesantes las preguntas que se plantea en referencia a los profesionales de la información y a la gestión del movimiento por el data sharing. Además, es especialmente útil para constatar que el contexto de oportunidades para profesionales de esta rama del conocimiento sigue siendo parecido en la actualidad. Cuando comienza a hablar de las posibles oportunidades, Torres-Salinas refiere, según el contexto de 2010, que, si los argumentos a favor del uso compartido de datos superan a los argumentos en contra y se convierte en una práctica extendida, se presentarían retos técnicos de manera inevitable. Aunque a lo largo de estos casi diez años se le ha intentado dar respuesta a estos retos, relacionados con aspectos que hemos ido comentando a lo largo de este documento (formatos apropiados para compartir datos, protocolos de metadatos, repositorios y sus peculiaridades, sistema de incentivos, protección de datos o las políticas institucionales y de las revistas), persisten las inseguridades y las reticencias, como también se ha explicado. Según la opinión de este autor y de otros como Hernández-Pérez y García-Moreno (2013), todas estas preguntas generadas a raíz de la promoción del data sharing constituyen un terreno de trabajo y una coyuntura favorable para documentalistas y profesionales de la información, comparable a la que supuso el OA para las bibliotecas universitarias. Un ejemplo de éxito que sirve como ejemplo es el de la Biblioteca de la University of Leeds de Gran Bretaña (https://library.leeds.ac.uk/info/14062/research_data_management/61/research_data_management_explained), que ofrece a los investigadores un servicio muy completo de gestión de datos, con múltiples enlaces, descripciones y herramientas para hacer efectivo el depósito y la reutilización. En la Figura 23 se presenta una captura de pantalla de la página de la biblioteca de la University of Leeds, donde se pueden observar todos los servicios que ofrece en relación a los datos de investigación, a su gestión y al uso compartido.

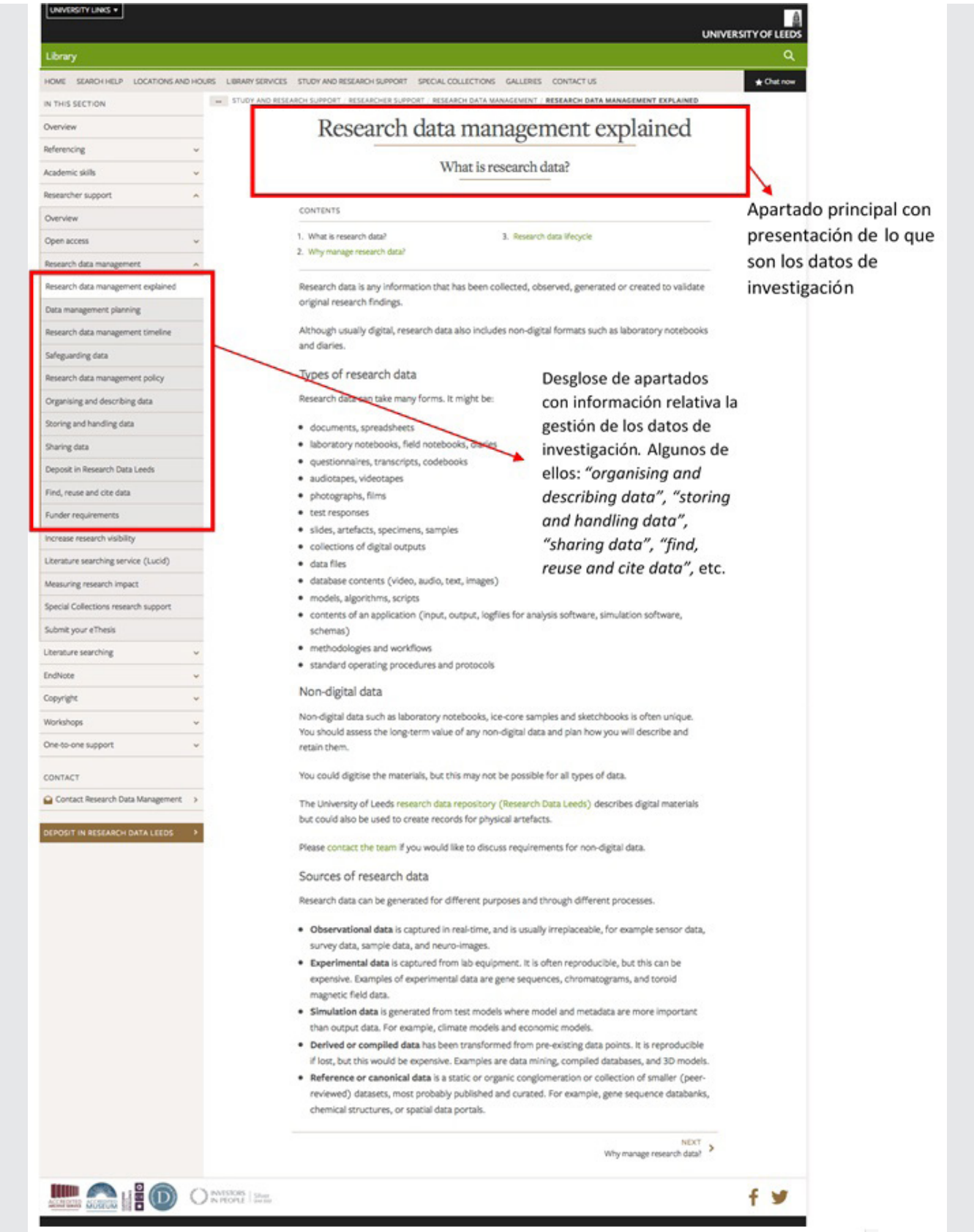

Figura 24. Captura de pantalla de la Biblioteca de la University of Leed.

8. REFERENCIAS BIBLIOGRÁFICAS

Aleixandre-Benavent, R., Lucas-Domínguez, R., Sixto-Costoya, A., Vidal-Infer, A. (2018). The Sharing of Research Data in the Cell y Tissue Engineering Area: Is It a Common Practice?. *Stem Cells and Development*, 27(11), 717–22. **https://doi.org/10.1089/scd.2018.0036**

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., y Ioannidis, J. P. A. (2011). Public availability of published research data in High-Impact journals. *PLoS ONE*, 6(9), 2009–2012. **https://doi.org/10.1371/journal.pone.0024357**

Andaur, G. (2016). Panorama de la gestión de datos de investigación en América Latina y el Caribe. Recuperado de **http://learn-rdm.eu/es/gestion-de-datos-de-investigacion-en-america-latina/**

Australian National Data Service. (s.f.). About us. Recuperado de **https://www.ands.org.au/about-us**

Ball, A., y Duke, M. (2015). *How to Cite Datasets and Link to Publications*. Edinburgh: Digital Curation Centre. **https://doi.org/10.1007/1-4020-5340-1**

Borgman, C. L. (2008). Data, disciplines, and scholarly publishing. *Learned Publishing, 21*(1), 29–38. **https://doi.org/10.1087/095315108X254476**

Canadian Institutes of Health Research. (2015). CIHR Open Access Policy. Recuperado de **http://www.cihr-irsc.gc.ca/e/46068.html#5**

Cho, J. (2019). Subject analysis of LIS data archived in a Figshare using co-occurrence analysis. *Online Information Review,* 43(2), 256–264. **https://doi.org/10.1108/OIR-12-2017-0369**

Comisión Económica para América Latina y el Caribe (2019). Gestión de datos de investigación. Recuperado de **https://biblioguias.cepal.org/gestion-de-datos-de-investigacion/citacion**

Crue Universidades Españolas, y Red de Bibliotecas (REBIUM). (2016). Cita tus datos de investigación. Recuperado de **http://repositori.urv.cat/media/upload/domain_378/arxius/REBIUN_castellà/Cita_tus_datos_de_investigacion.pdf**

Datasea. (s.f.). Preparando un plan de gestión de datos. Recuperado de **http://www.datasea.es/es/research-data/investigador/preparando-plan-gestion-datos**

Diario Oficina de la Unión Europea. Directiva (UE) 2019/1024 del Parlamento Europeo y del Consejo de 20 de junio de 2019 relativa a los datos abiertos y la reutilización de la información del sector público (versión refundida) (2019). Recuperado de **https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1561563110433yuri=CELEX:32019L1024**

Elsevier Connect. (2019). Tipos de Open Access: diferencias entre la "vía verde" y la "vía dorada." Recuperado de **https://www.elsevier.com/es-es/connect/actualidad-sanitaria/tipos-de-open-access-via-verde-y-la-via-dorada**

European Research Council. (2017). *Guidelines on Implementation of Open Access to Scientific Publications and Research Data in projects supported by the European Research Council under Horizon 2020.*

European Commission. (2019). H2020 Online Manual. Recuperado de **https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm**

European Commission. (s.f.). Data management. Recuperado de **https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm#**

European Commission. (2016). H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020, (July 2016). Recuperado de **http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf?utm_content=bufferc22c5yutm_medium=socialyutm_source=twitter.comyutm_campaign=buffer**

European Animal Research Association. (2015). El Principio de las 3Rs. **Recuperado de http://eara.eu/es/el-principio-de-las-3rs/**

Ferrer-Sapena, A., Sánchez-Pérez, E.-A., Aleixandre-Benavent, R., y Peset, F. (2016). Cómo analizar el impacto de los datos de investigación con métricas: modelos y servicios. *El Profesional de La Información, 25*(4), 632-641. Recuperado de **http://www.elprofesionaldelainformacion.com/contenidos/2016/jul/13.pdf**

Generalitat de Catalunya. (s.f.). ¿Qué son los datos abiertos? Recuperado de **http://governobert.gencat.cat/es/dades_obertes/que-son-les-dades-obertes/**

Gewin, V. (2016). An open mind on open data. *Nature, 529*(7584), 117–119. **https://doi.org/10.1038/NJ7584-117A**

Gómez, N.-D., Méndez, E., y Hernández Pérez, T. (2016). Data and metadata research in the social sciences and humanities: An approach from data repositories in these disciplines. *El Profesional de La Información, 25*(4), 16992407. **https://doi.org/10.3145/epi.2016.jul.04**

González-Alcaide, G., y Ferri, J. G. (2014). La colaboración científica: principales líneas de investigación y retos de futuro; Scientific collaboration: main research lines and future challenges. *Revista Española de Documentación Científica 37(4), Octubre-Diciembre 2014, 37*(4), e062. Recuperado en **https://doi.org/10.3989/redc.2014.4.1186**

Henning, P. C., Ribeiro, C. J. S., Da Silva Santos, L. O. B., y Dos Santos, P. X. (2019). GO FAIR e os princípios FAIR: o que representan para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. *Em Questão, 25*(2), 389–412. Recuperado en **https://doi.org/10.19132/1808-5245252.389-412**

Kim, Y., y Burns, C. S. (2016). Norms of data sharing in biological sciences: The roles of metadata, data repository, and journal and funding requirements. *Journal of Information Science, 42*(2), 230–245. **https://doi.org/10.1177/0165551515592098**

Knoppers, B. M., Harris, J. R., Budin Ljøsne, I., y Dove, E. S. (2014). A human rights approach to an international code of conduct for genomic and clinical data sharing. *Hum Genet, 133*, 895–903. **https://doi.org/10.1007/s00439-014-1432-6**

Melero, R. (2005). Acceso abierto a las publicaciones científicas: definición, recursos, copyright e impacto. *El Profesional de La Información, 14*(4), 255–266. Recuperado en **http://www.elprofesionaldelainformacion.com/contenidos/2005/julio/3.pdf**

Michener, W. K. (2015). Ten Simple Rules for Creating a Good Data Management Plan. *PLoS Computational Biology, 11*(10), 1–9. **https://doi.org/10.1371/journal.pcbi.1004525**

Ministerio de Industria, Economía y Competividad. Plan Estatal de Investigación Científica y Tecnológica y de Innovación 2017-2020 (2017). España. Recuperado de **http://www.ciencia.gob.es/stfls/MICINN/Prensa/FICHEROS/2018/PlanEstatalIDI.pdf**

National Institutes of Health. (2015). National Institutes of Health Plan for Increasing Access to Scientific Publications and Digital Scientific Data from NIH Funded Scientific Research. Recuperado de **http://www.ncbi.nlm.nih.gov/pmc/**

National Institutes of Health. (2018). Request for Information (RFI) on Proposed Provisions for a Draft Data Management and Sharing Policy for NIH Funded or Supported Research. Recuperado de **https://grants.nih.gov/grants/guide/notice-files/NOT-OD-19-014.html**

Nature Communications. (2018). Data sharing and the future of science. *Nature Communications, 9*(1), 9–10. **https://doi.org/10.1038/s41467-018-05227-z**

Office Management and Budget. Uniform Administrative Requirements for Grants and Agreements With Institutions of Higher Education, Hospitals, and Other Non-Profit Organizations (1999). EEUU.

OpenAire. (2019). OpenAIRE mission and vission. Recuperado de **https://www.openaire.eu/mission-and-vision**

Peset, F., y González, L.-M. (2017). *Ciencia abierta y gestión de datos de investigación*. Gijón: TREA.

Peset, F., Aleixandre-Benavent, R., Blasco-Gil, Y., y Ferrer-Sapena, A. (2017). Datos abiertos de investigación. Camino recorrido y cuestiones. *Anales de Documentacion, 20*(1), 1–12. **https://doi.org/10.6018/ANALESDOC.20.1.272101**

Popkin, G. (2019). Data sharing and how it can benefit your scientific career. *Nature, 569*(7756), 445–447. **https://doi.org/10.1038/d41586-019-01506-x**

Portal Europeo de Datos. (s.f). Cómo elegir el formato correcto para los datos abiertos. Recuperado de **https://www.europeandataportal.eu/elearning/es/module9/#/id/co-01**

Red Española sobre Datos de Investigación en Abierto (MareData). (2018). *Recomendaciones para la gestión de datos de investigación dirigidas a investigadores*. **https://doi.org/http://hdl.handle.net/10261/173801**

Riley, J. (2017). *Understanding metadata. What is Metadata, and what is for?* Baltimore: National Information Standards Organization (NISO). Recuperado de **http://www.niso.org/apps/group_public/download.php/17446/Understanding Metadata.pdf**

Sociedad Max Planck. (2003). La Declaración de Berlín sobre acceso abierto. *GeoTrópico, 1*(2), 152–154. Recuperado de **https://openaccess.mpg.de/67627/Berlin_sp.pdf**

SPARC Europe y Digital Curation Centre. (2019, August 28). An Analysis of Open Science Policies in Europe v4. Zenodo. **http://DOI.org/10.5281/zenodo.3379705**

Stanley, B., y Stanley, M. (1988). Data sharing: The primary researcher's perspective. *Storage Management Solutions, 12*(2), 173–180.

Thelwall, M., y Kousha, K. (2017). Do journal data sharing mandates work? Life sciences evidence from Dryad. *Aslib Journal of Information Management, 69*(1), 36–45. **https://doi.org/10.1108/AJIM-09-2016-0159**

Torres-salinas, P. D. (2010). Compartir datos (data sharing) en ciencia: contexto de una oportunidad. *Anuario ThinkEPI*, 258–261.

Torres-Salinas, D., Robinson-García, N., y Cabezas-Clavijo, Á. (2012). Compartir los datos de investigación en ciencia: introducción al *data sharing*. *El Profesional de la Información, 21*(2), 173–84. Recuperado en **http://eprints.rclis.org/16786/1/data%20sharing.pdf**

Vidal-Infer, A., Aleixandre-Benavent, R., Lucas-Domínguez, R., y Sixto-Costoya, A. (2019). The availability of raw data in substance abuse scientific journals. *Journal of Substance Use, 24*(1), 1–5. **https//DOI.org/10.1080/14659891.2018.1489905**

Wilkinson, M. D., Dumontier, M., Aalbersberg, J. I., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 1–9. **https://doi.org/10.1038/sdata.2016.18**