UNIVERSIDAD DE SALAMANCA
FACULTAD DE TRADUCCIÓN Y DOCUMENTACIÓN
GRADO EN INFORMACIÓN Y DOCUMENTACIÓN

Trabajo de Fin de Grado

DATA SHARING:

Análisis de repositorios de datos de investigación

Manuel Gutiérrez Gallego

Tutor: Críspulo Travieso Rodríguez

Salamanca, 2015

GUTIÉRREZ GALLEGO, Manuel

DATA SHARING: Análisis de repositorios de datos de investigación/ Manuel Gutiérrez Gallego; Tutor: Críspulo Travieso Rodríguez. – Salamanca : Universidad de Salamanca, Facultad de Traducción y Documentación, 2015, 57h.

Trabajo de fin de grado – Grado en Información y Documentación 1. Bibliometría. 2. Evaluación de la ciencia. 3. Data Sharing. 4. Repositorios de datos. I. Travieso Rodríguez, Críspulo, dir. II. Título

025

RESUMEN

En el contexto actual de la e-ciencia se está generalizando la práctica de compartir los datos de investigación publicados en acceso abierto. Aunque cada vez son más las iniciativas de todo tipo que pretenden impulsar esta práctica, no se logra vencer la reticencia de los investigadores a publicar los datos de sus investigaciones. El objetivo de este trabajo es ofrecer una visión global de la práctica del *data sharing* en la comunidad científica. Se pretende averiguar si los principales repositorios de datos ofrecen información clara a los investigadores que ayude a superar la incertidumbre que supone publicar sus datos. En concreto información sobre sus políticas de datos, cuestiones relativas a su gestión, visibilidad y protección de autoría. Para ello se ha llevado a cabo una revisión bibliográfica y una evaluación y análisis comparativo de los repositorios de datos: *Dryad*, *Genbank*, *SIMBAD*, *PANGAEA*, *Zenodo*, *Figshare* y *wwPDB*. Los resultados muestran que aunque sí ofrecen información básica, esta es escasa y fragmentada, además de existir un amplio margen de desarrollo.

PALABRAS CLAVE: Evaluación de la ciencia, Data sharing, Repositorios de datos, Datos de investigación, Datos primarios, Citación de datos, Métrica de datos.

ABSTRACT

In e-science's current context,sharing published research data in open access is a general practice. Although more and more initiatives of all kinds seeking to promote this practice appear, they fail to overcome the researchers' reluctance to publish their research data. The objective of this paper is to provide an overview of the practice of data sharing in the scientific community in. It questions whether the main repositories of data provide clear information to the researchers to help them overcome the uncertainty that means publishing their data; in particular, information about their data policies, issues management, visibility and protection of authorship. For this we have conducted a bibliographic review and an evaluation and comparative analysis of data repositories: *Dryad*, *Genbank*, *SIMBAD*, *PANGAEA*, *Zenodo*, *Figshare* and *wwPDB*. The results show that even though they offer basic information, it is scarce and fragmented, plus, ample development margin exist.

KEYWORDS: Evaluation of scientific literature, Data sharing, Data repositorie, Research data, Raw research data, Data citation, Data metrics.

ÍNDICE

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

INTRODUCCIÓN

El presente estudio es el resultado de la investigación del alumno bajo la supervisión del tutor Críspulo Travieso Rodríguez. Se trata de la realización del Trabajo de Fin de Grado necesario para poner fin a los estudios de Grado de información y Documentación de la Universidad de Salamanca. Para ello el alumno se ha servido de todos los conocimientos adquiridos a lo largo de la carrera, especialmente de las competencias y contenidos que aporta la asignatura de Bibliometría y evaluación de la ciencia, cursada en el 4º curso. La realización del trabajo ha supuesto también la necesidad de aplicar otras competencias transversales como las relacionadas con metodologías de investigación e inglés.

Para la realización del trabajo se encontró con el problema de que no existía ninguna obra de carácter general que recogiera una visión global del tema del *data sharing*. La mayor parte de la bibliografía se encontraba en inglés y se trataba de artículos puntuales que trataban algún aspecto concreto del *data sharing*. En concreto podríamos hablar, de forma general, de tres tipos de documentos:

- Los documentos que recogían algún pequeño estudio muy delimitado, como por ejemplo algún análisis de una herramienta concreta, alguna encuesta realizada a investigadores, estudios que se referían a algún repositorio de datos específico o a un directorio de repositorios de datos, etc.
- Los que se centraban en reflexionar o mostrar algún problema o ventaja concreta de la práctica del *data sharing* o su relación con el ámbito de la e-ciencia y el acceso abierto.
- Los documentos realizados por instituciones u organizaciones científicas que en algunos casos tenían relación con alguna administración pública. Todo tipo de informes, directrices y políticas, etc.

Lo que en principio suponía un problema paso a convertirse en la oportunidad de realizar un trabajo que permitiera ofrecer una modesta visión global del panorama que ofrece el *data sharing*.

El trabajo está estructurado en cuatro partes:

- Una primera parte muestra los objetivos, justificación y la metodología.
- La segunda parte es el marco teórico o estado del arte del *data sharing* que muestra una visión global del *data sharin*g.
- La tercera parte es el estudio que analiza los repositorios de datos elegidos para saber si ofrecen o no información clara que ayude a los investigadores a superar su reticencia a publicar los datos de sus investigaciones.
- Por último estarían las conclusiones finales.

1. OBJETIVOS, JUSTIFICACIÓN Y METODOLOGÍA.

1.1 OBJETIVOS Y JUSTIFICACIÓN.

El **objetivo general** es ofrecer una visión global de la práctica del *data sharing* en el contexto actual de la *open science* revisando la literatura científica y analizando los principales repositorios de datos de investigación. Se realiza una evaluación y análisis de siete de los principales repositorios de datos de investigación para comparar la información que aportan a los investigadores sobre la publicación de datos en sus repositorios.

Uno de los principales problemas a los que se enfrenta el *data sharing* es la reticencia de los investigadores a publicar sus datos de investigación para que sean compartidos con el resto de la comunidad científica. Ante este problema interesa saber si los principales repositorios de datos de investigación ayudan a paliar esta situación. En concreto, se pretende averiguar si muestran información clara y concisa para evitar la incertidumbre entre los investigadores que se decidan a publicar los datos de sus investigaciones.

Para cumplir con este objetivo se han de cumplir los siguientes **objetivos específicos**:

- Contextualizar el *data sharing* en la situación actual de e-ciencia.
- Entender qué son los datos de investigación. Sus diferentes tipos, formatos y origen.
- Conocer los antecedentes, evolución y conceptualización del *data sharing*.
- Ofrecer un panorama general de los repositorios de datos existentes.
- Aplicar un modelo de evaluación sobre los principales aspectos detectados en la bibliografía relacionados con la comunicación de información sobre sus políticas de datos, preservación y reutilización, aspectos legales y protección de autoría. También aspectos relacionados con la visibilidad, impacto y reconocimiento de los investigadores.

La gran cantidad de datos que se están generando continuamente en todas partes gracias las nuevas tecnologías junto con la gran capacidad de almacenamiento y gestión de datos masivos de los últimos años da lugar al *big data*. Los datos se generan en muchos ámbitos distintos y la mayoría sin que nos demos cuenta. En la ciencia siempre se han generado datos y se les puede considerar la base fundamental de esta. En la actualidad se está creando un nuevo entorno mucho más abierto del que forma parte el *data sharing.* Si se generaliza esta práctica, se tendrá acceso a una gran cantidad de datos de investigación que en su conjunto se podrán considerar datos masivos. Incluso los datos generados al realizar el *data sharing* también contarán para revelar realidades. Esta perspectiva va más allá de los aspectos del *data sharing* que pueden interesar, de forma más inmediata, a los investigadores; como son el el acceder y disponer de datos de investigación de forma fácil y gratuita. Aún así, es lo que ha motivado el interés por averiguar más sobre este tema y conocer que proyeción y alcance ofrece el entorno del *data sharing* para los profesionales de la información.

1.2. METODOLOGÍA.

La realización del trabajo se llevó a cabo en dos fases. Primero se realizó una búsqueda bibliográfica en las bases de datos *Scopus*, *Web Of Science*, *LISA* y la base de datos del *CSIC*. Para una primera búsqueda se empleó el término *data sharing** y luego se realizaron una serie de búsquedas posteriores combinándolo con otros términos relacionados semánticamente como: *research data*, raw research *data*, *data repositories, data citation, data metrics, open data,* entre otros. El objetivo era obtener una amplia bibliografía para revisar lo más ampliamente posible todo lo que la literatura científica ofrecía sobre el *data sharing*. También se realizó una búsqueda en *Google Scholar* para revisar distintas publicaciones de diverso tipo que hablaran sobre *data sharing* o temas relacionados. Una vez revisada toda la bibliografía y literatura científica se procedió a redactar la parte teórica del trabajo que ofrece una visión global del concepto de *data sharing.*

La segunda fase es la que tiene que ver con la evaluación y análisis comparativo de una serie de repositorios de datos de investigación para completar el estado de la cuestión con una aproximación práctica al objeto de estudio. Primero se procedió a decidir qué repositorios serían los elegidos para el análisis. Esta selección se llevo a cabo teniendo en cuenta la bibliografía revisada previamente. Los siete repositorios elegidos son: *Dryad*, *Genbank*, *PANGAEA*, *SIMBAD*, *Zenodo*, *Figshare* y *wwwPDB*. Son los repositorios que más veces aparecen mencionados y que son considerados como importantes referentes en sus campos o disciplinas; suelen coincidir además con los más veteranos. También se incluyen los de reciente creación que se consideran referentes a tener en cuenta por sus características innovadoras y su potencial; como son *Zenodo* y *Figshare*. Todos son de carácter internacional ya que se descartó la idea de seleccionar alguno de los pocos repositorios de datos españoles por su escasa repercusión. Además todos pertenecen también, junto con otros, a la selección que hizo la revista PLOS (*Public Library Of Science*) de los repositorios de datos más confiables; recomendando su uso a los investigadores.

Una vez escogidos los repositorios se procedió a realizar una breve descripción de los mismos. Para ello, se elaboró un cuestionario que recogería 29 criterios necesarios para evaluar y comparar los repositorios; cuya respuesta era del tipo SI/NO en la mayoría de ellos. Para la elaboración de este cuestionario se tuvo en cuenta la *Guía para la evaluación de repositorios institucionales científicos* (2014). Se trata de una guía que ofrece los recursos necesarios para evaluar la calidad de los repositorios institucionales científicos. Se trabajó con repositorios de datos de investigación así que hubo que llevar a cabo una adaptación.

Para elaborar los criterios nos centramos en la necesidad que se tenía de evaluar la calidad de la información que ofrecían los repositorios de datos científicos desde el punto de vista de la promoción de la cultura del *data sharing.* En concreto, los aspectos relevantes de la información que ofrecen los repositorios de datos que podría ayudar a evitar la reticencia de los investigadores a publicar sus datos.

Se establecieron 6 categorías que recogían los 29 criterios de evaluación:

Políticas, cuyos criterios van enfocados a determinar si el repositorio ofrece o no información clara sobre sus políticas de datos, sobre todo en cuanto a preservación y reutilización.

Aspectos legales y protección de autoría, para determinar si el repositorio ofrece información que pueda aportar garantías de autoría al investigador depositante.

Accesibilidad y disponibilidad, para saber si cumple con uno de los requisitos fundamentales del *data sharing,* este es*,* que los datos estén fácilmente accesibles y disponibles*.*

Visibilidad, Impacto y reconocimiento, para averiguar si ofrece información que sirva de ayuda al investigador depositante para el aumento del impacto de su investigación.

Seguridad, autenticidad e integridad de los datos, cuyos criterios van enfocados a averiguar si provee información sobre estas garantías concretas.

Interfaz y software, determinar si utiliza herramientas que provean una navegación fácil, útil y sencilla.

Una vez se tuvo elaborado el cuestionario se procedió a realizar el análisis de los repositorios y se elaboró una tabla comparativa que contiene las respuestas a los criterios de evaluación establecidos previamente. Por último se redactó el análisis y conclusiones de la tabla obtenida.

Para la elaboración de la bibliografía se siguió la norma UNE-ISO 690:2013

2. APROXIMACIÓN AL CONCEPTO DATA SHARING.

2.1. CONTEXTO DE LA E-CIENCIA.

En el contexto de la revolución digital que se ha dado en los últimos años se han producido nuevas formas y prácticas de acceso y uso de la información científica.

El acceso abierto, más conocido como *open acces*, es un movimiento que pretende hacer disponible el conocimiento científico permitiendo que cualquier usuario pueda acceder a la literatura científica de forma abierta y gratuita con lo que se amplia la distribución y difusión del conocimiento científico. Esto reporta beneficios tanto para el autor como para la sociedad en general: para el primero la investigación puede ser más consultada y citada, y, para la sociedad, se permite acceder a los resultados de las investigaciones financiadas con dinero público. El acceso abierto no es sólo acceder sin restricciones a las publicaciones sino también la posibilidad de su reutilización y de poder compartirlas. Gracias al acceso abierto una enorme cantidad de recursos distribuidos por Internet están accesibles para la la comunidad científica de una forma mucho más accesible; en gran medida debido al desarrollo de aplicaciones e infraestructuras que lo permiten. Es lo que se conoce como e-infraestructuras (infraestructuras electrónicas).

Las actividades científicas desarrolladas con los recursos distribuidos en internet en acceso abierto y apoyadas en las e-infraestructuras han dado lugar a la e-ciencia la cual está cambiando las prácticas de los investigadores de todas las áreas del conocimiento. Según González et al. (2013) e-ciencia son "los métodos de trabajo colaborativos con un uso intensivo de tecnología". Los fenómenos de acceso abierto y la e-ciencia permiten que se genere y transfiera conocimiento de una forma nueva que incita a la colaboración y que crea una nueva visión de la ciencia, ya que no se trata sólo de acceder a las publicaciones sino además de crear un contexto nuevo más transparente, visible y eficiente con vistas a ahorrar esfuerzos y duplicidades y evitar fraudes. Dicho de otro modo, puede ayudar a generar conocimiento más rápido pero también más fiable ya que se puede detectar antes errores, fallos o fraude.

Puede parecer que el acceso abierto sólo se limita a las publicaciones científicas pero los datos también han cobrado relevancia en los últimos años en el contexto de la e-ciencia, surgiendo muchas iniciativas que abogan por ponerlos en acceso abierto como recursos digitales con entidad propia. El *open data* o datos abiertos es esa filosofía, corriente o movimiento dentro del contexto del acceso abierto que trata de que los datos sean también accesibles a los ciudadanos sin ningún tipo de restricción, ya que estos son importantes y cada vez cobran más relevancia no sólo en la ciencia sino en todos los ámbitos de la sociedad. No es extraño encontrarse últimamente con todo tipo de definiciones y conceptos relacionados con ellos: *Big data* (datos masivos), *open data* (datos abiertos), *data sharing* (datos compartidos), *data metrics* (métrica de datos), etc. Es importante tener en cuenta que no todos los datos se pueden considerar datos abiertos por el mero hecho de estar disponibles para su descarga en internet. Además del acceso, deben cumplir una serie de condiciones entre los que destacan las licencias legales para poderlos usar y reutilizar y que las mantengan al ser explotados y redistribuidos incluso en las obras derivadas de ellos. Un requisito importante es el de reconocer la autoría de los mismos (atribución). Son accesibles y se pueden reutilizar pero la reutilización suele estar controlada por las licencias.

El *Open Definition Advisory Coun*cil expone ciertos factores imprescindibles para que los datos sean considerados datos abiertos:

> *"...acceso (disponible integramente, a un coste razonable y de forma que pueda ser modificable), redistribución, reutilización, ausencia de restricciones tecnológicas, reconocimiento, integridad, sin discriminación de personas o grupos, sin discriminación de ámbitos de trabajo, distribución de la licencia, la licencia no debe ser específica de un paquete y la licencia no debe restringir la distribución de otras obras."* (Hernández-Pérez y García-Moreno, 2013)

Por otra parte, el esquema de 5 estrellas para datos abiertos de Tim Berners-Lee establece 5 categorías:

- 1 estrella: Publica tus datos en la Web (con cualquier formato) y bajo una licencia abierta.
- 2 estrellas: Publicados como datos estructurados (ej: Excel en vez de una imagen de una tabla escaneada)
- 3 estrellas: Usa formatos no propietarios (ej: CSV en vez de Excel)
- 4 estrellas: Usa URIs para denotar cosas.
- 5 estrellas: Enlaza tus datos a otros datos para proveer contexto.

Según Hernández-Pérez y García-Moreno (2013), a la hora de hablar de datos abiertos también hay que diferenciar entre datos gubernamentales y datos científicos. Los datos gubernamentales son los que generan y recopilan las administraciones públicas; aunque algunos datos científicos también podrían considerarse gubernamentales debido a que han sido producidos por investigaciones científicas financiadas por los gobiernos. Los datos abiertos de los gobiernos cada vez tienen más relevancia ya que su disponibilidad es síntoma de transparencia y participación social a la vez que promueve el poner a disposición del público datos reutilizables para el emprendimiento económico por sus potenciales usos comerciales. Son una fuente de recursos tanto para generar sociedades más democráticas, transparentes y participativas como para impulsar nuevas iniciativas y usos comerciales.

Los datos científicos por su parte tienen que ver en gran medida con los resultados de las investigaciones científicas, así como los datos generados durante el proceso de dichas investigaciones, que aunque siempre se han publicado debido a ser parte esencial de las publicaciones científicas en la actualidad está cobrando mucha relevancia el hecho de que estos datos sean publicados como recursos propios para que puedan ser compartidos y reutilizados por la comunidad científica y la sociedad.

2.2. LOS DATOS DE INVESTIGACIÓN.

Antes de abordar el concepto de *data sharing* hay que tener en cuenta qué son los datos de investigación y cuál es su tipología.

Los datos de investigación han sido centro de debate debido a su relevancia dentro de la comunidad científica. Desde hace tiempo se vienen considerando como recursos propios independientes de las publicaciones científicas. Estos datos sirven para dar validez a las investigaciones y publicaciones científicas a la vez que pueden servir a otros miembros de la comunidad científica para generar nuevos conocimientos y nuevos datos gracias a su reutilización. Pero el problema o la complejidad de está cuestión no está en la posibilidad de considerar a los datos científicos como recursos o fuentes propias independientes de las publicaciones científicas sino que, una vez considerados de tal forma, surge el problema de como asegurar su disponibilidad y accesibilidad al igual que lo están las publicaciones y artículos científicos. Los datos de investigación son variados y tienen una gran dependencia tanto de la disciplina del conocimiento a la que pertenecen o donde se generan como al ciclo de vida que se da dentro del proceso de investigación. Otra variable condicionante de los datos de investigación son el tipo de requisitos tanto legales como técnicos a los que están sujetos.

El problema de su explotación empieza ya en el hecho mismo de su naturaleza por lo que incluso definir qué es un dato de investigación no resulta tan simple y sencillo como cabe esperar, debido a su carácter heterogéneo y a una dependencia sujeta a toda la variedad de propósitos y procesos de las investigaciones; cuando hablamos de datos científicos hablamos de objetos complejos que son la esencia de una investigación o publicación.

Es importante tener claro qué se entiende por dato de investigación pues es el primer paso para que se puedan compartir de forma eficiente y eficaz, para que funcione el *data sharing*. Las definiciones sobre lo que es un dato de investigación son muchas y variadas pero todas se encuadran en uno de los dos puntos de vista generalizados diferenciados por los que consideran datos de investigación todos los datos recogidos durante el proceso de investigación y los que sólo consideran datos de investigación los datos finales de la investigación por ser quienes la validan. A continuación mostramos dos ejemplos de cada punto de vista:

Según el Grupo de trabajo de "depósito y gestión de datos en acceso abierto" del proyecto RECOLECTA (2012), la Universidad Australiana de Melbourne aporta la siguiente definición sobre datos de investigación:

> *"Los datos de la investigación son hechos, observaciones o experiencias en que se basa el argumento, la teoría o la prueba. Los datos pueden ser numéricos, descriptivos o visuales. Los datos pueden ser en estado bruto o analizado, pueden ser experimentales u observacionales."* (Grupo de trabajo de "depósito y gestión de datos en acceso abierto" del proyecto RECOLECTA, 2012)

Además señala que se puede incluir como datos de investigación todos los recogidos durante el proceso de investigación en cualquier formato o estado. Eso incluye todo tipo de cuadernos de trabajo (de laboratorio o de campo), datos de investigación primaria (tanto en papel como en soporte informático), cuestionarios, cintas de audio, vídeos, fotografías, películas, etc. Las colecciones datos para la investigación pueden incluir diapositivas; diseños y muestras. Considera datos de investigación hasta el código de software utilizado para generar, comentar o analizar los datos.

Según Torres-Salinas, Robison-García y Cabezas-Clavijo (2012), una de las definiciones que más consenso tiene y que ha sido adoptada por diversas entidades como el National Institutes of Health (NIH) de EEUU o la OECD (2007), es la que considera dato de investigación todo el material registrado durante la investigación y que sirva para apoyar y certificar los resultados de la investigación.

También especifica que debe provenir de una fuente única y deben ser difíciles o imposibles de obtener de nuevo por pertenecer a un momento o circunstancias irrepetibles de una forma exactamente igual. Esta última definición no considera datos de investigación los recogidos durante el proceso de investigación como borradores, notas, análisis, etc. Sólo se centra en los datos finales y sus características. Esta definición choca con otras opiniones que tienen en cuenta también los datos recogidos durante el proceso de investigación ya que también pueden ser relevantes y ayudar a hacer la ciencia más eficiente reduciendo los esfuerzos y costes económicos.

Como hemos comprobado, la variedad es una de las características de los datos de investigación y esto influye a la hora de intentar hacer una clasificación de estos. Son muchas las maneras de categorizarlos; una podría ser según su soporte (papel, digital) también su formato (texto, números, imágenes...) si son descriptivos: cualitativos o cuantitativos, geográficos, espaciales, etc. Pero en el contexto del *data sharing,* donde prima el compartir los datos, es importante conocer qué se va a compartir por lo que una de las clasificaciones que se impone, según Torres-Salinas, Robinson-García y Cabezas-Clavijo (2012), es la clasificación que ofrece la *Research Information Network,* la cual establece tres criterios o tipos de datos para categorizarlos ofreciendo una visión de su variedad y de las necesidades que comportan a la hora de su gestión. Este esquema presenta los datos según su origen u obtención, según el objetivo por el cual se han recogido los datos y según el tratamiento que han recibido:

Por su origen. Aquí podemos encontrar tres tipos: Experimentales, observacionales y simulaciones.

Los *experimentales* son los datos que se dan a lo largo de una investigación desde su inicio hasta sus resultados, es decir, los datos que se obtienen como resultado de un experimento. Suelen ser reproducibles. Abarca todos los datos registrados desde la planificación de la investigación hasta que se obtienen los resultados.
Por ejemplo; cualquier investigación en un laboratorio químico que durante su proceso lleve una serie de experimentos para obtener su objetivo producirá datos durante su proceso de investigación y experimentación como pueden ser el control y registro de reacciones químicas y evolución del estado físico de la materia cada cierto periodo de tiempo.

Los *observacionales* son los datos fruto de la observación de un hecho o fenómeno concreto lo que los convierte en datos únicos. Son registros históricos que sólo se pueden conseguir una vez por lo que su preservación es muy importante ya que si se pierden es una información que no se puede reemplazar ni volver a conseguir debido a que se obtienen de forma directa en un lugar y tiempo concretos. A diferencia de los experimentales estos no se pueden repetir. Por ejemplo; una encuesta de opinión hecha por sociólogos en Albacete en 1986 si se pierde no se puede volver a repetir.

Las *simulaciones* son los datos computacionales que acompañan a la aplicación de simulaciones o modelos dónde lo que menos importa son los resultados del modelo ya que se pueden reproducir si se tiene los datos de entrada y los programas que los generan. Por lo tanto consta en gran parte de metadatos y el modelo que se aplique. Por ejemplo; un modelo informático para predecir el clima. No importa los resultados ni los datos de entrada y salida sino el modelo en sí.

Por el objetivo por el cual se han recogido los datos. Aquí también encontramos tres tipos: específicos, de alcance medio y de interés general.

Datos específicos: Son datos cuyo alcance está limitado a un proyecto concreto y su valor no va más allá de servir a los objetivos de ese proyecto. Son necesarios y tienen valor sólo para el investigador para llevar a cabo su investigación o en concreto una parte de ella durante el proceso de investigación.

Datos de alcance medio: Serían datos similares a los específicos pero su alcance abarca a una comunidad de investigadores de una misma disciplina o especialidad a los cuales les podría servir para sus investigaciones.

Datos de interés general: Son los más relevantes porque se trata de aquellos cuyo valor va más allá de la utilidad para un investigador o para el desarrollo de una disciplina concreta. Son los datos que son importantes para toda la comunidad científica y el desarrollo de la ciencia.

Por el tratamiento que han recibido. Si son datos brutos o primarios que aún no han recibido ningún tratamiento podemos hablar de datos primarios. (*raw data*)

Cuando los datos primarios han sido manipulados, procesados y combinados dando como resultado datos que justifican la investigación llevada a cabo y que son la razón de ser de dicha investigación hablamos de datos finales de la investigación. (*final research data*).

Esta tipología de datos presentada en este trabajo responde a un intento de categorizar los datos de investigación de una forma que ofrezcan una visión general pero es una clasificación que sirve simplemente para hacerse una idea general del contexto de los datos de investigación y sus características más comunes lo cual servirá a la hora de crear un estándar para poder compartirlos mejor. Pero como hemos mencionado anteriormente, la variedad de datos y su dependencia de las respectivas disciplinas a las que pertenezca el campo de investigación y la metodología llevada a cabo en las investigaciones generará datos que puede que no se adscriban exactamente a esta tipología con lo cual quedará en manos de cada disciplina qué categorización es la que mejor les conviene para establecer que se debe compartir o no.

2.3 DATA SHARING.

2.3.1. ANTECEDENTES Y ORIGEN DEL DATA SHARING.

Si los datos de investigación y en concreto los datos finales de la investigación son considerados la base material imprescindible de la que se nutre la e-ciencia es lógico pensar que un acceso abierto a ellos facilitará la labor de muchos investigadores que ahorrarán tiempo y recursos a la hora de llevar a cabo sus propios proyectos gracias a la reutilización de estos datos. Compartir los datos de forma abierta los vuelve accesibles y disponibles de forma rápida, lo que puede hacer que el coste de una investigación sea menor tanto en esfuerzos como desde un punto de vista meramente económico, algo relevante en el contexto actual de crisis económica donde la inversión en ciencia es mucho menor. También se presupone que esto aceleraría los avances en ciencia y su progreso, además de dotarla de una mayor transparencia evitando el fraude en las publicaciones. Teniendo en cuenta lo anterior vemos como compartir los datos finales de las investigaciones se ha convertido en una necesidad dentro de la comunidad científica por lo que el debate acerca del *data sharing* continúa creciendo.

En los últimos años el debate de si compartir o no los datos finales de investigación y cómo hacerlo para que puedan ser reutilizados de una forma que contente a todos y no perjudique a los autores ha ido en aumento. Se puede ver en el hecho de que ya son numerosas las revistas prestigiosas que han tratado el tema e incluso han dedicado números especiales, destacando el de la revista *Nature* de 2009 que para muchos ha marcado un punto de inflexión en la difusión de esta cuestión.

Pero aunque pueda parecer que es un tema actual, debido al contexto digital que ofrecen las nuevas tecnologías de la información, lo cierto es que la necesidad de compartir datos de investigación es un tema que ya se ha tratado anteriormente. Buscando antedecentes históricos sobre este debate nos podemos remontar hasta 1901 donde según Torres-Salinas, Robinson-García y Cabezas-Clavijo (2012) Galton (a quien también se asocia al nacimiento de la cienciometría) señalaba en la revista *Biometrika* que nadie debería publicar resultados sin depositar una copia de sus datos en algún lugar para poder ser consultados y comprobados por otros interesados. Dejando a un lado el debate sobre la cuestión del *data sharing* y centrándonos más en los hechos históricos observamos que el punto clave es la creación de los bancos de datos; en concreto el *Protein data bank* (PDB) en 1971.

También es relevante que en 1983 la revista *Journal of biological chemistry* fue la primera en exigir los datos de las investigaciones como requisito para la publicación. Torres-Salinas, Robinson-García y Cabezas-Clavijo (2012) siguiendo a Hrynaszkiewicz (2009) nos recuerda que mucho antes de la existencia de los repositorios, como herramienta imprescindible para el *data sharing*, la comunidad científica ya compartía datos como práctica inherente a la ciencia y nos señala como ya era habitual el uso de la vía informal a través de canales no controlados señalando las Ciencias Sociales y Humanas como ejemplos de disciplinas que no disponían de repositorios, por lo que era habitual recurrir a esta práctica que según Torres-Salinas, Robinson-García y Cabezas-Clavijo (2012) siguiendo a Pinowar (2008) muestra las dos vías informales básicas:

- Por petición a otros colegas compartiendo los datos de forma privada.

- De forma descentralizada; disponiendo de los datos en páginas webs personales de investigadores o de grupos de investigación.

Según un informe de Costas et al. (2013) el *data sharing* cobra relevancia a partir de los 80 destacando que a mediados de esta década ya había gran variedad de informes que recogían los beneficios y problemas a los que se enfrentaba el intercambio de datos científicos y la importancia de este en el desarrollo científico, señalando algunas buenas prácticas, directrices e incluso que actores y políticas serán necesarias para su proceso.

Por tanto, aunque el término y debate del *data sharing* pueda parecer algo surgido en el contexto de la ciencia actual, como vemos en los párrafos anteriores, lo cierto es que esta necesidad de compartir los datos de investigación se remonta mucho tiempo atrás y es una cuestión que aparece de forma casi periódica a lo largo del siglo XX recobrando interés según avanza la ciencia y aparecen nuevas tecnologías que lo permiten de una forma más eficaz y eficiente.

2.3.2. DELIMITACIÓN DEL CONCEPTO.

Cuando Costas et al. (2013) trataron el estado del arte de la investigación sobre el *data sharing* vieron que al realizar una búsqueda exploratoria en la base de datos de *Web of Science* con el término de búsqueda “data sharing*” obtuvieron un total de 1460 resultados observando que la práctica totalidad de la literatura al respecto respondía a tres grupos de términos concretos y bien diferenciados una vez que realizaron un mapa conceptual con todos los términos. Por un lado encontraron términos relacionados con la investigación, las políticas, la ciencia y el trabajo científico en general. Por otro lado un grupo de términos técnicos y tecnológicos, incluyendo términos como "aplicación", "sistema", "usuario", también "ordenador", "arquitectura", "algoritmo ", etc. Un tercer grupo más pequeño comprendía términos relacionados con los propios datos: "datos brutos", "datos científicos", "datos experimentales",etc. Al observar esto llegaron a la conclusión de que la investigación sobre *data sharing* o que implique el *data sharing* comprende dos dimensiones: una científica-política y otra de carácter tecnológico. Y entendieron el tercer grupo como la conjunción práctica de las otras dos dimensiones, a través de la organización de los repositorios y la consideración de las motivaciones de los científicos a compartir sus datos a través de estos nuevos enfoques y tecnologías.

Atendiendo a la exploración anterior sobre la literatura científica al respecto uno se podría hacer una idea del estado actual de la cuestión del *data sharing* viendo que hay un largo recorrido de investigaciones sobre cuestiones de carácter científico, político y tecnológico y uno menor, en estado embrionario, sobre nuevos enfoques y prácticas a la hora de materializar el intercambio de datos que correspondería con el contexto actual de la e-ciencia.

Ante el concepto *data sharing*[1] no encontramos ninguna definición específica de ningún autor que muestre el verdadero alcance del término con todos sus matices y apreciaciones. Las definiciones son escasas y muy escueta. Como concepto general nos encontramos con definiciones del tipo: "la publicación de los datos de investigación para su uso por los demás" (Borgman, 2012), “La acción de compartir con otros colegas los ficheros de datos (o raw data) generados durante el curso de una investigación.” (Torres-Salinas, 2010). Desde un punto de vista etimológico son definiciones correctas que van a la par con el término *data sharing* (que simplemente significa compartir datos).

Pero para entender el verdadero alcance del concepto, y en concreto de la práctica del *data sharing,* hay que ir más allá del simple hecho de compartir datos. Por esta razón, se ha creído conveniente señalar una serie de elementos y factores comunes encontrados en la literatura científica acerca de esta práctica en un intento de mostrar un alcance más amplio y matizado. Se parte de la idea de que todos estos elementos en conjunto configuran y confieren carácter al simple hecho de compartir los datos; que es en lo único en que coinciden las escasas definiciones o menciones a este concepto. Estos serían:

- Los datos como recurso propio. Los datos finales de la investigación (también algunos datos de tipo primario) considerados la parte fundamental de la ciencia se empiezan a tratar como un recurso propio, independiente de la publicación o artículo científico al que van ligados como parte esencial de la investigación.

- El contexto actual en el que se desarrolla la e-ciencia. Caracterizado por las nuevas e-infraestructuras y actitudes de los investigadores que en el contexto del o*pen access* y o*pen data* hacen que se desarrolle una nueva estrategia que ayuda a crear una ciencia más abierta *(open science)*. Se están generando nuevas prácticas que hacen que el desarrollo de la ciencia se haga de forma más colaborativa, más transparente, reduciendo los obstáculos que dificultan a los investigadores el desarrollo de sus proyectos de una forma más eficaz y eficiente. Como resultado se genera, de forma general, una buena práctica científica. Aunque todavía queda camino por recorrer y materializar.

- El acceso y disponibilidad de forma accesible de los datos finales de investigación. Deben ser fáciles de encontrar, acceder y disponer de ellos de forma fácil y gratuita. Por esto se desarrollan iniciativas tanto de protección de los autores y gestión de licencias como tecnológicas que evitan restricciones y fomentan la interoperabilidad.

- La provisión voluntaria de los datos por los investigadores. Poco a poco van superando las reticencias a publicarlos, pasando de publicarlos sólo dentro de sus respectivas disciplinas y especialidades a publicarlo de forma que estén disponibles para toda la comunidad científica. Cada vez más la publicación no se debe solo la voluntariedad de los investigadores sino a imposiciones o requisitos de algunas agencias o administraciones públicas que lo imponen como condición para recibir su financiación.

1 En castellano el término *Data sharing* significa compartir datos. Para el trabajo se ha preferido utilizar siempre el término en inglés por tener un mayor alcance. Abarca aspectos más allá del simple hecho de compartir los datos.

- La reutilización de los datos de investigación. Se publican de forma accesible y disponible para que puedan ser compartidos y reutilizados por la demás comunidad científica. Se genera el debate de la protección de autoría de los datos (atribución) y se ponen en marcha iniciativas e incentivos para reducir las reticencias a compartirlos.

- Colaboración científica. El intercambio de los datos necesita de la colaboración por lo que compartir es una nueva práctica más extendida que ayuda al progreso de la ciencia reduciendo costes económicos y esfuerzos que evitan las duplicidades.

- Transparencia de la ciencia. Se evita más el fraude y se detectan antes los errores.

Atendiendo a todos estos elementos y factores comunes se intenta ahora ofrecer una definición de *data sharing.* El *data sharing* como la acción o práctica de publicar de forma voluntaria los datos de investigación como recursos propios independientes (tanto los datos finales de las investigaciones como los de carácter primario generados en el transcurso de una investigación que potencialmente puedan tener alguna relevancia y utilidad) con el objetivo de que puedan ser intercambiados y compartidos; principalmente por la comunidad científica pero también por la sociedad en general. Implica su reutilización, con la finalidad de que sirva para el avance de la ciencia gracias al ahorro de tiempo, esfuerzos y costes económicos, lo que supone la colaboración al compartir los datos y su fortalecimiento gracias a la transparencia que ofrece esta práctica al facilitar que se pueda detectar mucho antes errores en las investigaciones o el fraude mismo. Por lo que el *data sharing* no es sólo una práctica o movimiento colaborativo que facilita el intercambio de datos sino también un mecanismo de transparencia que ayuda al control de la ciencia fortaleciéndola.

Todo esto conlleva que los datos deben ser tratados como recursos propios independientes, es decir, como objetos digitales independientes de la publicación o investigación a la que van ligados por lo que necesitan de su propio tratamiento para su preservación, distribución y difusión y de un plan de gestión de datos atendiendo también a su propio ciclo de vida. También deben estar accesibles, siendo fáciles de encontrar, acceder y disponer de ellos de forma fácil y gratuita y de una actitud más colaborativa y voluntariosa por parte de los investigadores que deben participar de esta nueva cultura generada en el contexto de la e-ciencia. A ello hay que sumar la intervención de la administración pública que debe seguir desarrollando iniciativas encaminadas a exigir la publicación de los datos de investigación de todas aquellas investigaciones llevadas a cabo bajo la ayuda de su financiación.

2.4. PLAN DE GESTIÓN DE DATOS.

Para facilitar la reutilización, los datos necesitan de unas infraestructuras (repositorios, etc.) que faciliten su almacenamiento, preservación, distribución y difusión para el intercambio, así como una de serie de iniciativas tanto políticas como culturales que generen el marco que pueda permitir una correcta gestión de los datos. Iniciativas políticas que faciliten, incentiven o impongan la publicación de los datos y que también protejan la autoría de los datos e iniciativas culturales que promuevan un cambio de mentalidad y actitud más colaborativa que se deshaga de las reticencias a compartir los datos por parte de los autores.

De este desarrollo tecnológico, político y cultural enfocado a la promoción del intercambio de datos deben salir los estándares y referencias que guíen a la comunidad científica a la hora de realizar una buena gestión de datos.

El contexto tecnológico que crean las e-infraestructuras y el político/cultural facilitan el intercambio y reutilización de los datos pero una vez se dispone de este contexto, teniendo en cuenta que los datos o *datasets* son objetos digitales independientes, es necesario atender a su tratamiento. Por lo cual, para que la gestión de los datos se desarrolle de forma correcta y se pueda facilitar su reutilización es importante elaborar un plan de gestión de datos.

Antes de tratar en asunto del plan de gestión de datos es conveniente tener en cuenta que actores principales son los involucrados en la gestión de los datos científicos:

Científicos e investigadores	Crean, producen y utilizan datos de investigación
Gobiernos y sector público.	Financian proyectos de investigación
Editoriales y repositorios de datos	Difusión de las investigaciones científicas

Tabla 1. Actores más relevantes en la gestión de datos.

El Grupo de trabajo de "depósito y gestión de datos en acceso abierto" del proyecto RECOLECTA (2012) menciona los siguientes agentes involucrados en la gestión de datos de investigación:

Investigadores/productores de datos	Productores, autores, y usuarios de los datos de investigación.
Universidades y Centros de Investigación	Establecen la política interna de gestión de los datos científicos.
Repositorios institucionales	Almacenamiento de los datos a corto plazo.
Centros de datos	Selección de los datos que deben preservarse a largo plazo.
Gestores de datos	Gestión y promoción del uso de datos desde su creación. Asegurar su uso, disponibilidad, localización y reutilización.
Usuarios que reutilizan los datos	Deben respetar la atribución y las licencias.
Agencias de financiación	Políticas de datos con los actores implicados

Tabla 2. Actores más relevantes en la gestión de datos según Grupo de trabajo de "depósito y gestión de datos en acceso abierto" del proyecto RECOLECTA (2012)

Además de los agentes o actores involucrados en la gestión de datos de investigación, hay que tener en cuenta que la infraestructura que se necesita para los repositorios de datos no es muy distinta de la de los repositorios de publicaciones. En los repositorios de datos se almacenan y gestionan los *datasets* que son los conjuntos de datos recopilados a lo largo de una investigación. En estos repositorios, los *datasets* se tratan como un recurso con entidad propia en el que encontramos todo tipo de datos y que pueden caracterizarse por diferentes elementos como distintos formatos.

El *dataset* (conjunto de datos) siempre va ligado a una publicación pero no es necesario que ambos, publicación y dataset, estén ubicados en el mismo sitio debido a que a este se le puede tratar como a un objeto digital independiente. El tratamiento de los *datasets* es de vital importancia para una correcta gestión de los datos y su publicación en los repositorios está marcada por dos problemas: la reticencia de los investigadores a publicar sus datos o *datasets* y la escasa formación en gestión en datos en los actores que deben gestionarlos correctamente (Hernández-Pérez y García-Moreno, 2013).

Todo lo tratado en los apartados anteriores hay que tenerlo presente para entender qué circunstancias y elementos influyen a la hora de elaborar un plan de gestión de datos. Centrándonos en el tratamiento de los *datasets* y la necesidad de desarrollar un plan de gestión de datos, Hernández-Pérez y García-Moreno (2013) siguiendo a Michener y Jones (2012) hablan de 8 fases en el ciclo de vida de los datos a los que hay que atender para gestionarlos desde su creación y durante todo su proceso hasta que son reutilizados:

Planificación	Donde se marcan todos los objetivos, métodos, procesos, políticas y recursos necesarios para la correcta gestión de los datos durante todo el ciclo de vida de estos. Fase crucial.
Recolección	Ajustándose a lo planificado y teniendo en cuenta los elementos y tipos que se han establecido.
Control de calidad	Para evitar errores.
Descripción de datos	Destaca la importancia de los metadatos.
Preservación	Plan de preservación a corto y largo plazo conforme a estándares internacionales. Copias de seguridad. Tener en cuenta la obsolescencia del hardware y software.
Descubrimiento	Accesibilidad y visibilidad de los datos de investigación para poder disponer de ellos de forma fácil .
Integración	Capacidad de poder combinarse con otros datos de otras fuentes.
Análisis	Fase final en la que se utilizan los datos para mostrar conclusiones de las investigaciones, refutar hipótesis, etc. Es la razón de ser de los datos de investigación.

Tabla 3. Fases del ciclo de vida de los datos.

Por lo tanto, para que el intercambio de datos mediante el *data sharing* se pueda implementar no sólo es necesario el contexto que ofrecen las e-infraestructuras (toda la infraestructura electrónica que ofrecen las tecnologías de la información) y tampoco basta con que se desarrollen toda una serie de políticas e iniciativas, ni el cambio cultural necesario en la comunidad científica; sino que, aprovechando todo este contexto, los actores o agentes implicados deben desarrollar un buen plan de gestión de datos. Esta cuestión es el pilar fundamental para que el intercambio de datos se pueda desarrollar de una forma eficaz y eficiente.

Observando la literatura científica sobre *data sharing* comprobamos que quienes tienen la responsabilidad de gestionar los datos son los investigadores por ser quienes los crean, pero también los profesionales de la información (bibliotecarios, servicios informáticos, etc.) debido a la escasa formación de los creadores de datos para la gestión de los datos desde un punto de vista de la gestión de información. La creación de los datos, como es obvio, corresponde a los investigadores que son los que generan datos pero la gestión del ciclo de vida de los datos debería recaer en los profesionales de la información por tener las competencias necesarias como gestores de información. El papel que se atribuye a las instituciones es el de proporcionar los recursos necesarios para esta gestión facilitando así la creación de un plan de gestión cuya gestión de los datos abarque todo el proceso de la investigación desde antes de la creación de los datos, cuando se crean y usan y a lo largo de todo su ciclo vital.

El plan de gestión debe incluirse en toda propuesta de financiación y debe tener presente una serie de cuestiones que, partiendo del *dataset* como objeto digital a tratar, se deben tener en cuenta. A continuación recogemos en una tabla las cuestiones imprescindibles a las que debe atender el plan de gestión:

La organización y documentación de los datos
• Mecanismos de almacenamiento y de seguridad para compartir los datos. • Atender a la variedad de tipos de formatos. • Tener presente siempre los estándares internacionales. • Compresión de los datos para que ocupen el menor espacio posible. • Control de versiones. Importancia de la nomenclatura para identificar contenidos. • Uso de registro de metadatos para la documentación aportada por los investigadores. • Priorizar la normalización para conseguir la interoperabilidad con otros sistemas de gestión de datos. • Uso de identificadores digitales únicos en forma de URI asociados a los datasets almacenados.
Marco legal
• Derechos legales sobre los datos y *datasets*. • Propiedad intelectual. • Confidencialidad, privacidad y protección de datos. • Licencias de acceso y uso de los datos.
Plan de preservación de datos
• Copias de seguridad de forma regular. • Estrategia de almacenamiento de datos. • Tener siempre presente la obsolescencia del *hardware* y *software*.

Tabla 4. Cuestiones imprescindibles a las que debe atender el plan de gestión.

Al hilo de los apartados anteriores, la gestión y desarrollo del plan de gestión de datos debe estar en manos de los investigadores y los profesionales de la información. En el caso de los primeros, por ser quienes los crean y en el de los gestores de información por ser quienes tienen competencias útiles que pueden servir en la correcta gestión de los datos. Por esta razón, según González et al. (2013) el papel de los gestores de información, en concreto de los relacionados con la documentación, podría evolucionar en este contexto hacia la nueva figura del *data manager* o *data curator* debido a que muchas de las actividades que han realizado las bibliotecas a lo largo de su historia están vinculadas a las actividades que se deben desarrollar a la hora de gestionar correctamente los datos o los objetos digitales llamados datasets. La actividad del *data curator* se centraría sobre todo en la preservación digital de los datos de investigación como garantía de su reutilización.

> *"Para la preservación y la reutilización de datos no sólo hay que resolver aspectos técnicos, sino también de organización y de procedimiento, económicos, financieros y de personal, administración de su propiedad, obligaciones legales, requisitos de auditoria, restricciones sociales de uso, etc. Este conjunto de acciones es lo que se denomina data curation"* (González et al., 2013)

Aunque no es el objeto de estudio de este trabajo, cabe señalar que en el ámbito del *data curation* también se habla del ciclo de vida de los objetos digitales para tenerlo en cuenta a la hora de su gestión. Destacamos la gestión y tratamiento de los metadatos que son quienes garantizan el acceso y la reutilización, claves para el intercambio de los datos mediante *data sharing*. Un referente en *data curation* es el *Data Curation Center.* Un centro británico que se ha convertido en toda una autoridad en el asesoramiento de expertos y la preservación digital. Dedicado a asegurar la mejora y el uso a largo plazo de los datos digitales, el estudio de métodos, asesoramiento, creación de guías de buenas prácticas, etc. En su web podemos encontrar múltiples guías referentes a planes de gestión de datos, cómo citar los objetos digitales o *datasets*, etc.

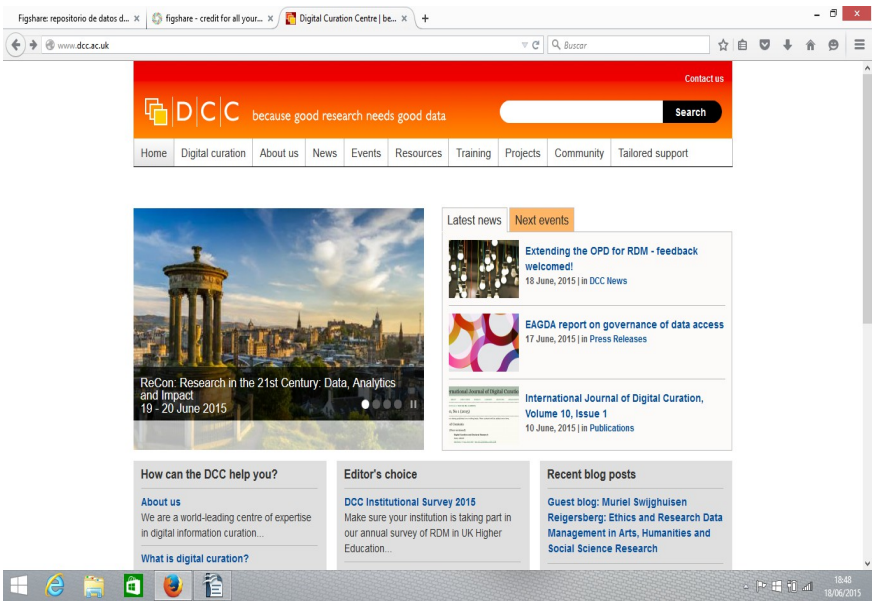

Ilustración 1. DCC (*Digital Curation Centre*)

2.5. INFRAESTRUCTURA TECNOLÓGICA.

2.5.1. CARACTERÍSTICAS GENERALES.

En cuanto a las infraestructuras necesarias para asegurar un intercambio de datos que mantenga la integridad de los datos y ofrezca estabilidad se debe tener en cuenta:

- Preservación de datos a largo plazo incluyendo mecanismos de autenticidad y de control.
- Acceso a los datos (ciclo de vida de los datos), capacidad de computación y servicios de *data curation*.
- Distribución de los datos.

Además es importante que el software sea capaz de gestionar el ciclo de vida de los datos, sistemas de almacenamiento masivo de datos y redes de alta capacidad para la transmisión de estos.

Para el almacenamiento, depósito y difusión de los datos hay una gran variedad de formas y tecnologías: repositorios, bancos de datos (data banks), directorios de bancos de datos, plataformas web, de editoriales, etc. Los repositorios o los bancos de datos son los más utilizados y no difieren mucho de los repositorios dedicados a publicaciones científicas, comparten características similares con una base de datos *online* con niveles de acceso y documentos descritos según estándares pero se diferencian en el hecho de que a diferencia de los documentos, los *datasets* y los datos son más heterogéneos, por lo que existe más variedad de repositorios o *data banks*. Esto sumado a diferencias de depósito y de recuperación hace que el uso de los repositorios de datos este caracterizado por más complejidad que el de los repositorios de publicaciones científicas, por lo que es necesario especializarse en su uso (Torres-Salinas, Robinson-García y Cabezas-Clavijo, 2012).

En la actualidad son muchas las formas, tecnologías y medios para compartir datos pero podemos hacer una primera clasificación general en la forma de compartirlos en función de si se hace por vía formal o informal. Como ya mencionamos anteriormente, la vía informal ya se daba mucho antes de la aparición de los primeros repositorios y se caracterizaba por la petición directa y privada entre investigadores o disponiendo de ella en portales web de investigación. Los problemas de compartir datos de forma informal son que la dificultad de reutilización de esos datos o directamente la imposibilidad de reutilizarlos debido a que son datos que no están normalizados de forma estándar sino que siguen criterios propios de cada autor. Al no ser tratados están más expuestos al riesgo de no poder acceder a ellos debido a la obsolescencia de la tecnología, etc. Por lo que está vía es sólo práctica a un nivel muy concreto y puntual de necesidad circunstancial, pero a medio y largo plazo no es eficaz y eficiente al no asegurar ni la conservación, acceso, disponibilidad y reutilización de los datos, aspectos claves para el *data sharing.*

En cuanto a la a la vía formal, según Borrego (2012) a la hora de crear una clasificación se podría hablar de dos rutas; una verde para *datasets* depositados en repositorios o bancos de datos específicos o la ruta dorada que es la que se depositan en las plataformas que tienen las editoriales junto a la publicación. Otros, según González et al. (2013), proponen otra clasificación que según ellos es más práctica. Estaría basada en servicios nacionales de datos, plataformas por disciplinas (más específicas) y plataformas generales orientadas a usuarios investigadores. Otro intento de clasificación sería distinguir cuatro modelos según la propiedad de los datos depositados: centralizados, descentralizados, federados y ciberestructuras. Esta clasificación fue creada por Kowalczyk y Shankar (2011):

- Modelo centralizado: Se correspondería con los bancos de datos que están controlados por una sola institución como por ejemplo el repositorio de una universidad.

- Modelo descentralizado: Participación de varias instituciones pero con una única ubicación física de los datos.

- Modelo federado: Caracterizado por la participación de varias instituciones pero a diferencia del modelo descentralizado aquí los datos están físicamente distribuidos entre los participantes.

- Ciberestructura: Datos distribuidos en redes de ordenadores. Accesibles en todo momento y ubicados en distintos lugares.

Con todo lo anterior presente aún así habría muchos tipos más ya que se podría hacer clasificaciones atendiendo al tipo de formato de los datos, disciplinas o ámbitos de investigación, etc.

2.5.2. REPOSITORIOS DE DATOS DE INVESTIGACIÓN.

El número de repositorios de datos es elevado y de muy variado tipo. Nos encontramos con repositorios: institucionales, generales, por disciplinas, nacionales, internacionales, consorcios, instrumentales, editores... Y a la hora de crear una infraestructura de intercambio de datos como los repositorios, se puede decidir entre crear un repositorio independiente solo para datos, de los cuales hay miles de ejemplos (como por ejemplo *Dryad*) o implementar uno para datos en algún repositorio ya existente de publicaciones, como es el caso de la Universitat de Barcelona, en cuyo repositorio digital se ha incorporado un apartado para los datos de investigación. También nos encontramos con herramientas que surgen influenciadas por el contexto actual de *open science* dónde los investigadores comparten sus investigaciones y publican sus datos de investigación. Destacamos el repositorio de datos Figshare.

Los repositorios o bancos de datos son ya tan numerosos y variados que sólo vamos a presentar una pequeña muestra de algunos ejemplos. Estos se han convertido en referentes de sus respectivos campos o disciplinas en el ámbito internacional y son parte de la selección que hizo la revista PLOS (*Public Library Of Science*) de los repositorios de datos más confiables:

Nombre	**Disciplina**	**Tipo de datos compartidos**	**Estadísticas**
Worldwide protein data bank	Proteómica	Estructuras de macromoléculas, gratuito y accesible a todo el mundo.	110.907 depósitos entre los años 2000-2015 en tres bases de datos: 84.009 en RCSB PDB, 20.061 en PDBj y 16837 en PDBe
GenBank	Genómica	Secuencias genéticas. Producido por los NIH, es de acceso público.	193.921 millones de bases en 185 millones de secuencias en GenBank y más de un billón de bases y 258 millones de secuencias en la división WGS.
Dryad	Biociencias	Datos referenciados en artículos científicos de biociencias.	8.674 paquetes de datos y 27.311 ficheros de datos
Pangaea	Geociencias	Datos georeferenciados sobre investigaciones geológicas, en acceso abierto.	222 proyectos de investigación; entre ellos 65 proyectos europeos
Simbad astronomical database	Astronomía	Información básica, identificaciones cruzadas, bibliografía y medidas de objetos astronómicos de fuera del sistema solar.	7.983.555 objetos, 22,264,673 identificadores, 305,087 referencias bibliográficas y 11,875,721 citas de objetos en publicaciones

Tabla 5. Principales repositorios de datos internacionales. (Datos estadísticos consultados el 21 junio 2015.)

En el caso español observamos que apenas existen repositorios de datos de investigación; ya sean independientes o integrados como parte de un repositorio digital ya existente. Realizando una búsqueda por países en el directorio de repositorios de datos de investigación Re3data obtenemos 13 resultados para España. De estos 13 resultados sólo 6 son íntegramente nacionales; los 7 restantes son repositorios dónde se colabora con otros países. Habría que sumar el caso del repositorio *Dipósit Digital de la UB* que dispone de un repositorio de datos integrado pero no aparece en Re3data.

Repositorios	Tipo	Disciplina	Situación	Estadísticas
Digital.CSIC.	Institutional	Multidisciplinar	Integrado en un repositorio digital ya existente	Desconocido
UPF Digital Repository - Recursos i dades primàries	Institutional	Multidisciplinar	Integrado en un repositorio digital ya existente	163 datasets
CEACS Data Library	Institucional	Temático. Ciencias sociales	independiente	2000 datasets
Banco de Datos Específico de Estudios Sociales - ARCES del CIS. CIS Data Bank	Institucional	Temático. Ciencias sociales	independiente	Desconocido
Herschel Science Archive	Disciplinar	Temático. Astrofísica y astronomía	Independiente	Desconocido
AMIGA	Disciplinar	Temático. Astrofísica y astronomía	Independiente	Desconocido
Dipósit Digital de la UB	Institucional	Temático. Humanidades y Ciencias sociales.	Integrado en un repositorio digital ya existente	134 datasets

Tabla 6. Principales repositorios de datos españoles. (Datos consultados el 22 junio de 2015)

En cuanto a software utilizado Kowalczyk y Shankar (2011) destacan 3 sistemas de depósito digital de código abierto que se utilizan o pueden utilizarse para aplicar las colecciones de datos: DSpace, Fedora y iRODS.

2.5.3. DIRECTORIOS DE REPOSITORIOS DE DATOS.

Respecto a los directorios de conjunto de datos han surgido iniciativas similares a los directorios de repositorios de acceso abierto como son ROAR (*Registry of Open Acces Repositories*) y OpenDoar (*Directory of Open Acces Repositories*) pero ofreciendo como resultado registros de repositorios de datos o conjuntos de datos. Su objetivo es realizar un intento por visibilizar y facilitar la búsqueda de repositorios de datos y por consiguiente facilitar el encontrar los *datasets* pertinentes. Entre estos directorios destacan a nivel internacional *Re3data y dataHub.* A nivel nacional existe un proyecto español reciente llamado *ODISEA*:

- *Re3data*: De origen alemán aunque de carácter internacional, es una herramienta que permite la búsqueda de más de 1200 repositorios de datos de investigación. Está en proceso de fusión con *Datacite* pasando a ser una aplicación de búsqueda de repositorios de esta. También ha integrado la herramienta *Databib* que identifica y localiza repositorios en línea de datos de investigación. Para ello cuenta con un catálogo de búsqueda y se fundamenta en los registros de repositorios que son descritos por los usuarios. La fusión con *Datacite* y *Databib* convierten a *Re3data* en una herramienta muy potente. Realiza las búsquedas por: temas, contenido y países.

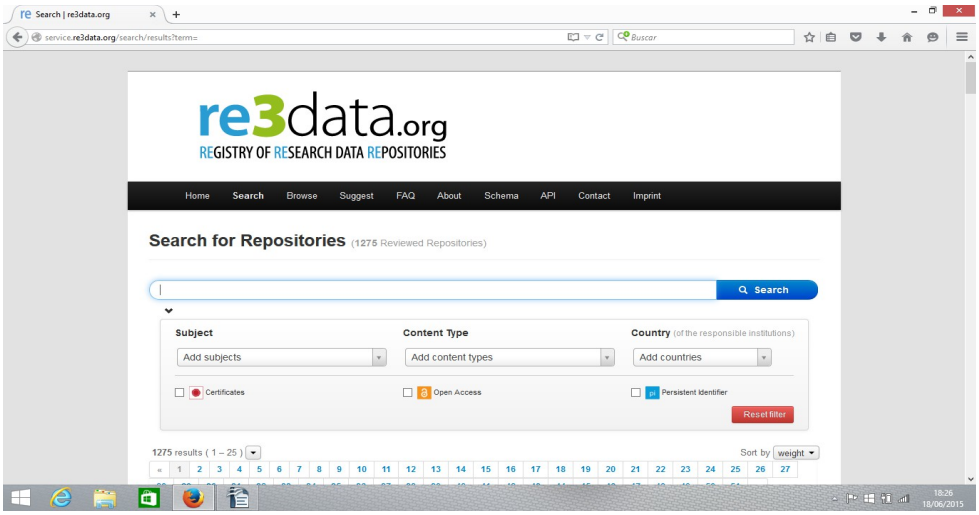

Ilustración 2. Portal del directorio *Re3data.*

- *DataHub*: Una plataforma que gestiona y facilita la búsqueda de repositorios, registros y conjuntos de datos de forma abierta. Es de carácter internacional, general y multidisciplinar. Provee datos de todo tipo y de distintas disciplinas, instituciones y niveles de gobierno.

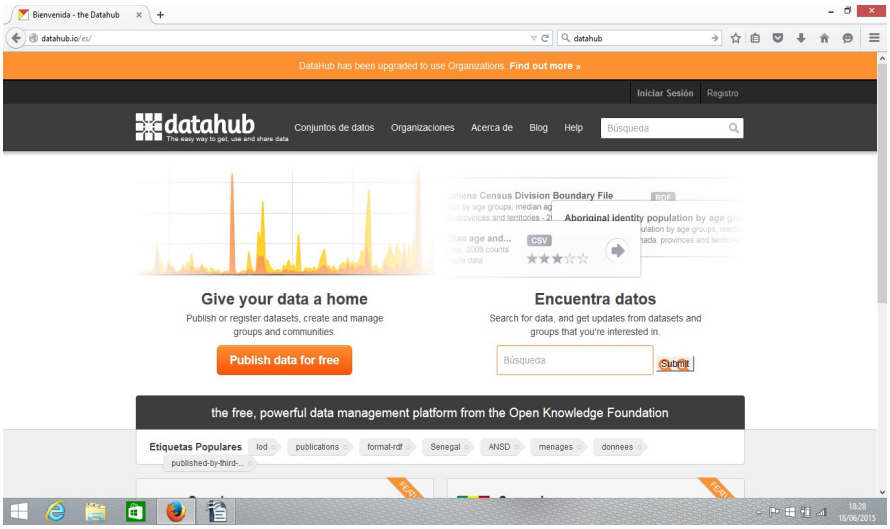

Ilustración 3. Portal del directorio *Datahub*.

- *Odisea: International Registry on Research Data*, Es un proyecto realizado por 5 universidades españolas (Universitat de Barcelona, Universidad Politécnica de Valencia, Universidad de Valencia, Universidad Católica de Valencia y Universidad de Murcia.) con el objetivo de crear un directorio global que recoja de forma centralizada todos los repositorios de datos de investigación e intentar clasificarlos por disciplinas para facilitar la búsqueda de estos y sus *datasets*. Funciona a modo de inventario internacional de depósitos de datos de investigación especializados en la preservación digital de datos. Utiliza *Drupal* para el web y la base de datos y dispone de un sistema de búsqueda en los registros aplicando distintos tipos de filtros. Actualmente cuenta con 183 depósitos y utiliza para la clasificación las áreas de conocimiento del *Essential Science Indicators de la Web of Knowledge.*

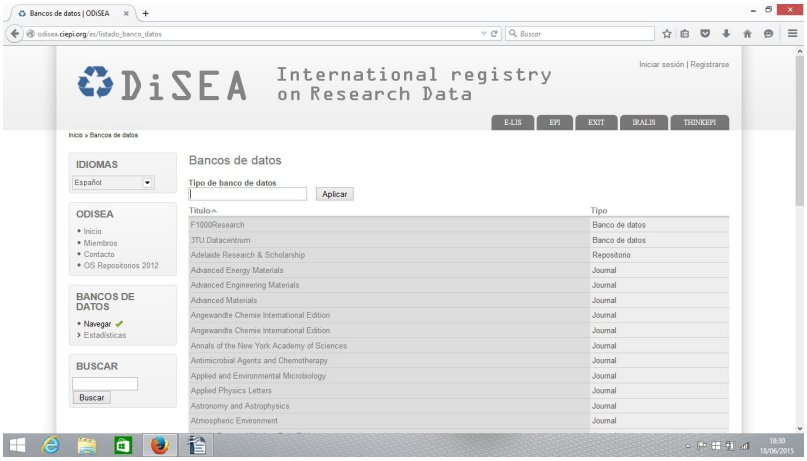

Ilustración 4. Portal del directorio ODISEA.

ODISEA es un proyecto que surgió debido a que, a pesar de haber surgido muchos repositorios de datos en los últimos años, estos no realizan la gestión, preservación y difusión de los datos o conjuntos de datos de una forma sistemática con acciones encaminadas a normalizar unos estándares. Esto provoca que la información no sea homogénea ni siga patrones estandarizados lo que sumado a que los *datasets* y el manejo de las plataformas no son más que comprensibles por los usuarios especializados en cada herramienta o disciplina a la que pertenezcan se dificulta la búsqueda de los datos de investigación y por tanto la práctica eficaz y eficiente del *data sharing*.

2.6. POLÍTICAS Y CULTURA DEL DATA SHARING.

2.6.1. PRINCIPALES POLÍTICAS.

Kowalczyk y Shankar (2011) utilizan el término "infraestructura humana" para referirse a todos los temas sociales y culturales relacionadas con el *data sharing* y las políticas, iniciativas y prácticas que lo desarrollan. Los problemas sociales o culturales a los que se enfrenta el *data sharing* tienen que ver en gran parte a la reticencia de publicar los datos de investigación por parte de los investigadores y los autores de los datos. Pero también se dan otro tipo de problemas, como puede ser la gestión correcta de los datos complejos. Para gestionar algunos tipo de datos es necesaria una mínima formación debido a su complejidad; en muchos casos son incomprensibles para los que no son especialistas en la disciplina a la que pertenecen.

Además se dan problemas que tienen que ver con la protección a la intimidad; como podría ser los casos en los que investigaciones que tienen que ver con el ADN de las cuales se pueden extraer datos que revelen la identidad de los pacientes con enfermedades genéticas. También referentes a la protección de la autoría de los datos, de reconocimiento, etc. Para superar la reticencia de los investigadores a publicar sus datos de investigación es necesario despejar el *data sharing* de toda la incertidumbre que generan estos problemas. En este camino va orientado el desarrollo de directrices y políticas que abarcan toda la problemática del *data sharing,* con el objetivo de crear un entorno que ofrezca garantías tanto a investigadores como a usuarios.

Opinan Kowalczyk y Shankar (2011) que hay tres dominios donde se generan, se ejecutan y se hace cumplir las políticas, directrices y normas para el ejercicio del *data sharing*. Estos son de tipo gubernamental, institucional y cultural. Las distintas entidades correspondientes a cada uno de esos dominios son las que establecen dónde generar datos, albergarlos, donde publicarlos y bajo que condiciones, como manejarlos, diferentes tipos de imposiciones o restricciones en su uso, etc. En algunos casos simplemente se tratan de vagas directrices o guías de buenas prácticas que orientan y aconsejan como gestionar mejor el intercambio de datos y su uso pero cada vez más se implementan normativas, leyes y reglamentos que obligan a la publicación de los datos de investigación e imponen y establecen como se ha de hacer y en que condiciones.

En su mayoría son de tipo gubernamental; como exigir planes de gestión de datos de investigación a los investigadores que quieran recibir dinero público para sus investigaciones. Un ejemplo es el caso de EE.UU donde se exige que para solicitar una subvención de más de medio millón de dólares es obligatorio presentar un plan de gestión de datos. Pero también en el ámbito institucional como en universidades y otras instituciones de investigación que, cada vez más, siguen este camino debido a las presiones. También ocurre en el ámbito cultural donde son ya numerosas las revistas que establecen criterios en cuanto a la publicación de datos de investigación, aunque varían mucho de unas a otras.

En cuanto a los organismos de financiación la principal razón es la que ya hemos comentado en el párrafo anterior de la exigencia de publicar los datos de las investigaciones que hayan sido financiadas con fondos públicos. Los documentos que establecen políticas, o simplemente directrices, empiezan a ser abundantes a partir del año 2000, sobre todo en EEUU y Reino Unido, destacando la relevancia del organismo *National Science Foundation* que publicó el documento *NSF data sharing policy and data management plan requirements 2001*.

Otro documento a nivel nacional, en este caso español, es el *Protocolo de remisión, almacenamiento y difusión de datos antárticos en España, 2004-2007* del Centro Nacional de Datos Polares. En él se expone todo lo que los investigadores deben realizar para la gestión de los datos indicando el tipo de datos, sus formatos, procedimientos y descripción de los mismos con todo lo referente al manejo de metadatos, prioridades de situaciones legales, periodo de carencia, la accesibilidad y visibilidad de los datos. En cuanto al envío de los datos, se detalla que si no son proporcionados el investigador principal será excluido en futuros proyectos.

La OCDE (Organización para la Cooperación y Desarrollo Económicos) es la mayor impulsora del *data sharing* a nivel internacional, sobre todo desde que publicó los *OCDE Principles and guidelines for acces to research data from public funding* (2007) donde expresa una serie de directrices y recomendaciones para los gobiernos. En cuanto a la Unión Europea actualmente la situación se encuentra supeditada al programa marco 2014-2020 llamado *Horizon 2020* que incluye el requisito de que los resultados y datos de investigaciones de diversas áreas y disciplinas del conocimiento deben ser publicadas en repositorios en acceso abierto.

A pesar de ser numerosos los documentos que publican políticas, directrices o recomendaciones de buenas prácticas no es aún algo generalizado por la mayoría de las organizaciones que financian las investigaciones por lo que se hace palpable el lento desarrollo de la normalización del *data sharing.* También la necesidad de más contundencia por parte de los actores implicados en el desarrollo de la ciencia y más sensibilidad sobre la importancia de compartir los datos por parte de los actores encargados de generar esos datos y publicarlos. Según Melero y Hernández-San-Miguel (2014) “tan solo un 26% de las agencias que financian la investigación requieren el depósito de los *datasets* en repositorios de acceso abierto, un 10% lo recomiendan y el resto no hacen alusión al depóstio de datos”

Lo comentado en párrafos anteriores es sólo una pequeña muestra representativa de lo más relevante en cuanto a las principales políticas sobre *data sharing.* Y a pesar de que la mayoría de organizaciones no muestran el requerimiento de la publicación de los datos de investigación esta tendencia va desarrollándose y aumentando; por lo que el volumen de documentos sobre políticas sigue creciendo.

2.6.2 LAS REVISTAS CIENTÍFICAS.

Las revistas científicas también contribuyen a fomentar la cultura del *data sharing* a través de las instrucciones que ofrecen para que los autores publiquen los datos y los compartan indicando donde tienen que depositarlos. En algunos casos supone una obligación para los autores el compartir los datos si quieren que sus trabajos sean publicados. Es el caso de la revista *Plos one,* que rechaza los trabajos de aquellos autores que no los comparten bajo los estándares que indica la revista.

En las normas para autores de las revistas es donde se pueden encontrar las políticas de la revista sobre *data sharing.* Uno de los principales requisitos que piden es que se debe proporcionar el *accesion number* o número de registro, también llamado DOI (*Digital Object Identifier*). Es utilizado para identificar los ficheros de datos de forma unívoca para poder buscarlos y recuperarlos. También se pueden encontrar recomendaciones para los autores como que tengan una actitud proclive a la publicación de los datos o que denuncien en caso de que un autor no los publique de forma accesible para poder compartirlos. Son las revistas científicas sin interés comercial las que más firmemente se han posicionado a este respecto.

Además se comienza a observar que las revistas con políticas definidas sobre *data sharing* tienen un mayor factor de impacto (Piwowar, 2008), pero aún queda mucho recorrido para estudiar si este fenómeno puede considerarse sistemáticamente cierto.

La citación de datos sigue la idea de que los *datasets* deben publicarse al igual que otros tipos de productos académicos, siendo considerados también como productos de investigación, tanto desde la perspectiva política, social y de financiación. Melero y Hernández-San-Miguel (2014) exponen esta idea afirmando que los *datasets* pueden ser citados como cualquier otro objeto digital y por lo tanto entrar en el ciclo de las citas de forma individual o por el artículo al que están vinculados. Exponen además una serie de ejemplos de cómo publicar los datos de investigación en acceso abierto hace que aumente la citación, como por ejemplo la revista *Paleoceanography,* que tuvo un 35% más de citas en los artículos que contenían enlaces a sus *datasets* albergados en *Pangea*. También cuenta el caso de la revista *Astrophysical Journal* cuyos artículos vinculados a *datasets* depositados en el *Astrophysical Data System* tuvieron un 50% más de citación que los que no tenían publicados los datos en acceso abierto.

Aun con todo lo mencionado en párrafos anteriores, las políticas sobre *data sharing* de las revistas no han conseguido evitar la baja participación de los investigadores en el *data sharing*. Y es que todavía no hay un consenso sobre la métrica necesaria. Esta cuestión es importante si tenemos en cuenta que una de las formas de romper con la reticencia de los investigadores a publicar sus datos es garantizarles, entre otras cosas, algún tipo de compensación o reconocimiento. Porque aunque el *data sharing* ofrece beneficios importantes para el progreso científico y el avance del conocimiento su adopción de forma generalizada se ve limitada en gran parte por no haber un desarrollo de la métrica de datos, *data metrics*. El desarrollo de un métrica de los datos podría solucionar la cuestión del reconocimiento de los autores lo que motivaría más a la tendencia de publicar e intercambiar los datos.

Según Costas et al. (2013) la publicación de los datos y la citación no es un elemento considerado para la promoción y evaluación de la investigación y la cita de los datos no es un comportamiento habitual en los trabajos académicos, lo cual afecta el desarrollo de un sistema de recompensas ya que si los investigadores no publican y citan los *datasets* de manera sistemática y estandarizada, el desarrollo de métricas de datos será difícil y probablemente no fiable.

Una de las ventajas de publicar los datos, que ya hemos mencionado anteriormente, es la de evitar errores o fraude en las investigaciones al estar accesibles los datos de la investigación. Esto no quiere decir que los investigadores los cometan de forma intencionada pero en algunos casos se ha observado que en los trabajos con mayor número de errores correspondían a autores que se negaban publicar los datos como el caso de unas revistas de Psicología donde se llevó a cabo un análisis para detectar errores estadísticos y coincidió que los trabajos con mayor número de errores eran los de autores que no quisieron publicar datos de investigación. Destaca *"el caso de fraude de Holanda en la psicología donde Diederik Stapel, un científico muy respetado en la Universidad de Tilburg, fabricaba conjuntos de datos fraudulentos"* (Costas et al., 2013)

Vemos que la publicación de datos de investigación de forma abierta en las revistas no sólo fomenta el desarrollo de la ciencia gracias a la reutilización de los datos que pueden servir para generar nuevas investigaciones y nuevos datos, sino que la fortalece gracias a la transparencia que otorga este sistema que permite detectar antes el fraude y los errores. Además sirve para el aumento del factor de impacto tanto de revistas como de autores de los trabajos gracias a que los datos también se pueden citar. Por tanto, las políticas de *data sharing* de las revistas están teniendo un impacto importante y los editores deben seguir requiriendo los datos de los trabajos y seguir desarrollando e implantando las directrices para poder disponer de esos datos fortaleciendo así el *data sharing*. Tal es la relevancia que los datos han generado para las revistas que ya hay proyectos de revistas dedicados exclusivamente a los datos, como la revista *Journal of Open Public Health Data* o la revista *Scientific Data* del grupo *Nature*.

Ilustración 5. Portal de la revista *Journal of Open Public Health Data*.

2.6.3. CULTURA DEL DATA SHARING: EL LENTO CAMINO HACIA LA NORMALIZACIÓN.

A pesar de toda la promoción que se ha hecho del *data sharing* y de sus virtudes durante los últimos años, el no proporcionar los datos de investigación por parte de los autores sigue siendo la pauta más generalizada y esto tiene que ver con sus percepciones y con las cuestiones culturales relacionadas con estas actividades. Tenopir et al. (2011) realizaron una encuesta para explorar las percepciones de los investigadores con respecto a intercambio de datos y la publicación de datos. Encontraron que las barreras importantes percibidas por los investigadores son la falta de tiempo y la falta de financiación.

El informe de Costas et al. (2013) también señala otros aspectos como:

- La "pérdida de control" sobre los datos por parte de los creadores de los datos sumado a un posible uso indebido.
- La posible disminución de la calidad de la ciencia.
- Problemas con la propiedad de los datos y el permiso para la publicación de los datos.
- Una fuerte percepción cultural entre los científicos de que los datos son suyos, que no es totalmente defendible desde el punto de vista de la responsabilidad de la financiación pública.
- Temor a que los posibles errores los pueden exponer.
- Una resistencia por parte de los investigadores a compartir sus datos con otras disciplinas.

Por tanto, el desarrollo del *data sharing* como práctica generaliza es lento a pesar de los grandes avances de los últimos años y el desarrollo tanto de todas las políticas y directrices como del cambio cultural propio de la e-ciencia. Las reticencias de los investigadores, en muchos casos no exentas de cierta justificación, impiden y obstaculizan el intercambio de datos, generando un círculo vicioso debido a que si no aumenta esta práctica no se desarrollarán, entre otras cosas, una buena métrica de datos que solucionaría parte de los problemas que apuntan los investigadores. Las políticas y directrices tienen que ir también enfocadas a tratar la creación de estándares de citación de datos y desarrollo de una métrica de datos, (d*ata metrics*).

A continuación se muestra una tabla con los actores implicados y las acciones recomendadas para promover una cultura del *data sharing:*

Investigadores/productores de datos. (Científicos)	Aplicar la citación de datos como forma de reconocimiento. Ver el *data sharing* como una buena práctica científica.
Universidades y Centros de Investigación	Promoción de políticas y estándares de intercambio de datos.
Editoriales revistas científicas	Implementar como requisito para la publicación el aportar los datos de forma abierta para fomentar el intercambio.
Centros de datos	Desarrollo de la normalización y estándares para el intercambio, tratamiento y métricas de datos.
Gestores de datos	Promoción, información y formación de todas las ventajas del data sharing. Desarrollo de métricas de datos.
Agencias de financiación	Requisitos para recibir financiación. Recompensar el intercambio de datos. Tener en cuenta la métrica de datos a la hora de evaluar planes de gestión de datos.
Usuario de los datos	Citar los datos o *datasets.* Respetar la autoría y las licencias.

Tabla 8. Actores implicados y las acciones recomendadas para promover una cultura del *data sharing*

3. APLICACIÓN DE UN MODELO DE EVALUACIÓN PARA REPOSITORIOS DE DATOS.

Ante la necesidad de afianzar y extender la cultura de *data sharing* surge el problema de la reticencia de los investigadores a la hora de publicar sus datos de investigación. Para superar esta incertidumbre se debe ofrecer una serie de garantías a los investigadores. Una de las formas de generar el clima de confianza necesario para que los investigadores sean más proclives a publicar sus datos es que los repositorios de datos provean información clara sobre las medidas que toman para ofrecer garantías. Se considera que esta medida es clave para la promoción de la cultura del *data sharing*

El objetivo de este estudio es saber si los principales repositorios de datos de investigación ayudan a evitar esta situación. Para ello se pretende averiguar si muestran información clara y concisa para evitar la incertidumbre entre los investigadores que se decidan a publicar los datos de sus investigaciones.

3.1.SELECCIÓN DE LOS REPOSITORIOS .

Para ver como responden los principales repositorios de datos ante esta cuestión, se han seleccionado 7 repositorios de datos para realizar una evaluación, comparación y análisis siguiendo una serie de criterios que se muestran más adelante en un cuestionario. Estos criterios están relacionados con los aspectos relevantes de la información que podría ayudar a evitar la reticencia de los investigadores a publicar sus datos. Las respuestas se recogen en una tabla para facilitar la comparación y su posterior análisis.

El análisis de los repositorios se llevó a cabo en la misma web de cada repositorio. También se comprobó la información recurriendo al directorio de repositorios de datos *re3data*, para ello se realizaba una búsqueda en *re3data* de cada repositorio y se comprobaba la información que mostraba referente al repositorio buscado. Hay que apuntar que no toda la información aparecía, ni el la web de los repositorios ni en la información que aportaba el directorio *re3data*. (Además alguna de la información que ofrecía *re3data* estaba obsoleta).

A continuación se ofrece una breve descripción de cada uno de ellos.

Dryad

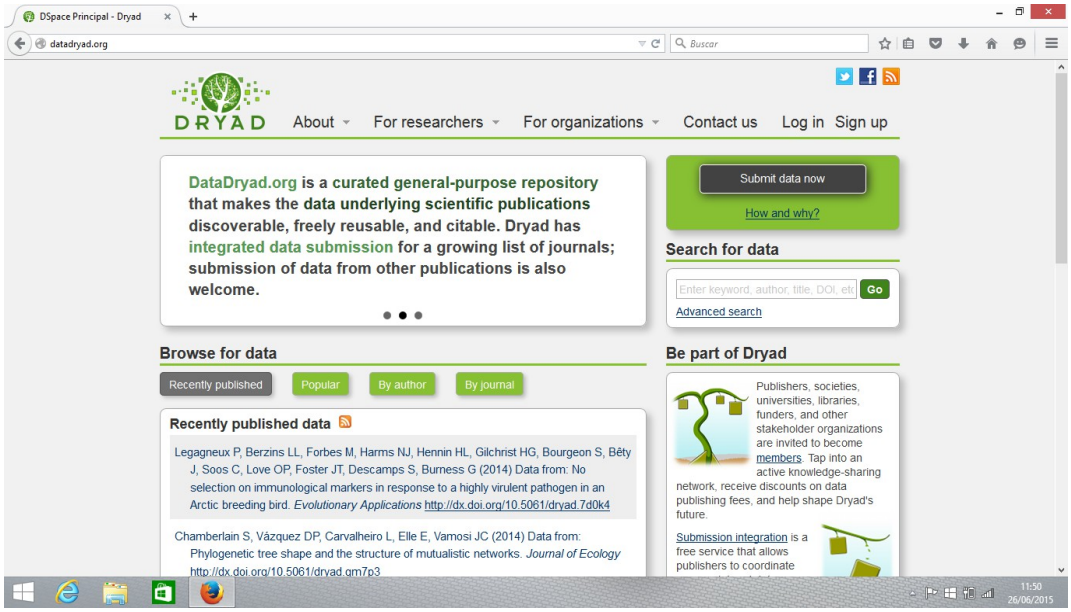

Ilustración 6. Portal del repositorio de datos de investigación *Dryad*.

Dryad es un repositorio digital de Biociencias en acceso abierto. Se encarga de proveer la infraestructura necesaria para publicar y albergar datos de investigación de la literatura científica y médica internacional. No persigue el beneficio económico sino ofrecer la posibilidad de publicar los datos de investigación para que estos puedan ser reutilizados y generen conocimiento. Tiene su origen en la iniciativa promovida por un grupo de revistas científicas del campo de la biología científica y la ecología para albergar conjuntamente los datos de las investigaciones. Alberga datos de muy diverso tipo y no hay limitación para ser miembro aunque entre los miembros predominan revistas, editoriales, instituciones de investigación, sociedades científicas, bibliotecas y organizaciones de financiación. Su objetivo es hacer el archivado de datos lo más simple y fácil posible a través de un conjunto de servicios que no ofrecen las revistas o instituciones. Participa en organizaciones como BioSharing, DataCite y Dataone.

Aunque no tiene fines de lucro necesita financiación para mantener sus funciones básicas de infraestructura, tratamiento y preservación de los datos. Para poder proporcionar el acceso libre a los datos necesita el apoyo financiero de los miembros mediante los Cargos de publicación de datos de Dryad (DPC) disponibles en su web. Cabe destacar que entre sus servicios también incluye ayuda para el desarrollo de un plan de gestión de datos para quién lo solicite, incluyendo la solicitud de fondos y la estrategia de gestión.

Genbank

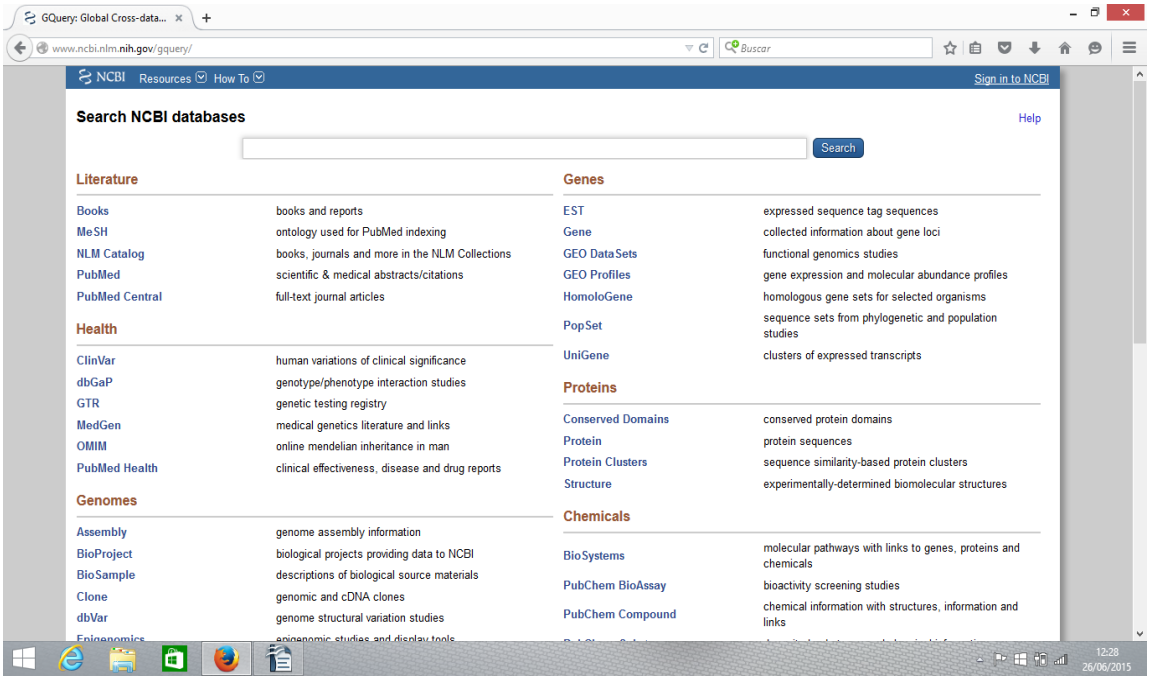

Ilustración 7. Portal del repositorio de datos de investigación *Genbank*.

Genbank es la base de datos de secuencia genética de los Institutos Nacionales de Salud de Estados Unidos (NIH). Los NIH están formados por un grupo de instituciones del gobierno de los Estados Unidos. El objetivo de Genbank se centra en la investigación médica y contiene una colección anotada de todas las secuencias de ADN disponibles públicamente. Además forma parte de la base de datos de secuencias de nucleótidos Internacional la cual es fruto de la colaboración de *Genbank* (NCBI), el banco de datos de ADN de Japón (DDBJ) y el Laboratorio Europeo de Biología Molecular (EMBL). Los tres intercambian datos en una base diaria.

GenBank proporciona y fomenta el acceso a la información más actualizada y completa de la secuencia de ADN. El uso o distribución de los datos es libre; aunque respeta los casos en que algunos autores disponen de patentes. Así como los casos en los que prima la autoría o propiedad intelectual sobre los datos por parte de sus autores. También aconseja sobre como obrar cuando se va a depositar datos que contengan información de una fuente a la que hay que proteger su intimidad y privacidad. Esto es debido a que al ser secuencias de ADN e información médica puede exponer a personas sobre las que se han llevado a cabo las investigaciones.

Utilizar *Genbank* resulta complejo si no se tienen conocimientos específicos de la disciplina por lo que su uso está muy limitado a la comunidad científica. Se puede buscar en su base de datos NCBI según las siguientes categoría: literatura, salud, genomas, genes, proteínas y químicos. Cada categoría contiene varias bases de datos especializadas en un tema concreto. También ofrece la posibilidad de buscar en la base de datos internacional *Nucleitode* en la que colabora.

PANGAEA

Ilustración 8. Portal del repositorio de datos de investigación *PANGAEA*.

PANGAEA es un archivo digital de geociencias surgido por iniciativa de la revista *Earth System Science Data* (ESSD) y un grupo de revistas de investigación del sistema terrestre. Su financiación se mantiene gracias a fondos aportados por la Comisión Europea, Investigación, Ministerio Federal de Educación e Investigación alemán (BMBF), Fundación Alemana de Investigación (DFG) y el Programa Internacional de Descubrimiento del Océano (IODP).

Es un archivo digital de acceso abierto destinado a albergar, publicar y distribuir datos georreferenciados y geológicos de la investigación del sistema terrestre. Garantiza la disponibilidad a largo plazo de los datos. Estos son de libre acceso y se pueden utilizar atendiendo a las licencias a los que estén sujetos.
Sólo están restringidos los *datasets* de proyectos en curso.

PANGAEA ofrece una interfaz de búsqueda de conjuntos de datos sencilla que ofrece la posibilidad de restringir la búsqueda por 4 opciones según el medio: agua, sedimento, hielo y atmósfera. La búsqueda avanzada hace posible la búsqueda de los *datasets* que se encuentran en una cobertura geográfica dada; aportando latitud/longitud en una ventana gráfica y una cobertura temporal: rango de fecha/hora. El motor de búsqueda utiliza el software de código abierto *Elasticsearch*.

La gestión de datos y el depósito se hace según las indicaciones de los Principios y Directrices de la OCDE para el acceso a los datos de investigación de financiación pública. Se adscribe a la Declaración de Berlín sobre Acceso Abierto al Conocimiento en Ciencias y Humanidades.

SIMBAD

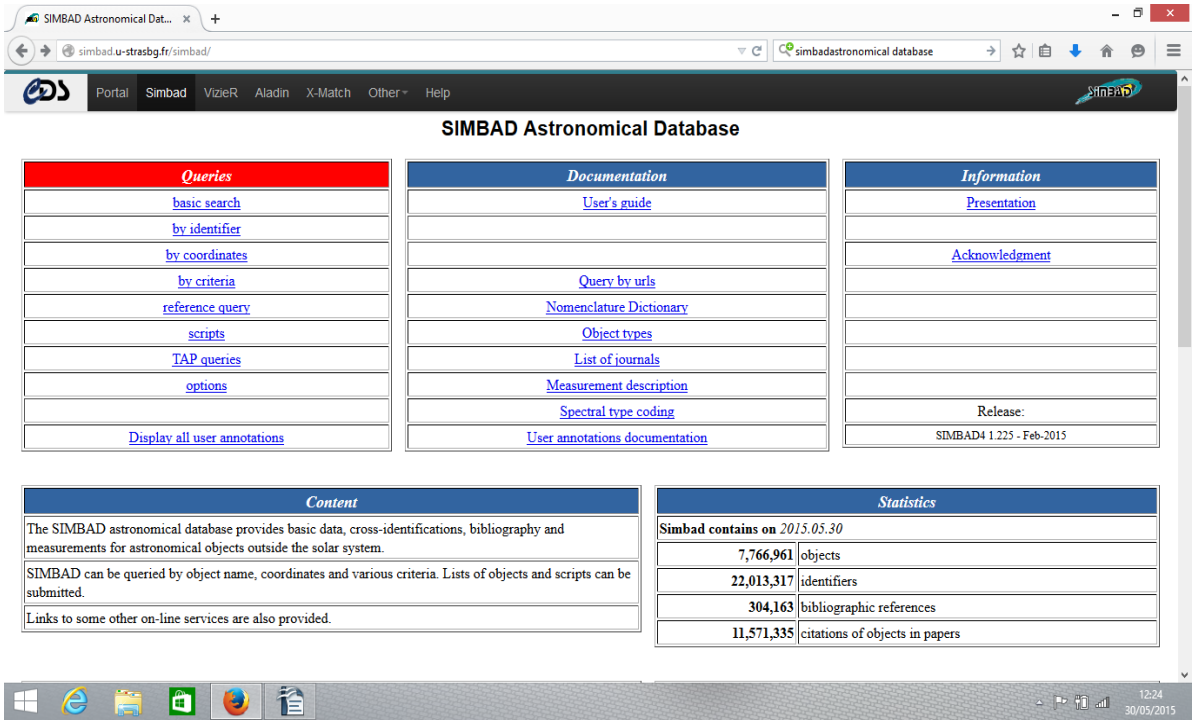

Ilustración 9. Portal del repositorio de datos de investigación *SIMBAD*.

SIMBAD Astronomical Database (*SIMBAD*, Conjunto de Identificaciones, Mediciones y Bibliografía de datos astronómicos) es una base de datos astronómicos francesa gestionada por el Centro de Datos Astronómicos de Estrasburgo. Alberga datos básicos, identificaciones cruzadas y mediciones de los objetos astronómicos que están fuera del sistema solar. Los únicos datos que no incluye son los datos astronómicos del sistema solar. Su actualización es continua gracias a la colaboración del Instituto de Astrofísica de París, del Observatorio de París y del Observatorio de Toulouse.

Su utilización es compleja incluso para los especialistas de la discipina. Para un buen uso del repositorio se provee de guías y mucha información con acrónimos y demás referencias especializadas. Se explica cómo acceder a la base de datos, informaciones generales y todo lo que se necesita saber para realizar consultas de una forma eficaz. También hay disponibles tablas para entender la descripción de los datos y de apoyo. Por ejemplo para los datos de observación o medidas hay unos 30 tipos de datos por lo que se necesita de una tabla para saber como es representado cada uno.

Podemos encontrar datos básicos sobre estrellas y galaxias. Los tipos de datos que ofrecen son entre otros: tipos de objetos, coordenadas, movimiento adecuado, paralajes, tipo espectral, magnitudes y fundentes para varias longitudes de onda y notas, velocidades radiales, tipo morfológico y dimensión integrados... Las búsquedas de los datos se pueden realizar por identificadores (nombres de objetos astronómicos), coordenadas (recuperación de todos los objetos en una dirección dada) y por conjunto de criterios físicos de muestreo.

Zenodo

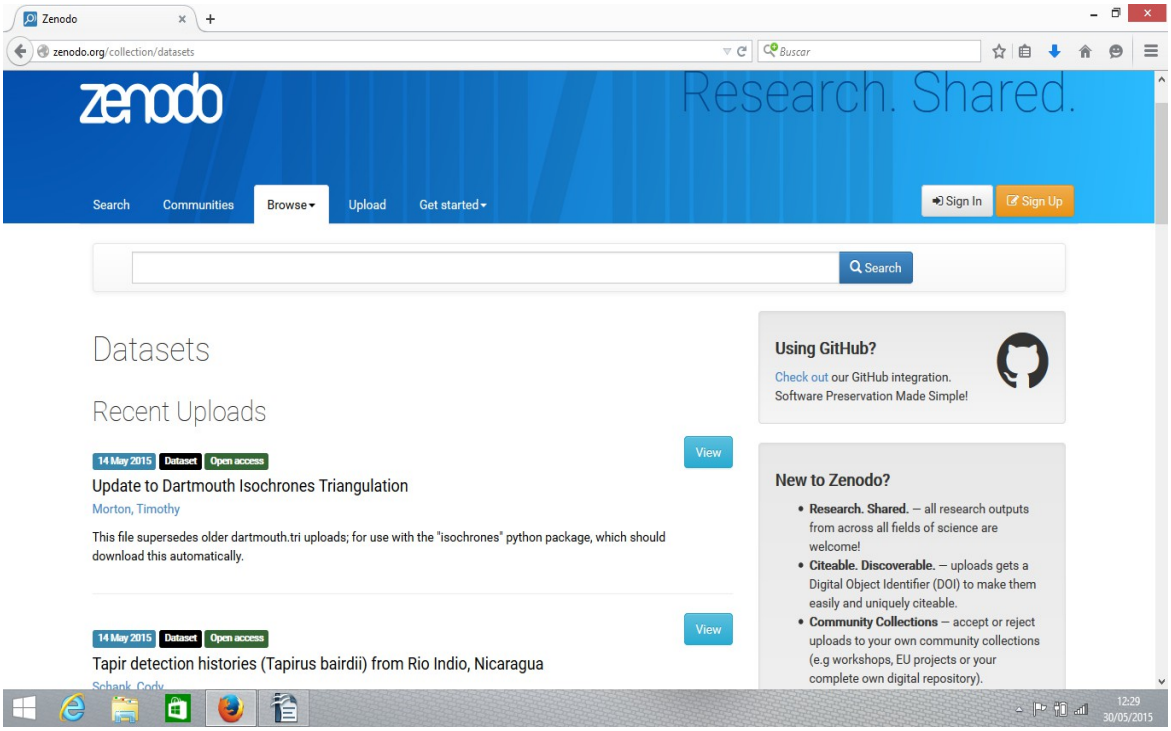

Ilustración 10. Portal del repositorio de datos de investigación *Zenodo*.

Zenodo es un repositorio de datos de investigación de carácter multidisciplinar creado por *Openaire* y el CERN. Debe su nombre a Zenodotus, bibliotecario de la antigua biblioteca de Alejandría. Se inició a raíz de un proyecto europeo (*OpenAireplus*) con financiación europea por lo que los primeros datos que se publicaron fueron provenientes de investigaciones financiadas con fondos europeos. A pesar de esto Zenodo esta abierto a cualquier tipo de datos provenientes de proyectos que no reciban financiación de fondos europeos. Es de reciente creación por lo que es pronto para aventurar la relevancia o impacto que pueda tener realmente, pero su objetivo es convertirse en un referente europeo. De ser así, a través de él, se podría acceder a la mayoría de datos de investigación de Europa y convertirse en una herramienta que ofrecería mucha visibilidad a los investigadores.

Es de acceso abierto y su objetivo es permitir a los investigadores el poder disponer de un espacio donde compartir y preservar los resultados de la investigación de cualquier disciplina, en cualquier tamaño y cualquier formato.

Para acceder a los datos dispone de un buscador simple con la posibilidad de realizar filtros por: *datasets*, imágenes, posters, presentaciones, video/audio, software y publicaciones. Acepta que los depositantes suban archivos de hasta 2 GB. En algunos casos permite tener más de un archivo de 2GB o si superan dicha capacidad permiten ponerse en contacto con ellos y tratar el asunto de forma favorable. Todos los datos se almacenan en el centro de datos del CERN.

Figshare

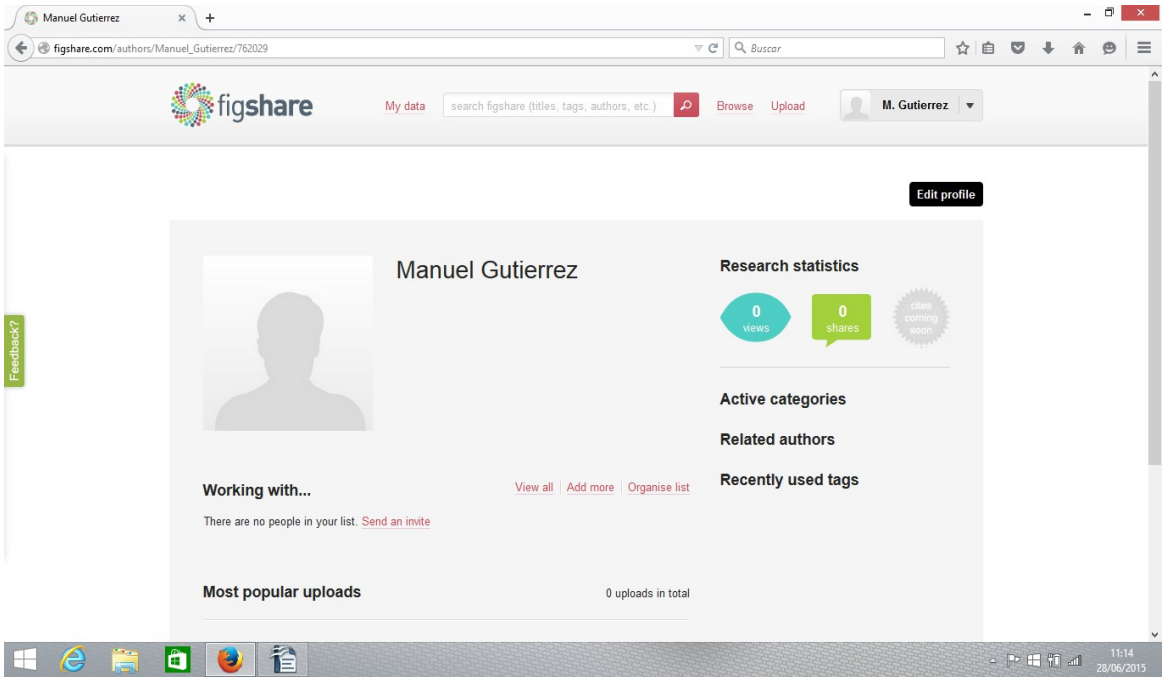

Ilustración 11. Portal del repositorio de datos de investigación *Figshare*.

Figshare es un repositorio de datos creado por Mark Hahnel. Comenzó su andadura en enero de 2011. Ofrece la posibilidad de publicar todos los datos de la investigación incluso los negativos. También otros materiales que se hayan producido durante la investigación. Todo lo publicado se hace de manera que pueda ser encontrado fácilmente, citado, compartido y reconocido. Sus usuarios comparten su material de forma abierta pero también da la posibilidad de compartir de forma privada con tus colaboradores de investigación y ofrece la posibilidad de 1GB para almacenar todo tu material de forma privada; esto te permite tener todo tu material almacenado de forma centralizada.

Lo que diferencia a *Figshare* de otros repositorios es el potencial de visibilidad que ofrece a las investigaciones y conjuntos de datos porque sus publicaciones aparecen posicionadas en los buscadores de la red debido a que sus contenidos están indexados. Permite integrar y distribuir contenidos en redes sociales, publicación de investigaciones negativas o erróneas que pueden ser útiles y el incremento del impacto de las publicaciones al contener estadísticas del impacto y citas de cada publicación; se puede medir el impacto social que pueda tener una publicación (*altmetrics*). Además está integrado en *ImpacStory.*

Su aspecto visual resulta muy agradable y da la opción de realizar unos posters visualmente muy atractivos donde aparecen los datos de la publicación. Su manejo y navegación es muy sencilla. Se podría decir que Figshare es un repositorio que fomenta la ciencia abierta y la colaboración de tal manera que su dinámica se asemeja un poco a la de las redes sociales; salvando las distancias. Sin duda su dinámica marca una gran diferencia con el resto de repositorios de datos analizados para la elaboración de este trabajo; mucho más estáticos.

wwwPDB

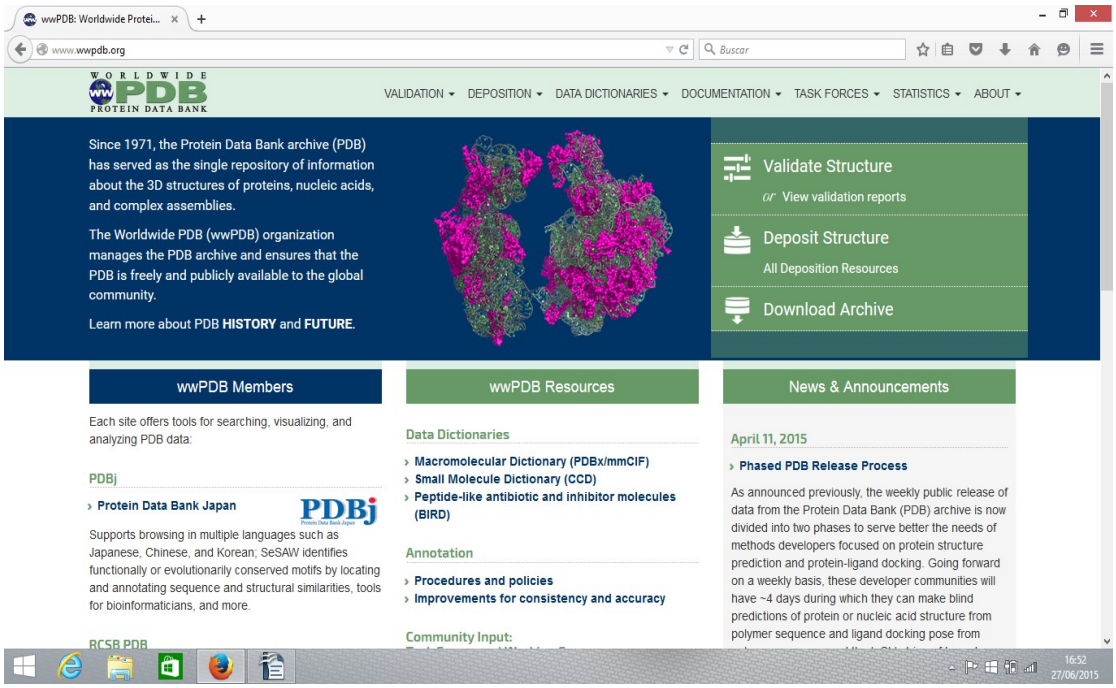

Ilustración 12. Portal del repositorio de datos de investigación *wwwPDB*

La *Worldwide Protein Data Bank* (wwPDB) es un banco de datos de Proteómica; gratuito y de acceso abierto. Desde 1971 archiva datos de estructuras de macromoléculas, proteínas, ácidos nucleicos y ensamblajes complejos. Está formada por cuatro miembros con sus respectivas bases de datos:

- *Research Collaboratory for Structural Bioinformatics Protein Database* (RCSB PDB)
- *Protein Data Bank in Europe* (PDBe)
- *Protein Data Bank Japan* (PDBj)
- *Biological Magnetic Resonance Data Bank* (BMRB)

La financiación la reciben de múltiples instituciones:

- RCSB PDB: Fundación Nacional de Ciencias, la Biblioteca Nacional de Medicina, el Instituto Nacional de Ciencias Médicas Generales, el Departamento de Energía, Instituto Nacional del Cáncer, Instituto Nacional de Trastornos Neurológicos y Accidentes Cerebrovasculares y el Instituto Nacional de la Diabetes y Enfermedades Digestivas y Renales.

- PDBE: Laboratorio Europeo de Biología Molecular, Biotecnología y Ciencias Biológicas de Investigación, los Institutos Nacionales de Salud y la Unión Europea.

- PDBj: Centro nacional de biociencia de Japón

- BMRB: Biblioteca Nacional de Medicina.

3.2. CRITERIOS PARA LA EVALUACIÓN DE LOS REPOSITORIOS.

A continuación se exponen los criterios escogidos para la elaboración de la tabla comparativa, que dividen en 6 categorías: Políticas, Aspectos legales y protección de autoría, Accesibilidad y disponibilidad, Visibilidad, impacto y reconocimiento, Seguridad, autenticidad e integridad de los datos, y por último, los que tienen que ver con Interfaz y *software*.

Políticas
1. Tipo de repositorio: institucional, disciplinar, multidisciplinar.
2. Expone de forma clara sus políticas de datos: SI/NO
3. El repositorio expone de forma clara la misión, objetivos, alcance y funciones del mismo: SI/NO
4. Expone de forma clara su política de preservación de datos: SI/NO
5. Expone de forma clara su política sobre reutilización de los datos o *datasets*: SI/NO
6. Presenta información clara sobre quién puede depositar los datos, la manera de hacerlo y en que formato: SI/NO
Aspectos legales y protección de autoría
7. Provee información sobre licencias a las que están sujetos los datasets: SI/NO
8. Indica como citar sus datasets: SI/NO
Accesibilidad y disponibilidad
9. El acceso a los datos se encuentra: abierto/restringido
10. Facilita buscadores para recuperar los *datastes*: SI/NO
11. Incluye funciones de búsqueda avanzada: SI/NO
12. Proporciona información adicional y ayuda para la búsqueda de *datasets* cuando la disciplina o especialidad a la que pertenecen se caracteriza por su complejidad: SI/NO
Visibilidad, Impacto y reconocimiento:
13. Ofrece estadísticas sobre sus datos: SI/NO
14. Ofrece estadísticas sobre total de *datasets* que contiene: SI/NO
15. Ofrece estadísticas sobre descargas realizadas totales: SI/NO
16. Ofrece estadísticas sobre descargas realizadas para cada *dataset*: SI/NO
17. Ofrece información sobre número de citas que recibe cada *dataset*: SI/NO
18. Ofrece información sobre el número de veces que un autor ha sido citado: SI/NO
19. Ofrece información sobre el número de veces que un autor ha depositado datos: SI/NO
20. Aparecen los *datasets* posicionados en buscadores de la internet: SI/NO
21. Ofrece posibilidad de integrar los *datasets* en redes sociales: SI/NO
22. Permite depositar *datasets* negativos: SI/NO
Seguridad, autenticidad e integridad de los datos
23. Dispone de un procedimiento sobre la elaboración de copias de seguridad de los *datasets*: SI/NO
24. Dispone de identificador digital para sus *datasets*: SI (especificar) /NO
25. Algún tipo de control de calidad de los datos publicados: SI /NO
Interfaz y *software*.
26. Tipo de *software*.
27. Dispone de una interfaz de búsqueda amigable: SI/NO
28. La navegación es sencilla: SI/NO
29. Visualmente presenta la información de forma clara: SI/NO

Tabla 9. Las 6 categorías con los 29 criterios escogidos para el análisis.

3.3. TABLA COMPARATIVA DE LOS REPSOSITORIOS DE DATOS.

	Dryad	Genbank	PANGAEA	SIMBAD	Zenodo	Figshare	wwPDB
1	Disciplinar	Disciplinar	Disciplinar	Disciplinar	Multicisciplinar	Multicisciplinar	Disciplinar
2	SI	NO	SI	NO	SI	SI	SI
3	SI	SI	NO	SI	SI	SI	SI
4	SI	NO	NO	NO	SI	NO	NO
5	SI	SI	NO	NO	SI	NO	NO
6	SI	SI	SI	SI	SI	SI	SI
7	SI	SI	SI	NO	SI	SI	NO
8	SI	NO	NO	NO	NO	SI	NO
9	Abierto	Abierto	Abierto	Abierto	Abierto	Abierto	Abierto
10	SI	SI	SI	SI	SI	SI	SI
11	SI	SI	SI	SI	NO	SI	SI
12	NO	SI	SI	SI	NO	NO	SI
13	SI	SI	NO	SI	NO	SI	SI
14	SI	SI	NO	SI	NO	NO	SI
15	SI	NO	NO	NO	NO	NO	SI
16	SI	NO	NO	NO	NO	SI	NO
17	NO	NO	NO	NO	NO	SI	NO
18	NO	NO	NO	NO	NO	SI	NO
19	NO	NO	NO	NO	NO	NO	NO
20	NO	NO	NO	NO	NO	SI	NO
21	SI	NO	NO	NO	SI	SI	NO
22	NO	NO	NO	NO	NO	SI	NO
23	SI	No se sabe	No se sabe	No se sabe	SI	SI	No se sabe
24	DOI	No se sabe	DOI	DOI	DOI	DOI	No se sabe
25	SI	SI	SI	SI	SI	NO	SI
26	Dspace	Microsoft SQL server	Elasticsearch	PostgreSQL	Invenio	Figshare	No se sabe
27	SI	NO	SI	SI	SI	SI	NO
28	SI	NO	SI	SI	SI	SI	NO
29	SI	NO	SI	SI	SI	SI	SI

Tabla 10. Tabla comparativa con las respuestas a los criterios de evaluación de los repositorios de datos.

3.4. ANÁLISIS DE LA TABLA COMPARTIVA.

Al analizar los criterios que tratan sobre las políticas de los repositorios elegidos se observa que el tipo de repositorios que más abunda es el temático o disciplinar. Sólo *Zenodo* y *Figshare* son de carácter multidisciplinar probablemente debido a que son de reciente creación; a diferencia del resto que tienen mucho más tiempo. El contexto de *open science* ha propiciado que se rompa con la tendencia a trabajar abarcando sólo un campo o especialidad y fomentando repositorios multicisciplinares que promueven la colaboración; abarcando toda la ciencia.

En general, todos los repositorios de datos muestran de forma clara y bien visible la política de datos. Sólo *SIMBAD* y *Genbank* son más ambigüos y disponen de información más dispersa sobre el asunto. No es que no tengan o no muestren información sobre sus políticas pero es demasiado escasa y fragmentada. (En el caso de *SIMBAD* prácticamente inexistente). La misión, objetivos, alcance y funciones de los repositorios se muestran en todos salvo en *PANGAEA* lo cual sorprende ya que es algo que debería mostrar todo repositorio de forma explícita. *Dryad* y *Zenodo* son los únicos repositorios de la tabla que exponen de forma clara su política de preservación de datos. En cuanto a exponer de forma clara la política sobre reutilización de los datos o *datasets* sólo *Dryad*, *Zenodo* y *Genbank* muestran información al respecto. En lo único que todos coinciden, aunque de forma desigual, es en proporcionar información clara sobre quién puede depositar los datos, la manera de hacerlo y en que formato.

Se observa que no hay homogeneidad a la hora de presentar las políticas por parte de los repositorios. La cantidad o claridad de información que se ofrece es bastante desigual o se limita a remitir a documentos de organizaciones internacionales o nacionales sobre las que siguen sus directrices o buenas prácticas. Suele presentarse dispersa en distintos apartados: en inicio, políticas, en apartado de preguntas frecuentes, etc. Sería bueno que los repositorios dedicaran tiempo a elaborar una buena presentación de las políticas que siguen sobre datos. Que estuviera bien visible y que contuviera toda la información posible de forma sintetizada; todo ello agrupado en un mismo apartado.

Los investigadores deben sentir que el repositorio les aporta garantías y estar bien informados sobre el alcance de las políticas de datos es una pequeña ayuda en ese sentido. Todos los repositorios poseen un apartado de contacto donde los investigadores podrían exponer sus dudas e incertidumbres pero es más práctico si todo lo básico está bien expuesto y detallado en la web. Algo en que fallan los repositorios analizados es que salvo *Dryad* y *Zenodo*, el resto no exponen de forma clara su política de preservación de los datos y ambos, junto con *Genbank,* son los que más claro muestran cual es su política de reutilización. Mostrar estas dos informaciones es necesario para ayudar a evitar ese clima de incertidumbre que crea reticencias en los investigadores a la hora de compartir sus datos.

Los repositorios *SIMBAD* y *wwwPDB* no proveen información clara sobre las licencias a las que están sujetos los datos que albergan. De todas formas esto no quiere decir que sus datos no estén sujetos a ningún tipo de licencia o propiedad intelectual que respete la autoría pero en sus webs no proporcionan esta información. Por otro lado los repositorios *Dryad* y *Figshare* son los únicos que proporcionan información acerca de como citar sus *datasets*.

Se observa que una vez más que no hay una forma de homogénea ni normalizada de informar acerca de dos cuestiones fundamentales para que la práctica del d*ata sharing* se generalice. Los datos deben estar protegidos con licencias que puedan asegurar la autoría de sus creadores y productores. Así mismo, los repositorios deben proporcionar la información sobre las licencias a las que están sujetos sus datos para que los investigadores puedan publicar sus datos con garantías y que los usuarios de esos datos las tengan en cuenta a la hora de su reutilización.

Indicar como citar los datos es otra práctica menos común aún en los repositorios analizados. Es necesario generalizar esta práctica e intentar normalizarla para que todos los repositorios puedan ofrecer una forma similar de citar; teniendo en cuenta siempre la variedad de datos y diferencias de estos entre las distintas disciplinas. La citación como factor que aumenta el impacto y el reconocimiento de los investigadores es uno de los estímulos que puede revertir la actual situación en que los investigadores recelan de publicar sus datos. Que las licencias a las que están sujetos los datos y la forma de citar este bien visible y clara fomenta, junto con las políticas, un entorno con más garantías y certezas. Los investigadores se pueden sentir más seguros y proclives a publicar sus datos de investigación.

Una de las características del *data sharing* es que los datos deben estar accesibles y fáciles de localizar. Esta condición se cumple para los repositorios analizados ya que todos son de acceso abierto. También facilitan buscadores para los *datasets* e incluyen opción de búsqueda avanzada. El único que no presenta búsqueda avanzada es *Zenodo*.

En cuanto a proporcionar información adicional cuando la búsqueda de datos implica cierta complejidad la respuesta es desigual. Su ausencia en los casos de *Zenodo* y *Figshare* puede ser por su condición de repositorios multidisciplinares; no especializados en ningún campo concreto. Tampoco lo ofrece *Dryad*. El resto de repositorios si ofrecen guías, documentos, diccionarios, listados, etc. que ayudan a sus usuarios. Esta ayuda es fundamental para la correcta búsqueda y localización de datos que son complejos hasta para los investigadores especializados. Tal vez por tratarse de una cuestión más práctica los repositorios cumplen de sobra con esta función necesaria para el *data sharing* ya que facilita que los datos pertinentes puedan ser fácilmente localizables.

Los únicos que no ofrecen estadísticas sobre sus datos son *PANGAEA* y *Zenodo*. Además *Figshare* no ofrece estadísticas sobre total de *datasets* que contiene. *Dryad* y *wwwPDB* no ofrecen información sobre estadísticas del total de descargas que se han realizado en sus repositorios. Ninguno de los repositorios consultados ofrece datos estadísticos sobre el número de veces que un autor ha depositado datos.
La posibilidad de comprobar todos estos datos estadísticos permite a un investigador decidir mejor el repositorio en que albergar sus datos; porque puede ser beneficioso para su publicación que sus datos estén depositados en un repositorio muy activo y con trayectoria.

Se observa que el repositorio que ofrece más información estadística sobre cada *dataset* o citas es *Figshare*, en menor medida *Dryad* y el resto nada. *Dryad* y *Figshare* ofrecen estadísticas sobre el número de veces que se ha descargado cada *datase.* Además *Figshare* es el único que ofrece información sobre número de citas que recibe cada *dataset,* sobre el número de veces que un autor ha sido citado, si aparecen los *datasets* posicionados en buscadores de la internet y el único que permite depositar *datasets* negativos. También *Figshare* y *Zenodo* permiten integrar los *datasets* en redes sociales.

Ante esto, creo que una de las formas que puede fomentar el *data sharing* es que la publicación de datos de investigación ofrezca una nueva manera de ofrecer visibilidad a las publicaciones de los investigadores. Ofrecer información acerca de estadísticas de descargas de datos, número de citaciones de un *datasets*, etc. ayuda a aumentar la visibilidad. Sobretodo porque ofrecen un nuevo canal para medir el impacto de una publicación (*almetrics*). Pueden influir en el impacto y por tanto el aumento de reconocimiento. Además si los contenidos se pueden integrar en redes sociales y aparecer posicionados en los buscadores de internet el potencial de visibilidad para un investigador aumenta. Por esta razón *Figshare* es el repositorio de todos los analizados en la tabla que más posibilidades ofrece en este sentido. El resto de repositorios analizados no ofrecen estas posibilidades lo que considero un error desde el punto de vista de intentar romper con las reticencias de los investigadores a publicar sus datos y crear una cultura de *data sharing.*

También el hecho de que se permita publicar datos negativos puede influir en la visibilidad del autor o investigación (*Figshare* lo permite). La colaboración es una de las características del *data sharing* y que alguien comparta sus datos negativos, los que contienen errores, permiten a otros no cometerlos. Los *datasets* negativos son parte de una investigación y su reutilización o descarga aumenta las estadísticas y por tanto podrían aumentar la visibilidad e impacto. Por esta razón, es algo a tener en cuenta.

Dryad, *Zenodo* y *Figshare* si muestran que disponen de un procedimiento sobre la elaboración de copias de seguridad de los *dataset*. *Genbank*, *PANGAEA*, *SIMBAD* y *wwwPDB* no muestran información al respecto. Todos comparten la característica de disponer de algún tipo de control de calidad de los datos publicados salvo *Figshare* que no muestra información clara. En cuanto a la cuestión de si utilizan algún tipo de identificador digital la respuesta es bastante unánime ya que casi todos lo utilizan y además el mismo. Utilizan el *Digital Object Identifier* (DOI). Ha sido imposible averiguar que identificador digital utilizan *Genbank* y *wwwPDB*. Para un investigador que quiera publicar sus datos es necesario ofrecerle alguna garantía de preservación digital por lo que los repositorios deberían mostrar información clara sobre si realizan copias de seguridad. En cuanto al control de calidad de los datos deberían ser aún más claros ya que los usuarios necesitan garantías de fiabilidad si deciden a reutilizar datos.

En cuanto a software las respuestas han sido muy variadas. Se esperaba que la mayoría dispusiera del software *Dspace* debido a que era el que más aparecía en la literatura científica sobre repositorios de datos pero no ha sido así. Sólo *Dryad* lo utiliza. El resto cada uno utiliza uno y en el caso de *Figshare* utiliza un software de desarrollo propio. *WwwPDB* ha sido imposible averiguar que software utiliza.
Casi todos utilizan una interfaz de búsqueda amigable, disponen de una navegación sencilla y presentan la información de forma clara por lo que buscar los datos resulta sencillo y fácil. Los únicos casos en los que no es así son *Genbank* y *wwwPDB*. Presentan interfaces muy obsoletas y visualmente poco amigables.

3.5 CONSIDERACIONES GENERALES.

Al analizar los resultados obtenidos se observa que en muchos criterios hay una respuesta bastante desigual. En general todos ofrecen información sobre las garantías que ofrecen a los investigadores respecto a sus datos de investigación pero en algunos casos siempre hay algún tipo de información necesaria que no se aporta.

Se puede decir que se muestra información a los investigadores sobre sus políticas de datos, preservación y reutilización pero hay algunos aspectos que no se aclaran o se nota su falta. No hay homogeneidad al respecto, la información que aporta uno otro no y viceversa. Ocurre lo mismo con la información sobre aspectos legales y protección de autoría. Tal vez una de las cuestiones más importantes a la hora de motivar a los investigadores a publicar sus datos junto con la categoría de visibilidad, impacto y reconocimiento.

En cuanto a información que pueda ayudar a los investigadores a aumentar su visibilidad, impacto y reconocimiento hay más homogeneidad y en general gana la respuesta negativa. Sólo ofrecen una información básica. Es otra de las cuestiones que podría animar a los investigadores a publicar sus datos, además de facilitar la métrica de datos, pero no es una información plenamente desarrollada. La posibilidad de obtener mayor visibilidad, impacto y reconocimiento supone uno de los mayores incentivos para publicar.

Hay más consenso sobre la información sobre seguridad, autenticidad e integridad de los datos. Sólo destaca por su ausencia la información que tiene que ver con la realización de copias de seguridad. En los otros dos criterios hay homogeneidad en las respuestas ya que la mayoría utiliza el mismo identificador digital y tiene controles de calidad. Aunque para cada criterio siempre como ocurre en las categorías anteriores, siempre hay un par de casos que no ofrece este tipo de información.

Por último las categorías de accesibilidad y disponibilidad de los datos e interfaz y software responden de forma positiva y la desigualdad de respuestas es menor. Sólo destaca la respuesta al criterio del tipo de software ya que muestra que cada repositorio utiliza uno distinto y no hay una tendencia a utilizar *Dspace* como se observó en la literatura científica.

Si nos ceñimos a los objetivos planteados al inicio de este trabajo, se podría decir que en general los repositorios de datos de investigación si ofrecen algo de información sobre las garantías que ofrecen a los investigadores para evitar las reticencias a publicar de estos. Pero no ofrecen toda la que se debe ofrecer y es una información demasiado escasa y fragmentada. Además la información que ofrece cada uno varía respecto a lo que publican otros, es desigual. En algunos casos incluso ambigüa o se desconoce. Se debería tender a la normalización y a un mínimo de estandarización de la comunicación de la información que ofrecen los repositorios para favorecer la promoción del *data sharing.* Una forma muy visual y que sintetiza mucho la comunicación de esas garantías podría ser el uso de una serie de iconos, dispuestos juntos en alguna parte de la página principal, que representen cada uno de los aspectos mencionados y que pueden servir incluso a modo de certificados.

4. CONCLUSIONES.

El objetivo de este trabajo, además de ofrecer una visión global del *data sharing*, era averiguar si una serie de repositorios de datos de investigación ofrecían una buena comunicación de la información acerca de las garantías que podrían ofrecer a los investigadores para que depositaran y reutilizarán los datos; y por tanto participaran de la cultura del *data sharing.* El estudio ha demostrado que si ofrecen información pero con deficiencias. El papel de los repositorios en este sentido es una pequeña aportación a la promoción del *data sharing* pero no por ello debe ser infravalorada o ignorada. Toda aportación a esta causa es importante por el motivo de que actualmente la reticencia a publicar los datos es uno de los mayores problemas a los que se enfrenta el *data sharing*.

La bibliometría es una disciplina cuya misión principal es el estudio y medición de la producción científica. Lo que en un principio era simplemente una forma de medir la ciencia acabó por influir en la forma de produción de la literatura científica. La acción de observación y medición de la producción científica modificó la misma. Algo que recuerda, salvando las distancias por supuesto, al principio de incertidumbre de Heisenberg y la paradoja del *gato* de Schrödinger. Después de la aparición de la bibliometría los investigadores están más condicionados en su forma de publicar y en el hecho mismo de obligarse a tener que generar publicaciones. Este imperativo o necesidad hace que la producción científica y la ciencia parezcan que están más condicionados por el progreso de las carreras de los investigadores que por el de la ciencia; prima el reconocimiento científico de los autores frente al progreso o avance científico colectivo. Esto puede ofrecer una visión gruesa de por qué existe tanta reticencia por parte de los investigadores a publicar los datos de sus investigaciones.

Se podría pensar que con el actual contexto de la e-ciencia y de la *open science* (donde cada vez se habla más de datos abiertos, compartir, colaboración, etc.) se puede dar un cambio de paradigma en el que se inviertan las prioridades. Pasando a ser el progreso o avance de la ciencia colectivo lo que prime frente al reconocimiento de los autores. Pero el hecho de que aún exista esa reticencia a publicar los datos demuestra que de momento no es así. Toda investigación conlleva un esfuerzo, tanto humano como económico, por lo que publicar los datos de investigación para que puedan ser reutilizados por otros investigadores que se ahorrarán ese esfuerzo desmotiva a sus creadores. Para motivar a publicar los datos se necesita un incentivo que responda más a los intereses profesionales de los investigadores.

Se ha comprobado en el trabajo que los datos o *datasets* se consideran recursos propios independientes que pueden ser tratados. Atendiendo a lo mencionado en los párrafos anteriores, considero que el desarrollo de mecanismos y métricas de datos (*data metrics*) es de vital importancia para solventar el problema de la reticencia a publicar datos. Si se pueden citar y medir su impacto el autor o creador puede obtener visibilidad y reconocimiento; un incentivo más ajustado a sus intereses profesionales. Al igual que la bibliometría influyó en la producción científica, la medición de la producción científica de datos de investigación podría influir en la producción y publicación de estos. Si esto ocurre puede que si se redefina un nuevo paradigma.

Existe otro elemento a tener en cuenta en el cambio de paradigma de la investigación científica donde la publicación de los datos de investigación va a jugar un papel fundamental. Se trata del *Big data* (datos masivos). La revolución que el *big data* está generando va a cambiar u ofrecer un nuevo punto de vista a la hora de realizar investigaciones científicas. El *big data* es el tratamiento de enormes cantidades de datos masivos que gracias a los avances de la tecnología en almacenamiento de información y la generación de complejos algoritmos para su tratamiento va a generar un conocimiento nunca antes visto en la historia de la humanidad. Dejando a un lado todo lo que supondrá el *big data* para la sociedad y centrándonos en su papel en la ciencia, podremos hablar de un cambio de paradigma a la hora de realizar una investigación y el papel necesario que juega la publicación de los datos de investigación.

Lo que va a cambiar o aportar el *big data* es un cambio frente a la forma clásica de investigación en la cual se parte de una hipótesis concreta y después se investiga para confirmar o refutar dicha hipótesis. Además cuando se investiga de esta forma se hace sobre algo concreto, las investigaciones están encaminadas a investigar algo con unos límites bien establecidos; es decir, se tiende a la especificidad. Un ejemplo de esto podría ser una investigación sociológica para la que se establece una hipótesis de partida y en la investigación el investigador tiene que especificar sobre que población concreta realiza su investigación y tratar de dar en su investigación un resultado muy específico para que la investigación establezca una respuesta específica a la hipótesis; una respuesta única. Con el *big data* no es necesario establecer hipótesis ni dar una respuesta única y específica. Lo que se hace es tratar, sin necesidad de plantear una hipótesis inicial, directamente los datos masivos con algoritmos y ver que nos muestran estableciendo tendencias que tienden a la probabilidad y no a la especificidad.

Para que el *big data* funcione se necesita de gran cantidad de datos y es aquí donde la publicación de los datos de investigación y el *data sharing* juegan un papel fundamental. Es necesario que la cultura del *data sharing* crezca y se consolide porque ello producirá que estén disponibles enormes cantidades de datos de investigación cuyo tratamiento y combinación, desde el punto de vista del *big data,* puede ofrecer unos resultados con un alcance y potencial revelador nunca visto antes en la historia de la ciencia. (También se presupone que los datos de investigación científicos reutilizados por el mundo empresarial mediante *big data* ofrecerá nuevas oportunidades económicas)

Uno de los criterios de evaluación y análisis de los repositorios de datos elegidos para este trabajo era si permitían depositar *datasets* negativos. Se realizó está cuestión porque una de las características del *big data* es que nunca se sabe que datos pueden ser útiles o que pueden mostrar; por lo que se tiende a almacenar todo tipo de datos. Los datos negativos además de ayudar a mostrar errores a los investigadores que practiquen el *data sharing* también tienen potencial para actividades de *big data*. Sólo *Figshar*e ofrecía esta opción.

Hemos visto en el trabajo la importancia y necesidad de un buen plan de gestión de datos y la relevancia de la actividad de *data curation* para una buena preservación digital. También que otro de los problemas del *data sharing* es la falta de formación para la gestión correcta de los datos de investigación. Es aquí donde los profesionales de la información tienen una oportunidad de desempeñar una labor importante. El *data sharing* ofrece un entorno con un potencial profesional a tener en cuenta por los bibliotecarios y estudiantes de grado de información y documentación.

Su papel en este nuevo contexto puede incidir en el desarrollo y establecimiento de políticas, directrices y guías de buenas prácticas sobre la gestión de los datos de investigación. Participar en la actividad de *data curation* en la preservación digital de los datos o como bibliotecario especializado en algún campo científico para servir de intermediario entre los investigadores y la información. Además de asesorar o formar en gestión de datos de investigación a los propios investigadores.

Estas competencias se pueden desempeñar en el entorno universitario donde se generan muchas de las investigaciones. Es necesario que las universidades empiecen a tener en cuenta los datos de las investigaciones que generan sus investigadores. Se ha observado en el trabajo que en el caso español la gestión de datos y creación de repositorios de datos de investigación es prácticamente inexistente.

Por último y con una visión global del *data sharing* surgen otras posibilidades de investigación sobre el *data sharing.* Se pueden proponer como líneas de investigación futuras los siguientes temas relacionados:

- Estudiar la situación de los repositorios de datos de investigación españoles.
- Analizar la gestión de datos de investigación en la la Universidad de Salamanca.
- Profundizar en los nuevos perfiles del profesional de la información ante el *data sharing*
- Desarrollar métricas de datos adecuadas a este nuevo contexto colaborativo.
- Establecer mecanismos para evitar errores y fraude en la ciencia en el contexto *data sharing*.

5. BIBLIOGRAFÍA.

ALEIXANDRE-BENAVENT, R., et al., 2015. Disponibilidad en abierto de los artículos y de los datos brutos de investigación en las revistas pediátricas españolas. Anales de Pediatría[en línea]. Madrid: Asociación Española de Pediatría, vol.82, no.1, pp .e90-e94 [consulta: 10 abril 2015]. ISSN 1695-4033. Disponible en: http://dx.doi.org/10.1016/j.anpedi.2013.11.014

ALEIXANDRE-BENAVENT, R., et al., 2013. Compartir datos de investigación en cardiología. Revista Española de Cardiología [en línea]. Madrid: Sociedad Española de Cardiología, vol. 66, no. 12, pp. 1007-1008 [consulta: 10 abril 2015]. ISSN 0300-8932. Disponible en: http://dx.doi.org/10.1016/j.recesp.2013.08.005

ARGUIMBAU, L., 2013. Les dades de recerca: una oportunitat professional per als gestors d'informació. Item: revista de biblioteconomia i documentació [en línea]. Barcelona: COBDC. no. 57, pp. 37-56. [consulta: 20 marzo 2015]. ISSN 1699-521X. Disponible en: http://www.raco.cat/index.php/Item/article/view/269702

BARRUECO CRUZ, J.M. (coord.), 2014. *Guía para la evaluación de repositorios institucionales de investigación* [e-book] Madrid: FECYT, RECOLECTA, CRUE [consulta: 20 mayo 2015]. Disponible en: http://recolecta.fecyt.es/sites/default/files/contenido/documentos/GuiaEvaluacionRecolecta_v.ok_0.pdf

BORGMAN, C.L., 2012. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology [en línea]. [Hoboken (New Jersey)]: Wiley-Blackwell, vol. 63, no. 6, pp. 1059-1078 [consulta: 25 marzo 2015]. ISSN 15322882. Disponible en: http://doi.wiley.com/10.1002/asi.22634

BORREGO, Á., 2012. Los retos de la gestión de datos de investigación. En: Blok de BiD [en línea]. Barcelona: Universitat de Barcelona [consulta: 28 marzo 2015]. Disponible en: http://www.ub.edu/blokdebid/es/content/los-retos-de-la-gestión-de-datos-de-investigación

COSTAS, R., et al., 2013. The value of research data - Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report [en línea]. [Leiden]: Leiden University [consulta: 15 marzo 2015]. Disponible en: http://www.knowledge-exchange.info/datametrics

GARCÍA-GARCÍA, A., et al., 2012. ODiSEA: International registry on research data. BiD: textos universitaris de biblioteconomia i documentació [en línea]. Barcelona: Universitat de Barcelona, vol. 29 [consulta: 17 marzo 2015]. ISSN 1575-5886. Disponible en: http://dx.doi.org/10.1344/BiD2012.29.12

GRUPO DE TRABAJO DE "DEPÓSITO Y GESTIÓN DE DATOS EN ACCESO ABIERTO" DEL PROYECTO RECOLECTA,2012. La conservación y reutilización de los datos científicos en España. Informe del grupo de trabajo de buenas prácticas [en línea]. Madrid: Fundación Española para la Ciencia y Tecnología [consulta: 10 marzo 2015]. Disponible en: http://hdl.handle.net/10261/65317

GONZÁLEZ, L.M., et al., 2013. Gestión de datos de investigación: infraestructuras para su difusión. El Profesional de la Información [en línea]. Barcelona: Editorial UOC, vol. 22, no. 5, pp. 415-423 [consulta: 17 marzo 2015]. ISSN 1386-6710. Disponible en: http://dx.doi.org/10.3145/epi.2013.sep.06

HERNÁNDEZ-PÉREZ, T. y GARCÍA-MORENO, M.A., 2013. Datos abiertos y repositorios de datos: nuevo reto para los bibliotecarios. El Profesional de la Información [en línea]. Barcelona: Editorial UOC, vol. 22, no. 3, pp. 259-263 [consulta: 12 marzo 2015]. ISSN 1386-6710. Disponible en: http://dx.doi.org/10.3145/epi.2013.may.10

KOWALCZYK, S. y SHANKAR, K., 2011. Data sharing in the sciences. Annual Review of Information Science and Technology [en línea]. [Medford (New Jersey)]: Information Today, vol. 45, no. 1, pp. 247-294 [consulta: 17 abril 2015]. ISSN 0066-4200. Disponible en: http://dx.doi.org/10.1002/aris.2011.1440450113

LABASTIDA, I., 2013. Les dades de la recerca: de la foscor a la claror. Item: revista de biblioteconomia i documentació [en línea]. Barcelona: COBDC, no. 57 [consulta: 20 marzo 2015]. ISSN 1699-521X. Disponible en: http://www.raco.cat/index.php/Item/article/view/269703

MELERO, R. y HERNÁNDEZ-SAN-MIGUEL, J., 2014. Acceso abierto a los datos de investigación, una vía hacia la colaboración científica. Revista española de Documentación Científica [en línea]. Madrid: CSIC, vol. 37, no. 4, pp. e066 [consulta: 12 marzo 2015]. ISSN 1988-4621. Disponible en: http://dx.doi.org/10.3989/redc.2014.4.1154

NINA-ALCOCER, V., BLASCO-GIL, Y. y PESET, F., 2013. Datasharing: Guía práctica para compartir datos de investigación. El Profesional de la Información [en línea]. Barcelona: Editorial UOC, vol. 22, no. 6, pp. 562-568 [consulta: 10 mayo 2015]. ISSN 1386-6710. Disponible en: http://dx.doi.org/10.3145/epi.2013.nov.09

PIWOWAR, H.A. ,2011. Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data. En: PLoS ONE [en línea]. [Cambridge]: PLoS, vol. 6, no. 7, pp. e18657 [consulta: 16 abril 2015]. ISSN 1932-6203. Disponible en: http://dx.doi.org/10.1371/journal.pone.0018657

PIWOWAR, H.A. y CHAPMAN, W.W. ,2010. Public sharing of research datasets: A pilot study of associations. Journal of Informetrics [en línea]. [Amsterdam]: Elsevier, vol. 4, no. 2, pp. 148-156 [consulta: 17 abril 2015]. ISSN 1751-1577. Disponible en: http://dx.doi.org/10.1016/j.joi.2009.11.010

PIWOWAR, H.A., CHAPMAN, W.W., 2008. Identifying data sharing in biomedical literature. AMIA Annual Symposium proceedings archive [en línea]. [S.l.]: AMIA, pp. 596-600 [consulta: 17 abril 2015]. ISSN 1942-597X. Disponible en: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655927/

TENOPIR, C., et al., 2011. Data Sharing by Scientists: Practices and Perceptions. En: PLoS ONE [en línea]. [Cambridge]: PLoS, vol. 6, no. 6, pp. e21101 [consulta: 26 abril 2015]. ISSN 1932-6203. Disponible en: http://dx.doi.org/10.1371/journal.pone.0021101

TORRES-SALINAS, D., 2010. Compartir datos (data sharing) en ciencia: contexto de una oportunidad. *Anuario ThinkEPI*. Barcelona: Editorial UOC, vol. 4, pp. 258-261. ISSN 1886-6344.

TORRES-SALINAS, D., 2010. Hacia la gestión de datos de investigación en las universidades: la Data Asset Framework. *Anuario ThinkEPI*. Barcelona: Editorial UOC, vol. 4, pp. 262-265. ISSN 1886-6344.

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 [en línea]. Barcelona: Editorial UOC, vol. 21, no. 2, pp. 173-184 [consulta: 10 marzo 2015]. ISSN 1386-6710. Disponible en: http://dx.doi.org/10.3145/epi.2012.mar.08

WALLIS, J.C., ROLANDO, E. y BORGMAN, C.L. ,2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. En: PLoS ONE [en línea]. [Cambridge]: PLoS, vol. 8, no. 7, pp. e67332 [consulta: 18 abril 2015]. ISSN 1932-6203. Disponible en: http://dx.doi.org/10.1371/journal.pone.0067332

6. RECURSOS.

Scopus: http://www.scopus.com/

Web Of Science: http://webofscience.com/

LISA: http://search.proquest.com/lisa

Base de datos del CSIC: http://bddoc.csic.es:8085/

Open Definition Advisory Council: http://opendefinition.org/od/

Berners-Lee, Tim. Esquema de desarrollo de 5 estrellas para Datos Abiertos: http://5stardata.info/

Open Data Handbook: http://opendatahandbook.org/guide/es/what-is-open-data/

OCDE (2007) *Principles and guidelines for acces to research data from public funding;* http://www.oecd.org/sti/sci-tech/38500813.pdf

Protocolo de remisión, almacenamiento y difusión de datos antárticos en España (2004-2007); http://www.idi.mineco.gob.es/stfls/MICINN/Investigacion/FICHEROS/2008-final-protocolo.pdf

Centro Nacional de datos polares en España: http://hielo.igme.es/index.php/es/9-sin-categoria/73-centro-nacional-de-datos-polares

Horizon 2020. The EU Framework Programme for Research and Innovation. http://ec.europa.eu/programmes/horizon2020/en/official-documents

DCC (Digital Curation Centre): http://www.dcc.ac.uk/

Re3data: http://www.re3data.org/

Datahub: http://datahub.io/es/

ODISEA: International Registry on Research Data: http://odisea.ciepi.org/es

Dryad: http://datadryad.org/

Genbank: http://www.ncbi.nlm.nih.gov/genbank

SIMBAD: http://simbad.u-strasbg.fr/simbad/

PANGAEA: http://www.pangaea.de/

Zenodo: http://www.pangaea.de/

Figshare: http://figshare.com/

wwPDB: http://www.wwpdb.org/

Scientific Data: http://www.nature.com/sdata/

Journal of Open Public Health Data: http://openhealthdata.metajnl.com/

Digital.CSIC: https://digital.csic.es/?locale=en

UPF Digital Repository -Recursos i dades primàries: http://repositori.upf.edu/handle/10230/5963

CEACS Data Library: http://www.march.es/ceacs/biblioteca/datalib/

Banco de Datos Específico de Estudios Sociales. *CIS Data Bank*: http://www.cis.es/cis/opencms/EN/NoticiasNovedades/InfoCIS/2014/PlataformaOnLineBancodeDatos.html

Herschel Science Archive: http://archives.esac.esa.int/hsa/aio/doc/

AMIGA. Analysis of the interstellar Medium of Isolated Galaxies: http://amiga.iaa.es/p/1-homepage.htm

Dipósit Digital de la UB: http://diposit.ub.edu/dspace/handle/2445/56364