UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DCADHA
DPTO. DE COMUNICACIÓN AUDIOVISUAL DOCUMENTACIÓN E HISTORIA DEL ARTE

eCALSI

Gestión de datos de la investigación

Tesina final de estudios Máster Oficial
CALSI

Autor:
Andreu Valentín Torrecilla

Bajo la dirección de las doctoras:
Antonia Ferrer Sapena
Fernanda Peset Mancebo

Valencia, Septiembre 2013

Resumen

Este trabajo trata sobre la gestión, la recuperación y la conservación de los datos primarios de la investigación.

El trabajo se divide en dos partes bien diferenciadas, en la primera parte se analiza y se estudia qué se entiende por dato de la investigación, enmarcando el concepto en un espacio y un tiempo determinados y estudiando las definiciones que han realizado diferentes instituciones y autores, todo ello con el objetivo de entender bien el término.

También se realiza una prospección general sobre las principales entidades internacionales que están trabajando en el campo de los datos de investigación, para ofrecer, de esta manera, una visión sintetizada de todo lo que se está realizando en este ámbito.

Una vez adquiridos los conocimientos sobre los datos de investigación y las buenas prácticas que se están llevando a cabo, se realizarán una serie de conclusiones para sintetizar en qué estado se encuentra la gestión de los datos de la investigación.

Índice:

1 - Justificación

La investigación básica se realiza para incrementar la cultura, pero si esa investigación no se difunde, si el conocimiento que esa investigación genera no se comunica, entonces no hay incremento de la cultura, y si no hay incremento de la cultura como producto de la investigación básica, entonces no hay investigación básica. No hay ciencia.

Para un investigador la difusión es un punto clave ya que le da sentido a su trabajo al permitir el registro, evaluación, diseminación y acumulación del conocimiento.

Esto ya lo puso de manifiesto Einstein (1948):

"Es muy importante que se tenga la oportunidad de conocer y comprender los resultados del trabajo de investigación científica. No es suficiente que el conocimiento adquirido sea registrado, desarrollado y aplicado sólo por algunos especialistas. La limitación del capital de conocimientos a su propio círculo es la muerte del espíritu filosófico de todo un pueblo y conduce al empobrecimiento intelectual."

Los avances tecnológicos están produciendo enormes cambios en el seno de la actividad científica, cada vez hay más necesidad de comparar, preservar y gestionar grandes cantidades de datos.

El número de iniciativas internacionales dedicadas a la gestión a largo plazo de la información científica ha crecido exponencialmente.

Por eso, y aunque el acceso abierto a las publicaciones de los investigadores, en cualquiera de sus vías, es el punto de partida del trabajo, no es el eje central del proyecto, debido a que es un tema que está ya muy tratado y debatido en nuestro ámbito profesional.

En España el desarrollo de repositorios institucionales y la realización de proyectos relacionados con ellos se encuentra en una etapa emergente y los datos varían en cortos periodos de tiempo.

Contamos con dos cosechadores como Hispana y Recolecta, incluso se ha creado un corpus legislativo español con la "Ley 14/2011 de la Ciencia, la Tecnología y la Innovación", que entró en vigor el 2 de diciembre de 2011, y el "Real Decreto

99/2011, de 28 de enero, por el que se regulan las enseñanzas oficiales de doctorado“.

El trabajo presta su atención en analizar y estudiar como se gestionan los datos primarios de la investigación, entendidos estos, no como resultados publicados, sino como la estadística u otro material suplementario necesario para redactar esas publicaciones.

La gestión de datos primarios de investigación es uno de los temas de actualidad en nuestro ámbito profesional.

1.1 - Objetivo

El objetivo principal del trabajo es conocer qué son los datos de investigación, elaborando un marco teórico que ayude a entender el concepto y su contexto actual.

Se propone:

- Obtener una visión general del movimiento OA
- Realizar una aproximación al tema de los datos de investigación
- Conocer las buenas prácticas y los proyectos que se están llevando a cabo en lo relativo a la preservación de datos de la investigación, centrándonos principalmente en dos proyectos de gran calado internacional como son el *Digital Curation Centre* (DCC) y el *Australian National Data Service* (ANDS), para conseguir una visión sintetizada del contexto general.
- Realizar un conjunto de conclusiones para sintetizar en qué estado se encuentra la gestión de los datos de la investigación.

1.2 - Fuentes y metodología

El proyecto gira en torno a la idea de conocer qué son los datos de la investigación, cómo se gestionan y cuáles son las organizaciones que sirven como ejemplo de buenas prácticas en este campo.

En primer lugar se realiza una aproximación al movimiento *Open Access* utilizando una metodología de consulta de fuentes de información, interrogando a diferentes bases de datos y a los motores de búsqueda con diferentes palabras clave, tanto en castellano como en inglés, estos son algunos de los términos buscados: *Open Access*, movimiento *Open Access, Budapest Open Access Initiative*, *Berlin Declaration*…

Para definir el concepto de dato de la investigación y todos los conceptos relacionados se consultan diferentes bases de datos y repositorios (bases de datos del CSI, *Web of Knowledge*, *Library and Information Science Abstracts*, el repositorio E-Lis…), también se utilizan los motores de búsqueda, plataformas como Recolecta y directorios como el DOAJ.

También se utiliza la bibliografía sobre el tema de la gestión de los datos de la investigación realizada por las tutoras de la tesina.

Para estudiar el *Digital Curation Centre* (DCC) y el *Australian National Data Service* (ANDS) se accede directamente a sus páginas webs y se utilizan los motores de búsqueda para encontrar estudios y proyectos relacionados con dichas organizaciones.

2. - Introducción:

2.1 – Movimiento OA

La comunicación científica sufrió un cambio en la última década del Siglo XX debido a la crisis del sistema tradicional de comunicación científica. Los motivos fueron:

- La escalada de fusiones y adquisiciones de empresas editoriales, lo que provocó que las empresas más pequeñas desapareciesen en manos de las más grandes, formando un mercado sin competencia.
- El incremento sostenido de los precios de las revistas científicas, denominado en la literatura especializada como 'crisis de las revistas' (*serial crisis*), en contraposición con el crecimiento nulo o el decrecimiento, en otros casos, de los presupuestos de las bibliotecas para adquirirlas.
- El aumento en las restricciones establecidas en las legislaciones sobre derechos de autor en lo relativo al acceso y diseminación de la información científica.
- Los problemas derivados del sistema de recompensa científica, enfocado más a la publicación en revistas "de impacto" que a la amplia difusión de los resultados científicos.

Estos factores provocaban que no se cumplieran los objetivos primarios de la comunicación científica, es decir, favorecer la diseminación y el intercambio de los resultados científicos para lograr la fertilización de la ciencia y el progreso científico-técnico y social de la humanidad.

Frente a esta situación, y con la intención de afrontar los diferentes problemas de la comunicación científica nace el movimiento *Open Access*, OA, (1997) que promueve la libre disponibilidad pública en Internet de los documentos de investigación científica, permitiendo a cualquier usuario la lectura, descarga, copia, distribución, impresión, búsqueda, o el vínculo a los textos completos de dichos artículos, la única restricción es dar a los autores control sobre la integridad de su trabajo y el derecho a ser reconocidos y citados.

El OA elimina las barreras de acceso a la literatura científica en Internet, tanto aquellas relacionadas con el precio como aquellas relativas a permisos y licencias.

El OA tiene sus orígenes o primeras propuestas en 1997 con la creación de la coalición SPARC, pero es en la declaración de *Budapest Open Access Initiantive*[1] (BOAI) en 2002, cuando ésta alcanza su pleno desarrollo, con la definición de la iniciativa OA "tendente a promover el acceso libre y gratuito a las publicaciones y que los autores conserven sus derechos de autor".

La iniciativa *Open Access* (OA) se perfiló mediante tres declaraciones realizadas en un período de dos años:

• *Budapest Open Access Initiative* (2002) : <http://www.soros.org/openaccess/index.shtml>.

• *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* (2003) [2] : <http://www.zim.mpg.de/openaccess-berlin>.

• *Bethesda Statement on Open Access Publishing* (2003)[3] : <http://www.earlham.edu/~peters/fos/bethesda.htm>.

También hay que destacar las recomendaciones de la Organización para la Cooperación y el Desarrollo Económico (OCDE) en 2004, que promovían el acceso abierto a los resultados de la investigación financiada con fondos públicos, o las realizadas por los *National Institutes of Health* (NIH) norteamericanos (2004), en las que se instaba a que cualquier investigación realizada con su financiación, debía ser publicada 6 meses después en *PubMed Central*.

[1] *Budapest Open Access Initiative* (BOAI) de finales de 2001; Disponible en: <http://www.budapestopenaccessinitiative.org/read >

[2] *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities;* Disponible en: <http://oa.mpg.de/openaccess-berlin/berlindeclaration.html>

[3] Texto de la *Bethesda Statement on Open Access Publishing;* Disponible en: <http://www.earlham.edu/~peters/fos/bethesda.htm>

La necesidad de desarrollar unas pautas de acción que sirvieran de guía a los responsables de las políticas de investigación, a las instituciones científicas y a todos los agentes culturales fue uno de los objetivos fundamentales de la Declaración de Berlín.

Internet y los nuevos medios de distribución de conocimientos como son webs, blogs, foros etc. suponen un punto de inflexión en la manera tradicional de difundir y distribuir el conocimiento y las publicaciones científicas, de ahí la necesidad de optar por nuevas técnicas en la edición y la publicación de los conocimientos.

La declaración de Berlín establece que las contribuciones basadas en el "principio de acceso abierto" deben satisfacer dos condiciones:

1. Los autores y los depositarios de la propiedad intelectual de las publicaciones deben garantizan a todos los usuarios el derecho de acceso gratuito, irrevocable y mundial, y el permiso para copiar, usar, distribuir, transmitir y exhibir la publicación para cualquier propósito responsable, todo ello sujeto al reconocimiento apropiado de autoría (los estándares de la comunidad científica continuarán proveyendo los mecanismos para hacer cumplir el reconocimiento apropiado y el uso responsable de las obras publicadas, como ahora se hace), lo mismo que el derecho de efectuar copias impresas en pequeño número para su uso personal.

2. Una versión completa de la publicación y de todos sus materiales complementarios, que debe incluir una copia de la autorización arriba mencionada, deberá de ser depositada, usando un formato electrónico estándar en, por lo menos, un depósito online que utilice estándares técnicos aceptables. El depósito debe ser gestionado y mantenido por una institución, una sociedad científica, una institución pública u otra organización bien establecida que busque implantar el "principio de acceso abierto", además se debe garantizar la distribución, la interoperabilidad y la capacidad archivística a largo plazo

Siguiendo con el tema de la publicación, en la *Budapest Open Access Initiative,* se establecieron dos rutas para alcanzar el OA que han ido desarrollándose desde entonces:

- **Vía dorada (revistas de acceso abierto):**

Se centra en la edición. Los autores publican en revistas OA, de modo que sus artículos son inmediatamente accesibles. En palabras de Remedios Melero y José Manuel Barrueco (2005), estas revistas oscilan entre dos extremos: "El caso más

puro sería el de aquella revista en la que ni el lector ni el autor pagan por publicar y son los autores los que retienen el copyright sobre sus trabajos, cediendo los derechos no exclusivos de publicación a la revista". En el otro extremo, "existen casos de revistas totalmente OA en las que el autor/institución paga por su publicación, como son las revistas de *BioMed Central* o *Public Library of Science*".

Para concretar, y dependiendo de la forma en la que se gestiona el coste de su publicación, podemos dividirlas en distintos grupos:

1. El caso más puro de acceso abierto serían las revistas en las que ni el lector paga por acceder ni el autor paga por publicar y, además, los autores mantienen el copyright sobre su trabajo, cediendo sus artículos, sin exclusividad, a la revista. Estas revistas suelen pertenecer a instituciones académicas o sociedades profesionales que son las que asumen su coste y mantenimiento, la mayoría de estas revistas se recogen en el DOAJ.

2. Revistas incluidas en plataformas o portales de acceso abierto con financiación pública. Esto ha sido posible gracias a la promoción que algunos países han empezado a hacer de esta forma de publicación. Este es el caso de *Scientific Electronic Library Online* (SciELO) que es un modelo para la publicación electrónica cooperativa de revistas científicas en Internet. Este modelo proporciona una solución eficiente para asegurar la visibilidad y el acceso universal a su literatura científica, especialmente en países en desarrollo.

3. Revistas con una adhesión explícita al movimiento de acceso abierto cuyo coste de publicación (costes de revisión, edición y difusión) paga el autor o la institución a la que pertenece.

4. Muchas empresas editoriales han creado un híbrido entre acceso abierto y tradicional, dando al autor la posibilidad de elegir cómo hacerlo. Este es el caso de *Springer Choice*, en el cual si el autor paga 3.000 euros aproximadamente puede publicar en abierto, o el de la editorial *Blackwell* con su *online open*. Los precios varían desde los 500 a los 3.500 euros.

Algunas editoriales y asociaciones profesionales también siguen la política de dejar sus publicaciones en libre acceso transcurridos 6, 12 ó 36 meses desde su publicación, a esto se le conoce como tiempo de embargo. Este tipo de publicación en libre acceso no llega a ser acceso abierto propiamente dicho, ya que aunque supone su disponibilidad en la Web gratis, la exclusividad de sus derechos de copyright la sigue manteniendo la editorial y no el autor (que generalmente ya ha cedido sus derechos en exclusividad a la editorial con la aceptación de su artículo para publicar).

- **Repositorios (vía verde):**

La vía verde se centra en el archivado. Los autores publican en revistas convencionales pero, además, permiten que sus artículos y sus datos estén libremente disponibles en la Red mediante la publicación en repositorios.

En estos repositorios se depositan copias de los artículos ya sea antes de su publicación y/o revisión (*preprints*) o después (*post-print*).

El repositorio, como ya se ha comentado, es un archivo digital de productos intelectuales gestionado por un organismo o institución y accesible a los usuarios finales. Es un archivo de su propio patrimonio investigador que tiene como objetivo poner a disposición de la sociedad y del resto de investigadores su producción científica para su beneficio mutuo. Además, como archivo abierto no debe ser tan sólo un depósito, sino que debe mantener una política preestablecida que regule cómo debe hacerse y en qué condiciones.

Para facilitar la interoperabilidad de los repositorios se utiliza el protocolo OAI-PMH (*Open Archives Initiative - Protocol of Metadata Haresting*). Este protocolo utiliza para los metadatos el esquema *Dublin Core* desarrollado por la *Online Computer Library Center* (OCLC) para describir cualquier objeto en la Web. Se fundamenta en dos tipos de servidores: *Data provider* (que tiene los documentos y metadatos) y el *Service Provider* (que recolecta los metadatos y ofrece opciones de búsqueda), es decir, ofrece a los usuarios servicios de valor añadido a partir de los metadatos que recolecta.

El uso de este mismo protocolo facilita el acceso a esta información desde múltiples puntos de búsqueda.

Los softwares libres más utilizados para la creación de repositorios son Eprints y Dspace. El proceso de autoarchivado se inicia con el registro del autor en el propio repositorio, donde se le otorga un espacio propio para que coloque sus documentos y sus datos.

El autoarchivado es un proceso muy sencillo, pero el autor debe conocer cuál es la situación de los derechos de autor en cuanto a su obra. Los repositorios facilitan a los autores una serie de servicios como son los datos estadísticos, que les facilita el número de consultas y descargas de su obra, qué países han consultado sus documentos, etc.

Además, producen un crecimiento exponencial de la visibilidad, al ser indizados por buscadores como Google y por recolectores de metadatos con protocolo OAI.

Los primeros repositorios fueron temáticos como el caso de *Pubmed Central*, desarrollado por *U. S. National Institutes of Health*, que archivaba artículos de

revistas con la participación de los editores. Más tarde surgieron los repositorios institucionales, que recogen la producción científica de su institución para hacerla visible.

Los repositorios institucionales están siendo promovidos mediante políticas institucionales que obligan a sus autores al autoarchivado, por lo que cada universidad y centro de investigación está creando su propio repositorio.

Existen registros internacionales de repositorios como el *Registry of Open Access Repository* (ROAR) elaborado por la Universidad de Southamptom (Reino Unido) o el *Directory of Open Access Repositories* (OpenDOAR), también hay un registro de las políticas institucionales clasificadas por países, *Registry of Open Access Repository Material Archiving Policies* (ROARMAP). Según este registro existen actualmente más de un millar de repositorios en el mundo.

En cualquier caso, para Peter Suber, profesor en la Universidad de Stanford y autor de algunos de los textos más citados sobre el estudio del acceso abierto a la ciencia, "la vía verde y la vía dorada son complementarias. Las revistas OA proporcionan *peer review*, los archivos OA no. Los archivos OA proporcionan difusión instantánea de nuevos descubrimientos, las revistas OA no. Los archivos OA también tienden a proporcionar preservación a largo plazo junto a su función de mejora del acceso".

Pero la investigación científica es algo más que las publicaciones, cuando se realiza una investigación los datos que finalmente se publican son tan solo una pequeña parte de todos los datos recolectados por los científicos durante el proceso de investigación.

Es importante que los datos de la investigación estén disponibles y accesibles en la red, de la misma manera que ocurre con las publicaciones científicas.

Borgman (2007) ya apuntó que los datos científicos estaban ganando mucho valor y empezaban a ser valorados como un producto final.

En Febrero de 2007, la Comisión Europea publicó una comunicación sobre la información científica en la era digital, en la que señalaba la importancia de poner en marcha una política referente al acceso, la difusión y la preservación de la información científica en toda la Unión Europea, tanto en lo referente a las publicaciones como a en lo referente a los datos fruto de la investigación.

Ese mismo año la Organización para la Cooperación y el Desarrollo Económico (OCDE) publicó una guía[4] en la que incluía recomendaciones generales para el acceso a los datos de información científica procedentes de la financiación pública.

En esa guía la OCDE defendía que el acceso efectivo a los datos de la investigación es una condición indispensable para poder aprovechar al máximo las nuevas oportunidades y los beneficios que ofrecen las tecnologías de la información con el objetivo de:

- Favorecer la buena gestión de la inversión pública
- Crear fuertes cadenas de valor en el ámbito de la innovación
- Aumentar la cooperación internacional
- Mejorar el acceso y el intercambio de los datos
- Reforzar la investigación científica abierta
- Alentar la diversidad de análisis y opinión
- Promover nuevas investigaciones
- Permitir la creación de nuevos conjuntos de datos mediante la combinación de datos de varias fuentes
- Facilitar la formación de nuevos investigadores

En el 2009, el informe del *National Research Council* también recogía premisas sobre la importancia de la protección, la accesibilidad y la custodia de los datos de la investigación.

Pero no hay que irse tantos años atrás para encontrar documentos y comunicaciones europeas referentes al acceso a la información científica, desde el 2010 el "*High level expert group on scientific data",* a petición de la Comisión Europea, está elaborando un informe[5] con su visión sobre el uso, el acceso, la calidad… de los datos de investigación científica.

[4] *Principles and Guidelines for Access to Research Data from Public Funding*, OECD, 2007; Disponible en: <http://www.oecd.org/sti/sci-tech/38500813.pdf>

[5] *High level expert group on scientific data: Riding the Wave: How Europe can gain from the rising tide of scientific data*; *European Union*, 2010; Disponible en: <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf >

Es un hecho que la sostenibilidad de estos datos interesa a las universidades, a las agencias de financiación de la investigación, a los gobiernos, a la Comisión Europea…

Por estas razones a la comunidad científica le surge la pregunta: ¿qué hacemos, cómo gestionamos esa cantidad ingente de datos fruto de la investigación?

2.2 - Dato de investigación

En primer lugar y antes de contestar a esa pregunta hay que saber que es lo que se entiende por dato de investigación, la Organización para la Cooperación Económica y el Desarrollo (OECD) considera que son datos de la investigación todos los materiales que hayan sido registrados durante la investigación, reconocidos por la comunidad científica y que sirven para certificar los resultados de la investigación que se realiza.

Por su parte, el *National Science Board* (NSBD) define los datos de la investigación de la siguiente manera:

"Cualquier información que se puede almacenar en formato digital, incluyendo texto, números, imágenes, vídeo, audio, software, algoritmos, ecuaciones, animaciones, modelos, simulaciones, etc. Estos datos se podrán obtener por diversos medios, incluyendo la observación, el cálculo y la experimentación"

Estos datos pueden tener distintos formatos y tipologías, pueden ser considerados como datos de la investigación:

- Datos numéricos
- Resultados fruto de la medida de instrumentos
- Datos de encuestas
- Imágenes digitales
- Audios digitales
- Vídeos digitales
- Documentación diversa como notas de investigación de campo, bases de datos de modelos genéticos, descripciones, informes...

Los datos de la investigación son considerados una fuente de conocimiento propia e independiente de las publicaciones, pueden ser utilizados para validar resultados de investigaciones publicadas o pueden ser re-utilizados para generar nuevo conocimiento.

Pero tan importante es saber que es lo que se considera "dato de la investigación", como aquello que no, por eso los *National Institutes of Health* (NIH) de Estados

Unidos han establecido una serie de elementos que no son considerados datos finales de investigación:

- Notas de laboratorio
- Sets de datos parciales
- Análisis preliminares
- Borradores de trabajos
- Planes para investigaciones futuras
- Informes que han tenido un proceso de revisión por pares
- Comunicaciones con colegas
- Objetos físicos
- Ejemplares de laboratorio

La *National Science Foundation* (2007) por su parte, categoriza los datos de investigación en tres grupos basándose en su origen:

- **Datos observacionales:** son registros históricos e irrepetibles que se obtuvieron en un lugar y en un momento concreto en el tiempo.

- **Datos experimentales:** son los datos que acompañan a los experimentos desde su planificación y preparación hasta la obtención de resultados. Los experimentos pueden repetirse y conseguir los mismos datos, pero el coste de repetir el experimento resta rentabilidad a la operación.

- **Datos computacionales:** son aquellos datos que acompañan a las simulaciones que suelen incluir datos de entrada, programas y resultados.

El tema de la necesidad de conservar los datos de la investigación está tan en auge hoy en día debido a que en los últimos tiempos se ha producido un aumento exponencial del volumen de datos valiosos y complejos por diversos motivos:

- Avance de las tecnologías de la comunicación e información.

- Proliferación de instrumentos científicos más potentes.

- Migración de los espacios físicos de trabajo hacia los espacios virtuales.

- Aceptación del movimiento *open access* por parte de científicos e investigadores.

Estos avances han revolucionado todas las actividades cotidianas, tanto las económicas, las culturas, las sociales… pero especialmente las relacionadas con el

ámbito científico y académico, y han provocado que aumente la necesidad de preservar, gestionar, acceder y conservar grandes volúmenes de datos.

Se han abierto nuevas formas y vías para utilizar las enormes masas de datos resultantes de los experimentos y observaciones de los procesos científicos.

Estos datos no deben de perderse, la falta de gestión de los datos es muy peligrosa, ya que puede provocar la pérdida de grandes cantidades de información a largo plazo.

Algunos datos de la investigación son únicos y no pueden ser reemplazados si se destruyen o se pierden.

Si el acceso tanto a las publicaciones científicas como a los datos de la investigación es sencillo y eficaz se puede acelerar la innovación y la investigación, y evitar la duplicación de los esfuerzos de investigación.

Conservar los datos de la investigación es esencial también para garantizar la trazabilidad y la repetibilidad de los experimentos.

Muchas instituciones e investigadores están obligados por ley a gestionar y conservar los datos de sus investigaciones.

Por todo esto surge la pregunta: ¿hay que cambiar el modo en que se hace la ciencia?

2.3 – E-ciencia, e-infraestructuras, data sharing.

Todos estos avances en las tecnologías de la información han afectado al modo en que se hace ciencia.

El concepto de Ciencia 2.0 o e-Ciencia (Shneiderman, 2008), está estrechamente relacionado con lo planteado anteriormente, Shneiderman habla de que la metodología de la ciencia debe de cambiar, ya no basta con que los procesos científicos sean controlados únicamente en condiciones de laboratorio, la actividad científica tiene la necesidad de ser cooperativa, el intercambio libre de conocimiento no puede limitarse exclusivamente al intercambio de los resultados finales en forma de artículos, dejando de lado los datos de la investigación o los detalles clave de los procedimientos, hay que crear infraestructuras científicas que permitan el acceso a bancos de datos muy voluminosos a través de Internet y de forma colaborativa, crear lo que se conoce como e-infraestructuras.

Estas infraestructuras tendrán como misión permitir el acceso, el uso y la reutilización de los datos, siempre respetando su autoría e integridad.

Para Martinez-Uribe y Macdonald (2008), la e-ciencia se produce "cuando la investigación multidisciplinar y en colaboración tiene lugar en diversas localizaciones, produciendo y utilizando grandes cantidades de datos".

A esa compartición de ficheros de datos generados durante el curso de una investigación con el resto de la comunidad académica o científica que fomenta la e-ciencia se le conoce como *data sharing.*

El concepto aunque está muy relacionado con el de Ciencia 2.0 o e-Ciencia no es nuevo, Galton en 1901 afirmaba que:

"Nadie debiera publicar resultados biométricos sin depositar una copia de sus datos bien redactada y presentada en algún lugar donde todo aquel que lo deseara pudiera verificar su trabajo"

El término que se suele utilizar actualmente para definir la gestión activa y prolongada de los datos científicos es "*data curation*" o "*digital curation*".

Este concepto hace referencia a "la labor de gestionar y promocionar el uso de datos desde el momento de su creación para asegurar su uso contemporáneo y su disponibilidad para ser localizados".

El *Digital Curation Center* (DCC), organismo británico que gestiona y promueve el uso de los datos desde su creación, define el concepto de *Data Curation* como aquella actividad centrada en el mantenimiento, la preservación y la adjudicación de valor añadido a los datos de la investigación durante su ciclo de vida.

“Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle”.

Ross Harvey (2010), amplía la definición dada por el *Digital Curation Center*:

"El Data Curation se encarga de la gestión activa de los datos durante el tiempo que siguen teniendo interés académico, científico, administrativo y personal, con el objetivo de favorecer su reproducción, su reutilización y agregándoles valor, los datos se gestionan desde su creación hasta que se determina que ya no son útiles, garantizando su accesibilidad a largo plazo, su conservación, su autenticidad y su integridad."

Es decir, el *Data Curation* se encarga de todos los procesos relativos a los datos desde su nacimiento hasta su almacenamiento.

Existe otro término muy relacionado con la corriente del *data curation*, el *sheer curation*, que aboga por integrar la gestión activa y prolongada de los datos científicos dentro del flujo normal de trabajo de los que crean y manejan datos dentro de otras aplicaciones, sin que éstos la noten.

Se basa en la hipótesis de que para la correcta gestión de los datos digitales es mejor comenzar a integrar la preservación cuando éstos datos están siendo creados, lo cual derivará más tarde en una buena práctica para compartir, publicar y/o preservar esos datos dentro del entorno digital, produciendo beneficio a largo plazo.

Además de unas infraestructuras adecuadas y una política clara entre los investigadores acerca de cómo conservar los datos de sus investigaciones, se necesita que tanto los científicos como los investigadores cambien su actitud, que se abran a la compartición, esto ayudará a que se puedan reunir gran cantidad de datos, ya sean datos utilizados por los propios investigadores en sus investigaciones como datos ajenos resultantes de otras investigaciones, de esta manera se puede conseguir que se vertebre un entorno colaborativo, en el que los datos se reutilizan y se combinan, incrementando la productividad y generando nuevo conocimiento a partir de ellos.

Movimientos como el OA ponen de manifiesto que hay muchos científicos con una mentalidad mucho más abierta en lo relativo a compartir sus hallazgos y sus investigaciones.

Compartir los datos proporciona a los autores de los mismos una audiencia mucho mayor, provoca que la difusión internacional sea superior y que se puedan alcanzar estratos sociales más amplios, acercándose de este modo a uno de los objetivos subyacentes de la ciencia.

Llegar a audiencias mucho más amplias aumenta la probabilidad de que los trabajos y los datos sean más leídos, citados y tengan más impacto.

Además el autor adopta un papel fundamental en el autoarchivo y en la autogestión de sus derechos de propiedad intelectual.

Pero la compartición de los datos y de las investigaciones no es algo que beneficie exclusivamente a los autores o a los investigadores, las entidades financiadoras también consiguen beneficios, ya que se posibilita la reutilización y un mayor uso y explotación de los resultados de la investigación por parte de un mayor número de usuarios.

Supone también un beneficio para las instituciones académicas ya que su uso y reutilización genera publicidad para la investigación desarrollada en la institución.

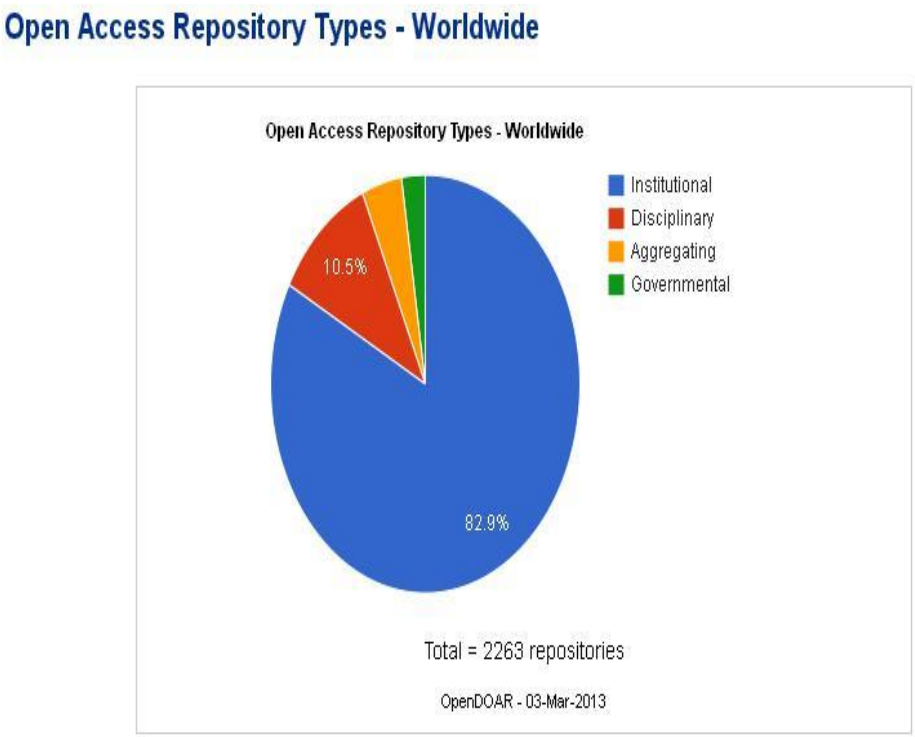
Pero, ¿se está realizando correctamente esta labor?

Para contestar a esa pregunta vamos a estudiar qué presencia tienen los datos de la investigación en los repositorios institucionales, para ello nos vamos a basar en los datos que ofrece el directorio internacional de repositorios académicos Opendoar

2.4 - Repositorios institucionales y contenidos.

Teniendo en cuenta los datos extraídos del prestigioso directorio internacional de repositorios académicos de acceso abierto Opendoar, gestionado por la Universidad de Nottingham del Reino Unido, a marzo del 2013 hay 2200 repositorios registrados, de los que 1875, casi un 83%, son institucionales.

Figura 1. Tipología de los repositorios registrados



Fuente: OpenDOAR

Los repositorios institucionales eran en sus orígenes bibliotecas digitales basadas en materiales de producción científica e institucional, pero esta tipología se ha ido ampliando a lo largo de los años para dar cabida a materiales diversos entre los que se encuentran los datos de investigación.

Los *datasets* son colecciones de datos compuestos y heterogéneos que se reúnen durante la ejecución de un proyecto, constituyen la base de una investigación y van asociados a una publicación científica.

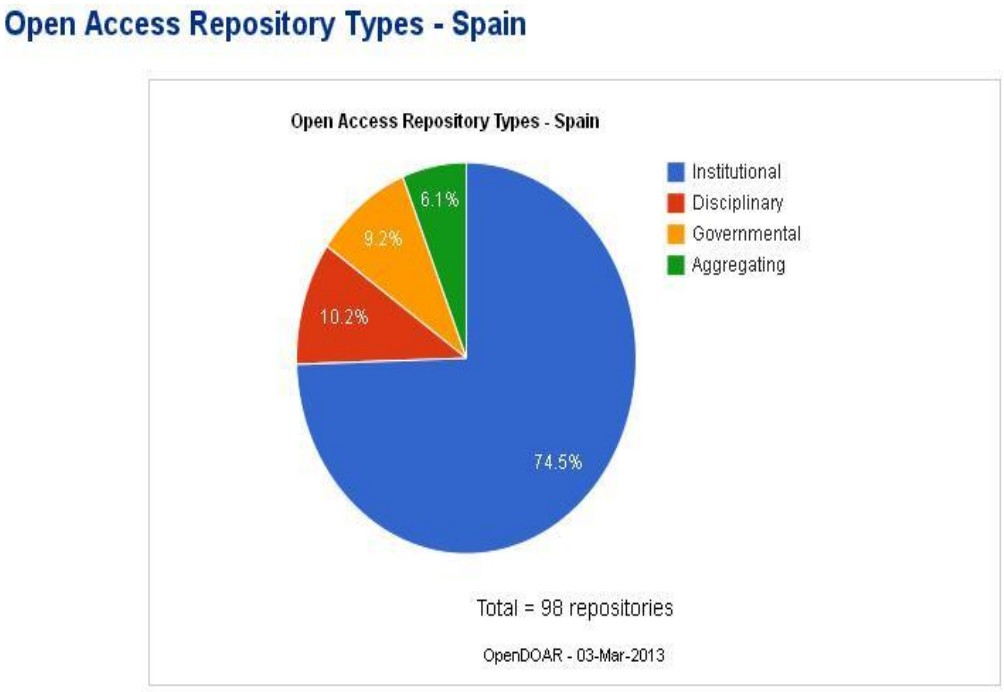
Los datos de investigación son uno de los tipos de información digital más complicados de gestionar, debido a que tanto los productores que los generan,

como los tipos de datos existentes, son muy variados, e incluyen objetos muy complejos.

Debido a su importancia, los repositorios institucionales se están convirtiendo, cada vez más, en herramientas esenciales para la comunicación académica en la era digital.

Los repositorios permiten el acceso abierto a los resultados de la actividad científica y académica de las distintas instituciones. Son importantes para las instituciones ya que ayudan a desarrollar estrategias para la captura, identificación, almacenamiento, conservación y recuperación de sus contenidos digitales.

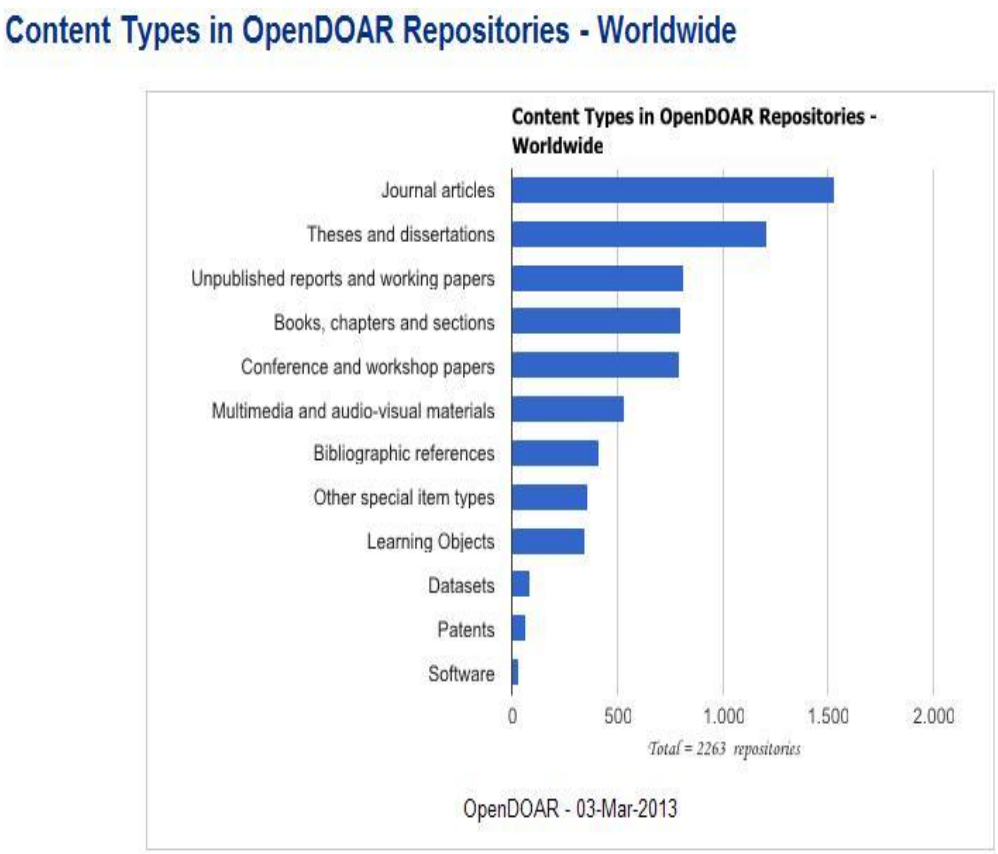
Figura 2. Tipología de los repositorios registrados en España



Fuente: OpenDOAR

En España el tanto por ciento de repositorios institucionales es algo inferior, pero se mantiene la tendencia internacional, de los 98 repositorios registrados, 73, un 74,5%, son institucionales

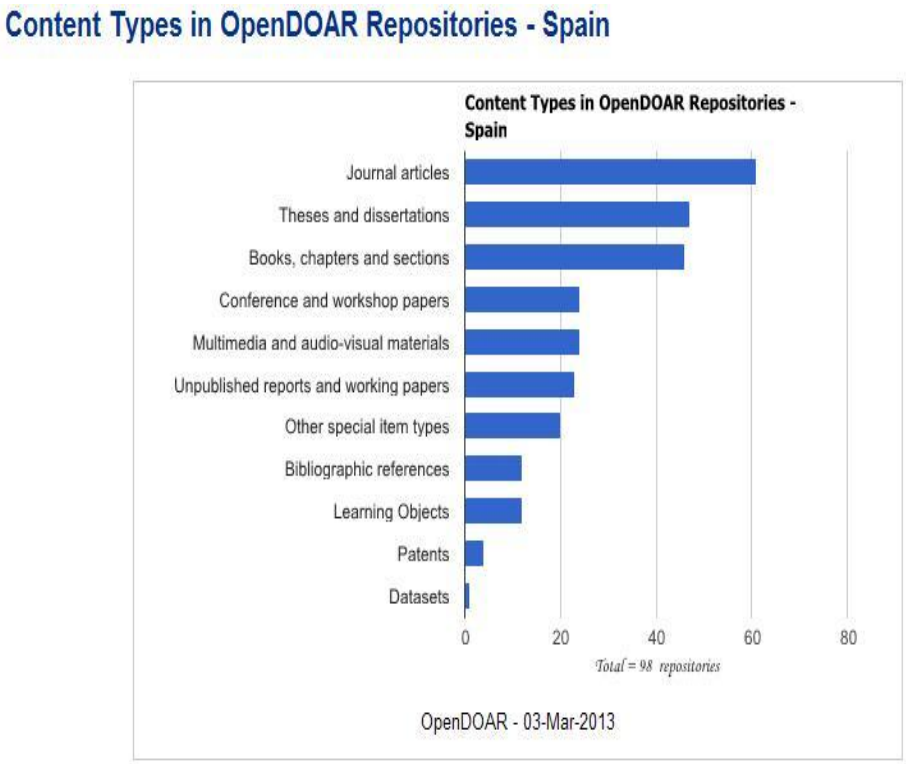
Figura 3. Tipos de contenido en los repositorios



Fuente: OpenDOAR

Como se puede observar en la figura 3, existe una clara tendencia hacia el predominio de los objetos digitales de tipo publicación, como son los artículos de revista, las tesis, los informes de proyectos, etc., en detrimento de otros formatos menos comunes como los datos de investigación, los materiales multimedia, los programas de ordenador, las patentes…

Figura 4. Tipos de contenido en los repositorios de España



Fuente: OpenDOAR

La situación de los tipos de contenido en los repositorios de España no difiere de lo observado a nivel mundial. Se puede ver una fuerte presencia de contenidos digitales tales como artículos y tesis doctorales, *working papers*, presentaciones a congresos y monografías… mientras que la presencia de los datos de investigación, *datasets*, es escasa.

Pero como hemos comentado en puntos anteriores, los proyectos, las iniciativas oficiales, los documentos y las comunicaciones referentes a la gestión de los datos de investigación han ido en aumento en los últimos años, además, la incorporación de los datos de investigación en infraestructuras científicas de acceso abierto es cada vez mayor.

Incluso se ha producido un cambio en los criterios tradicionales para la preservación a largo plazo, según el *Digital Preservation Coalition* (DPC), ya no se valoran tanto los criterios de valor, pertinencia y/o uso, hoy en día se tienen más en cuenta otros criterios, tales como:

- Que el formato del recurso sea legible actualmente y en un futuro
- Que el recurso esté en un soporte gestionable para su transferencia y/o almacenamiento
- Que la institución tenga pleno derecho a manipular los datos para asegurar su acceso en entornos informáticos del futuro, no se puede preservar un recurso reproduciéndolo o reformateándolo si no se tiene el permiso del titular
- Que el recurso disponga de documentación, incluyendo los metadatos

Por esta razón es interesante estudiar proyectos que permitan conocer casos de éxito relacionados con la gestión de datos de investigación.

3. Estudio de casos

El acceso a largo plazo y de forma sostenible a los datos actuales mediante la preservación, es una tarea enorme que muchos organismos internacionales, nacionales e institucionales están tratando actualmente. Los organismos, instituciones y centros de investigación más relevantes se encuentran ubicados en el Reino Unido, Australia y Estados Unidos.

Por eso, en este punto vamos a describir dos organismos significativos como son el *Digital Curation Centre* (DCC) y el *Australian National Data Service* (ANDS), que reflejan lo que se está realizando en el campo de la gestión de datos de investigación.

3.1 - Digital Curation Centre (DCC)

El *Digital Curation Centre* (DCC) es un consorcio liderado por la universidad de Edimburgo y financiado por el *Joint Information Systems Committee* (JISC).

Empezó su actividad con una primera fase de objetivos en noviembre de 2003. Nació como un centro para la preservación digital, con el objetivo de crear, capacitar y dotar de herramientas para la gestión de datos de investigación a la comunidad científica de las instituciones de educación superior en el Reino Unido.

Es uno de los organismos más importantes en lo referente a la preservación de datos que existe en el Reino Unido, y a nivel internacional.

El DCC apoya a las instituciones del Reino Unido que se encargan de almacenar, gestionar y preservar datos digitales.

También trabaja con otros profesionales para asegurar la mejora continua y el uso a largo plazo de los datos digitales.

Para el DCC la conservación digital no consiste simplemente en mantener y conservar la información digital, si no que también hay que añadirle valor tanto para su uso actual, como para su uso en un futuro.

Los objetivos que se marca el DCC son:

- Conseguir el liderazgo estratégico en la preservación digital para la comunidad investigadora del Reino Unido, haciendo especial hincapié en los datos de la ciencia.

- Influir en las políticas nacionales e internacionales mediante la creación de planes de gestión de datos de investigación.

- Proporcionar apoyo, asesoramiento especializado y orientación a los profesionales y los organismos de financiación.
- Elevar el nivel de conocimiento y experiencia entre los creadores de datos, los conservadores y otros individuos con un papel de conservación, mediante la creación de recursos y programas de formación.
- Fortalecer las redes de conservación y la colaboración.
- Fomentar la investigación en este campo.
- Desarrollar recursos, software, herramientas y servicios de apoyo.
- Desarrollar su modelo de gestión documental basado en el ciclo de vida de los datos

Antes de hablar del modelo de gestión del material digital del DCC, hay que dejar claros unos conceptos, por ejemplo: ¿qué se entiende por preservación digital?

La preservación digital se define como un conjunto de procesos dirigidos a conservar la información en formato digital. No existe preservación digital si no se mantiene la posibilidad de acceder a los recursos digitales. El objetivo de la preservación digital es permitir a los futuros usuarios recuperar, acceder, descifrar, ver, interpretar, entender y experimentar documentos y datos de forma significativa y válida (J. Rothenberg, 1995).

El material digital se enfrenta a diferentes amenazas:

- Soportes frágiles
- Rápida obsolescencia de los equipos y programas informáticos, debido a que los cambios en las tecnologías son frecuentes
- Incertidumbre en torno a los recursos, la responsabilidad y los métodos para su mantenimiento y conservación
- Falta de legislación que proteja estos procesos.
- Barreras al acceso: claves, cifrado, acceso restringido
- Descripciones inadecuadas que afectan a su recuperación
- Pérdida de información sobre el contexto

Neil Beagrie (2004) afirmó que la información digital nunca sobrevivirá accidentalmente.

La necesidad de realizar un modelo de conservación del material digital basado en el ciclo de vida de los datos fue tratado por Pennock (2007)[6].

El material digital, por su propia naturaleza, es inestable y es susceptible a los cambios tecnológicos desde el momento de su creación.

La curación y la preservación son actividades que si se realizan correctamente, en las diferentes etapas del ciclo de vida de los datos, pueden influir positivamente en la capacidad para cuidar de ellos con éxito en etapas posteriores.

Basarse en el ciclo de vida de los datos para la conservación del material digital asegura que todas las etapas de la preservación sean identificadas.

A estas etapas se les asigna una planificación y una serie de acciones relacionadas, en la secuencia correcta.

Esto puede ayudar a garantizar la autenticidad, la fiabilidad, la integridad y la facilidad de uso del material digital, además de asegurar que se maximiza la inversión empleada en su creación.

Muchos investigadores han tratado el enfoque del ciclo de vida para la gestión de activos digitales, y han ido desarrollado modelos de ciclo de vida específicos.

Sarah Higgins (2007), miembro del DCC, desarrolló un modelo de ciclo de vida que trataba específicamente las necesidades relacionadas con la "*digital curation*".

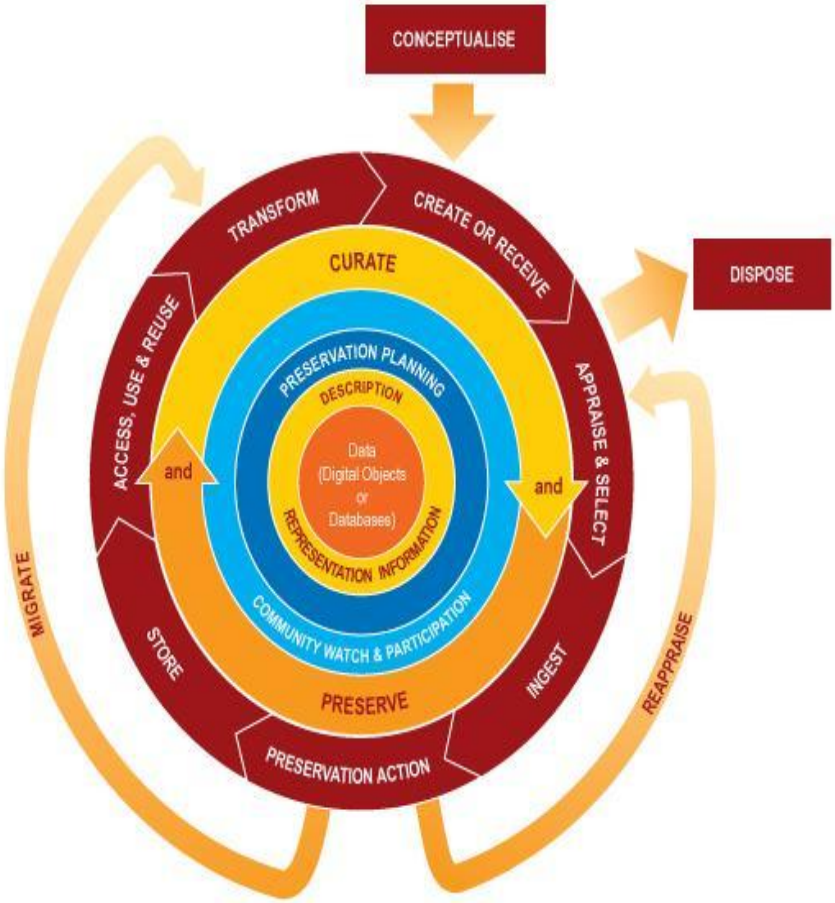
El proyecto *"DCC Curation Lifecycle Model"*, junto con una convocatoria abierta para comentarios, fue publicado en la "*International Journal of Digital Curation"* en Diciembre de 2007, y se hizo publico en el "*3rd International Digital Curation Conference*[7]*"*.

El DCC establece un modelo de gestión documental basado en el ciclo de vida de los datos, el objetivo principal es conseguir una correcta gestión de los datos a fin de garantizar su accesibilidad y su disponibilidad a lo largo del tiempo.

[6] Pennock, M. (2007). *Digital curation: A life-cycle approach to managing and preserving usable digital information*. Library and Archives Journal, Issue 1.Retrieved (preprint) June 18, 2008; Disponible: http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/libarch_curation.pdf

[7] *3rd International Digital Curation Conference*; Disponible en: <http://www.dcc.ac.uk/events/dcc-2007>

Figura 5. DCC Curation Lifecycle Model



Fuente: Digital Curation Centre

El modelo proporciona una visión general de las etapas requeridas para realizar una gestión activa y prolongada de los datos científicos

Es importante tener en cuenta que el modelo es un ideal que engloba todas las fases que debería tener un modelo de gestión documental para conseguir resultados positivos, pero, los usuarios del modelo, pueden empezar su actividad en cualquier etapa del ciclo de vida dependiendo de sus necesidades.

El modelo expuesto por el DCC permite definir roles y responsabilidades y construir un marco normativo.

También se puede utilizar para ayudar a las organizaciones a identificar los riesgos que corren sus activos digitales, proponiendo medidas adicionales y estrategias de gestión exitosas.

- **Elementos clave del ciclo de vida del DCC:**

El ciclo de los datos comienza con la descripción y la representación de la información.

Los datos de la investigación pueden adquirir gran variedad de formatos:

- **Objetos digitales:** objetos digitales simples (archivos de texto, archivos de imágenes o archivos de sonido, junto con sus identificadores y metadatos relacionados) u objetos digitales complejos (objetos digitales realizados mediante la combinación de diferentes objetos digitales).
- **Bases de datos:** colecciones estructuradas de registros o datos almacenados en un sistema informático.

En esta primera etapa será necesario colaborar muy de cerca con los creadores de los datos para comprenderlos bien y así poder conservarlos y preservarlos correctamente, las comunidades deben proporcionar a los investigadores un modelo de gestión de los datos para ahorrar tiempo y esfuerzo en el proceso de la investigación.

Por su parte, los investigadores deben proporcionar la información contextual necesaria para determinar el origen y el ciclo de vida de los datos.

- **Acciones a realizar:**

El modelo contiene distintas acciones de curación:

- Acciones aplicables a todo el ciclo de vida de los datos
- Acciones que pueden llevarse a cabo de forma secuencial
- Acciones ocasionales que se realizarán cuando las circunstancias lo exijan.

Acciones aplicables a todo el ciclo de vida de los datos	
Descripción y representación de la información	Asignar metadatos administrativos, descriptivos, técnicos, estructurales y de preservación, utilizando las normas apropiadas, para asegurar una descripción adecuada y el control a largo plazo. Recoger y asignar la información necesaria para comprender el material digital y los correspondientes metadatos.
Plan de preservación	Consiste en la elaboración de un plan de mantenimiento para todo el ciclo de curación, incluye la creación de planes para la gestión de datos, y para la correcta administración de todas las acciones del ciclo de vida.
Programas de vigilancia y participación.	Estar atento a las actividades de las comunidades de interés para la institución y realizar un seguimiento de las investigaciones relacionadas. Participar en el desarrollo de estándares comunitarios, herramientas y softwares.
Curación y preservación	Encargase de la gestión y de la administración de todas las acciones previstas para promover la curación y preservación durante todo el ciclo de vida.

La asignación de metadatos a los objetos digitales es un punto fundamental en este sistema de preservación digital ya que posibilita su gestión y preservación a largo plazo, su descripción, su diseminación y su recuperación.

Los metadatos que se asignen servirán después para poder localizar los objetos digitales depositados en los repositorios.

Los metadatos se suelen clasificar en:

- **Metadatos descriptivos:** describen e identifican los datos para permitir su búsqueda y su recuperación.
- **Metadatos estructurales:** facilitan la navegación y la presentación de los recursos electrónicos**,** proporcionan información sobre la estructura interna de los recursos, además, describen la relación entre los materiales, unen los archivos y los textos relacionados.
- **Metadatos administrativos:** facilitan la gestión y el procesamiento de los recursos digitales tanto a corto como a largo plazo, incluyen datos técnicos sobre la creación y el control de calidad, sobre la gestión de derechos y los requisitos de control de acceso y utilización, además de incluir información sobre las acciones de preservación.

Para la creación de metadatos se recomienda la automatización del proceso para minimizar el error humano.

Hay distintos ejemplos, estos son algunos:

- **Metadata Extraction Tool:** desarrollado por la Biblioteca Nacional de Nueva Zelanda, es un programa *open-source* multiplataforma que permite extraer metadatos automáticamente de los objetos digitales conservando su integridad. Dispone de interface en línea de comandos y en formato gráfico.
- **JHOVE:** realiza las funciones de identificación, validación y caracterización de formatos, es ideal para generar metadatos técnicos.
- **Xena:** es un software de código libre y abierto desarrollado por el Archivo Nacional de Australia. Permite la detección del formato de archivo de los objetos digitales y la conversión de objetos digitales en formato abierto para su ulterior conservación. Como resultado de la normalización se obtienen ficheros XML con los metadatos descriptivos y el contenido binario del archivo.

Las acciones secuenciales son las relacionadas con la conceptualización, están ligadas al proceso de almacenamiento de datos.

Acciones secuenciales	
Crear o recibir	Hace referencia al proceso de generación o de recepción de datos, donde se inicia la actividad de *data curation.* Es necesario asegurarse de que los datos son recogidos en un formato adecuado, y de que son descritos con metadatos apropiados
Evaluar y seleccionar	A partir de la documentación y de las políticas previamente establecidas, se realiza una evaluación de los datos para seleccionar aquellos que van a ser preservados a largo plazo.
Desechar	Todos aquellos datos que no son seleccionados en concordancia con la política de selección se deben de desechar. Los datos pueden ser trasladados a otro archivo, depósito, centro de datos… o bien pueden ser destruidos.
Traspaso	Consiste en la transferencia de los datos de investigación a un archivo, depósito, centro de datos…

Acciones secuenciales	
Acción de preservación	Engloba todas las acciones que se deben realizar para asegurar la preservación a largo plazo de los datos y su autenticidad, fiabilidad, reutilización e integridad. Incluye acciones de corrección de errores, limpieza de datos, validación de formatos, asignación de metadatos de preservación, realización de migraciones, copias de seguridad…
Almacenar	Es el proceso de almacenar los datos de forma segura siguiendo las normas. Hay que conocer las políticas del repositorio para evitar que puedan afectar al almacenamiento de datos a largo plazo, hay que saber, por ejemplo, cuáles son los formatos más adecuados. Hay que hacer que el proceso de almacenamiento sea sencillo y proporcionar apoyo y orientación siempre que sea posible, además, es importante tratar de automatizar los procesos. Hay que decidir quién es responsable de la garantía de calidad de los datos en el punto de depósito: el investigador, el archivo, el gestor de información, etc.
Acceso, uso y reutilización	El objetivo de esta etapa es que los datos sean accesibles y localizables para los usuarios, se deben de incluir metadatos y realizar indexaciones. Se debe controlar el acceso a los datos, así como establecer licencias y permisos.

Acciones secuenciales	
Transformación	Consiste en la creación de nuevos datos a partir de los originales: -Por la migración a un formato diferente, para evitar la obsolescencia. -Por la creación de un subconjunto de datos, fruto de la selección o de la consulta, con el objetivo de crear nuevos resultados derivados. Mediante esta acción los datos vuelven a encontrarse al inicio de su ciclo de vida.

Acciones ocasionales	
Reevaluación	Consiste en la recuperación de datos debido a fallos en los procedimientos de validación para su posterior evaluación y re-selección.
Migración	Consiste en migrar los datos a tecnologías o formatos más nuevos para garantizar su preservación y evitar así la obsolescencia del soporte físico o del software. También se puede realizar la migración de los datos para conseguir una homogeneidad en el entorno de almacenamiento. *En algunos casos se debe preservar la manera en la cual los datos originales fueron creados y presentados para que el objeto digital siga siendo accesible y significativo.

- **Proyectos de los que forma parte el DCC:**

Para favorecer el desarrollo del centro y de sus servicios el DCC colabora con distintas entidades en diferentes proyectos.

1. **Proyectos actuales:**

- **Proyecto 4c:**

Se trata de un programa que trata de ayudar a las organizaciones europeas a que inviertan más eficazmente en la preservación y la curación digital.

Cuando una organización invierte en estos aspectos busca lograr un beneficio, por lo que debe de apostar por el "valor", la "calidad" y la "sostenibilidad".

El objetivo es que las organizaciones sean capaces de controlar y gestionar eficazmente sus activos digitales a través del tiempo, pero también se busca crear nuevos servicios y soluciones rentables para los demás.

El proyecto 4c es co-financiado por la Unión Europea dentro del 7º Programa Marco de acciones de investigación y desarrollo tecnológico, se inició el 01 de febrero de 2013 y se desarrollará hasta el 31 de enero de 2015.

El Proyecto 4C cuenta con 13 socios de 7 países diferentes:

- Jisc (UK) (Project Co-ordinator)
- Deutsche Nationalbibliothek (Germany)
- Digital Preservation Coalition (UK)
- INESC-ID – Institute for System and Computer Engineering (Portugal)
- Keep Solutions (Portugal)
- KB – the National Library of the Netherlands (Netherlands)
- KNAW-DANS – the Royal Dutch Academy of Research Data Archive and Network Service (Netherlands)
- National Library of Estonia (Estonia)
- Secure Business Austria (Austria)

- Statens Arkiver – the State Archive (Denmark)
- University of Essex (UK)
- University of Edinburgh (UK)
- HATII, University of Glasgow (UK)

- **DaMSSI-ABC:**

Tiene como objetivo principal apoyar y mejorar el desarrollo, la difusión y la reutilización de los materiales de formación relacionados con la gestión de datos de investigación desarrollados por los proyectos RDMTrain CSAC.

Para conseguirlo se clasifican y depositan los distintos materiales de formación en JORUM, un repositorio institucional que recoge materiales de aprendizaje y enseñanza, para ayudar a que sean más fácilmente detectables y reutilizables.

- **Dryad Reino Unido:**

Consiste en el desarrollo y evaluación de un repositorio[8] de datos de investigación para las ciencias de la vida.

- **DigCurV:**

Se trata de un proyecto financiado por la Comisión Europea que pretende establecer un marco curricular para formar a profesionales en la curación digital.

- **CARDIO:**

Se trata de una herramienta de referencia realizada para desarrollar estrategias de gestión de datos, por lo general, se aplica a nivel de departamento o grupo de investigación.

CARDIO permite:

- Evaluar las necesidades de gestión de datos, las actividades y las capacidades de la institución o departamento.

[8] Disponible en: http://datadryad.org/

- Crear un consenso entre los creadores de datos, los gestores de información y los proveedores de servicios.
- Mejorar las actuaciones a la hora de la gestión de datos.
- Identificar ineficiencias operacionales y oportunidades de ahorro de costes.

- **Marco de Investigación Inteligente (SRF):**

Pretende desarrollar una infraestructura virtual colaborativa que proporcione las herramientas necesarias para la gestión de datos de investigación en distintos servicios desarrollados por la Universidad de Southampton (LabTrove, Blog3 y LabBroker).

2. **Proyectos antiguos:**

- **KRDS/I2S2:**

El proyecto fue financiado entre febrero y julio de 2011.

El objetivo fundamental del proyecto era desarrollar herramientas para evaluar los beneficios que supone la preservación digital de los datos de la investigación.

- **Closing the Digital Curation Gap:**

Se trataba de un proyecto colaborativo a nivel internacional que tenía como objetivo elaborar guías claras y comprensibles sobre la curación digital que sirvieran como ejemplo de buenas prácticas para los profesionales de la información que desarrollan su actividad en bibliotecas, archivos, museos y otros centros de información y repositorios.

- **Research Data Management Skills Support Initiative (DaMSSI):**

El proyecto se desarrolló en dos etapas, una entre noviembre de 2010 y agosto de 2011, y una segunda fase entre agosto de 2012 y agosto de 2013.

El proyecto DaMSSI tenía como objetivo facilitar el uso de herramientas como: *Vitae's Researcher Development Framework* [9] (RDF) y *Seven Pillars of Information Literacy model*[10].

Se pretendía ayudar a los investigadores a planificar la formación y el desarrollo de profesional en el área de la gestión de datos.

- **Incremental:**

Se trató de un proyecto en colaboración con la *Humanities Advanced Technology and Information Institute* (HATII), de la Universidad de Glasgow, y la *Cambridge University Library.*

Se desarrolló desde noviembre de 2009 hasta marzo de 2011.

El proyecto quería conocer y entender las preocupaciones y las necesidades de los investigadores a la hora de gestionar los datos, para ello había que responder a dos preguntas fundamentales: ¿cómo deben de crearse los datos para que puedan ser encontrados, entendidos y reutilizados a largo plazo?, y lo más importante, ¿cómo conseguirlo?

Para conseguir el objetivo se realizaron distintas actuaciones:

1- Elaboración de una guía sencilla y visual que reflejara buenas prácticas en lo referente a la creación, almacenamiento y gestión de datos.

2- Creación de tutoriales y recursos para formar a profesionales en la gestión de datos.

3- Se fomentó la interactuación entre los investigadores y el personal de apoyo.

4- Se desarrolló una infraestructura integral de gestión de datos tanto en Cambridge como en Glasgow.

5- Se desarrollaron políticas y planes sobre la gestión de datos en ambos centros.

6- Se crearon páginas web para promocionar las actuaciones llevadas a cabo.

[9] Disponible en: http://www.vitae.ac.uk/policy-practice/375-251231/The-Researcher-development-framework-RDF.html

[10] Disponible en: http://www.sconul.ac.uk/sites/default/files/documents/coremodel.pdf

- **I2S2:**

El objetivo del proyecto *Infrastructure for Integration in Structural Sciences* (I2S2) era descubrir qué se necesita para implementar una infraestructura de investigación basada en los datos de las ciencias estructurales.

Para conseguirlo, se estudiaron los datos en todas y cada una de las etapas de su ciclo de vida.

- **ERIM:**

El proyecto *Engineering Research Information Management* (ERIM) fue financiado por el JISC entre octubre de 2009 y marzo de 2011.

Se realizó en colaboración con el Innovative *design and Manufacturing Research Centre* (IdMRC) y el *United Kingdom Office for Library and Information Networking* (UKOLN), ambos de la Universidad de Bath.

El proyecto tenía varios objetivos:

1- Se estudió como realizar una gestión efectiva de los datos de la investigación en el campo de la ingeniería.

2- Se estudiaron las barreras y las oportunidades con las que se cuenta cuando se reutiliza información del campo de la ingeniería, incluyendo los resultados de la investigación llevada a cabo a partir de datos industriales altamente sensibles.

3- Se estudiaron los requisitos que son necesarios para una correcta reutilización de los conjuntos de datos fruto de la investigación.

- **Piloting the LIFE costs Tool in UK HEIs:**

La herramienta LIFE fue creada por el HATII como parte del LIFE3 project[11] que se desarrolló entre agosto de 2009 y septiembre de 2010.

LIFE, es una herramienta que permite a las organizaciones predecir el coste que puede suponer la preservación de los objetos digitales.

El DCC se encargó de realizar el testeo de la herramienta recogiendo la experiencia de distintos usuarios y repositorios institucionales.

[11] Disponible en: http://www.life.ac.uk/3

El objetivo fundamental era evaluar si la herramienta LIVE podía ser útil para los *UK HEI Repositories.*

Se quería conocer si con esta herramienta se podía tener una mayor comprensión de los gastos de funcionamiento de los repositorios, además de identificar errores en los procesos, todo ello con el objetivo de comprender los costos de los procesos para mejorar de esta manera la planificación, además de desarrollar políticas institucionales.

- **Case Studies in the Life Sciences:**

Estudio sobre los beneficios y las barreras que supone la utilización de la metodología de la "ciencia abierta" por parte de los investigadores.

- **SCARP:**

El proyecto consistió en la realización de diferentes estudios de caso para conocer como se trata el tema del depósito, el intercambio, la reutilización, la curación y la conservación de los datos en diferentes disciplinas científicas.

- **ERIS:**

El *Enhancing Repository Infrastructure in Scotland* (ERIS), se desarrolló desde abril de 2009 hasta marzo de 2011.

El propósito del proyecto ERIS era desarrollar, en estrecha colaboración con los investigadores y los administradores de los repositorios de las distintas instituciones, un conjunto de soluciones que motivasen a los investigadores a depositar sus trabajos en los repositorios.

Asimismo, se querían integrar los repositorios en los procesos de investigación y, en consecuencia, desarrollar un servicio de recuperación de recursos a través de repositorios que permitiera a los investigadores escoceses acceder a los resultados de las investigaciones.

3.1 - Australian National Data Service (ANDS).

Antes de explicar en profundidad el *Australian National Data Service* (ANDS) es importante conocer como se trata en Australia el tema de la gestión de los datos de la investigación.

En Australia todas las universidades e instituciones relacionadas con la investigación tienen políticas y directrices relativas a diferentes aspectos de la gestión de datos.

El *Australian Code for the Responsible Conduct of Research*[12] es el encargado de orientar a las instituciones y a los investigadores para que realicen prácticas de investigación responsables. El Código promueve la integridad en la investigación y explica lo que se espera de los investigadores.

Fue desarrollado conjuntamente por el *National Health and Medical Research Council*, el *Australian Research Council* y las universidades de Australia.

Está escrito específicamente para universidades y otras instituciones de investigación del sector público, pero es también una referencia para todas aquellas personas que están fuera de la comunidad investigadora y que necesitan información sobre los estándares relativos a la investigación dentro de Australia.

En el apartado segundo del *Australian Code for the Responsible Conduct of Research*, se habla de la gestión de los datos de la investigación

En él se desarrollan las políticas referentes a la propiedad de los materiales y los datos de la investigación, su almacenamiento, su mantenimiento más allá del final del proyecto, y todo lo relativo al acceso a ellos por parte de la comunidad científica.

La conservación y la gestión de los datos de la investigación es responsabilidad tanto de la institución como de los investigadores.

- **Responsabilidades de las instituciones**

Cada institución debe tener una política sobre la conservación de los materiales y los datos de la investigación. Las instituciones deben reconocer su papel en la

[12] Texto del *Australian Code for the Responsible Conduct of Research*; Disponible en: http://www.adelaide.edu.au/rb/code/#2

gestión de los mismos. La política institucional debe ser consecuente con las prácticas de la disciplina, la legislación pertinente, los códigos y las directrices.

En general, el periodo mínimo recomendado para la retención de datos de la investigación es de 5 años a partir de la fecha de su publicación. Sin embargo, el período durante el cual deben conservarse los datos debe ser determinado por el tipo específico de investigación.

Se recomienda:

- Para proyectos de investigación a corto plazo que tienen, solamente, fines de evaluación, como pueden ser los proyectos de investigación realizados por estudiantes, basta con conservar los datos de investigación 12 meses después de la finalización del proyecto.
- Para la mayoría de los ensayos clínicos, es necesario conservar los datos de investigación 15 años o más.
- Para áreas tales como la terapia génica, los datos de investigación deberán ser conservados de forma permanente.
- Si el trabajo o la investigación tiene valor patrimonial, los datos de investigación deben mantenerse permanentemente, preferentemente dentro de una colección nacional.

La institución debe garantizar una correcta eliminación de los datos de la investigación cuando el período especificado de conservación ha terminado.

Además, las instituciones deben tener instalaciones adecuadas para el almacenamiento seguro de los datos de la investigación, y contar con registros para saber donde se almacenan.

La propiedad de los datos de la investigación puede verse afectada por los acuerdos de financiación de los proyecto, pero, como regla general, los datos acumulados al final de un proyecto pueden ser propiedad de la institución que organizó el proyecto, de otra entidad con un interés en la investigación, o de un repositorio central.

La institución debe garantizar la seguridad y la confidencialidad de los datos, por eso debe de contar con una política estricta sobre la propiedad y el acceso a las bases de datos y a los archivos, que debe ser consecuente con los requisitos de confidencialidad, la legislación, las normas de privacidad y otras directrices.

- **Responsabilidades de los investigadores**

El investigador debe decidir qué datos y materiales se deben conservar, aunque en algunos casos, esto está determinado por la ley, por la agencia de financiación, por el editor o por las normas profesionales.

El objetivo es que los materiales y los datos que se conservan sirvan para justificar y defender los resultados de la investigación

Los datos deben estar disponibles para su uso por parte de otros investigadores, especialmente cuando son fruto de investigaciones difíciles o imposibles de repetir, a menos que haya impedimentos éticos, problemas de privacidad o cuestiones de confidencialidad.

Deberán conservarse, por lo menos, durante el periodo mínimo especificado en la política institucional.

Los investigadores deben gestionar los datos de la investigación de acuerdo con la política de la institución. Para lograr esto, deben de:

- Mantener registros claros y precisos de los métodos de investigación y de las fuentes de los datos, incluyendo todas las autorizaciones concedidas, durante y después, del proceso de investigación.
- Asegurarse de que los datos de investigación están almacenados de forma segura, incluso cuando no están en uso.
- Proporcionar el mismo nivel de atención y protección a todos los tipos de datos, desde los registros primarios de la investigación, hasta los datos fundamentales de las investigaciones.
- Conservar los datos de investigación de tal manera que sean duraderos y recuperables.
- Mantener un catálogo de los datos de la investigación que sea accesible.
- Administrar los datos de investigación de acuerdo con los protocolos éticos y la legislación pertinente.

En todo el proceso deberá mantenerse la confidencialidad de los datos debiendo usarse en la forma convenida.

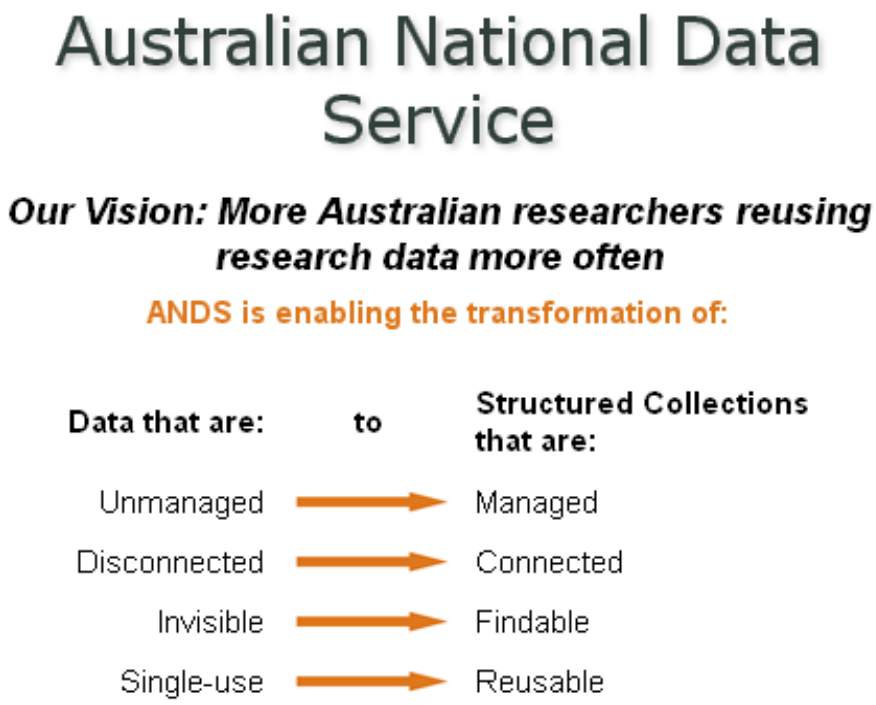
Una vez conocidas las políticas y las directrices australianas relacionadas con la gestión de los datos de la investigación vamos a pasar a explicar en profundidad el *Australian National Data Service* (ANDS).

El *Australian National Data Service* (ANDS) actúa como coordinador de la gestión de datos de la investigación australianos, se inició en 2009 con una financiación de 70 millones de dólares australianos para un periodo de 4 años.

Se trata de un proyecto nacional que pretende que los investigadores cuenten con datos de alta calidad que puedan ser reutilizarlos.

Para ello es necesario que los datos pasen de una situación en la que son inmanejables, invisibles, están desconectados y son de uso particular a convertirse en colecciones de datos estructurados, manejables, conectados, y que pueden ser encontrados y reutilizados.

Figura 6. Las cuatro transformaciones del ANDS



Fuente: ANDS

El ANDS apuesta por promover la creación y el fortalecimiento de infraestructuras institucionales para los datos de investigación, además de proporcionar guías y recomendaciones para la gestión, la producción y la reutilización de los datos…

El objetivo fundamental es facilitar a los investigadores australianos la publicación, el acceso y el uso de los datos de investigación.

Su lema es claro, contra mejores datos haya, mejor será la investigación.

Entre los proyectos que desarrolla el ANDS cabe destacar:

- ***Australian Research Data Commons***

Se presenta como un lugar de reunión para los investigadores y los datos australianos, pretende conseguir que se haga un mejor uso y aprovechamiento de los datos de investigación de Australia.

Reúne datos de las investigaciones de las universidades australianas, de los organismos de investigación financiados con fondos públicos y de las organizaciones gubernamentales.

El *Australian Research Data Commons* proporciona colecciones de datos compartibles y descritas, además de conectar los datos con los investigadores, la investigación, los materiales y las distintas instituciones.

Se pretende garantizar una mayor utilización y reutilización de los datos existentes, así como asegurar que se realiza una mejor gestión de los nuevos datos que se van generando fruto de las investigaciones australianas.

Representa un cambio en la forma en la que podemos acercarnos a los datos ya que permite el acceso a una fuente única y autorizada de datos.

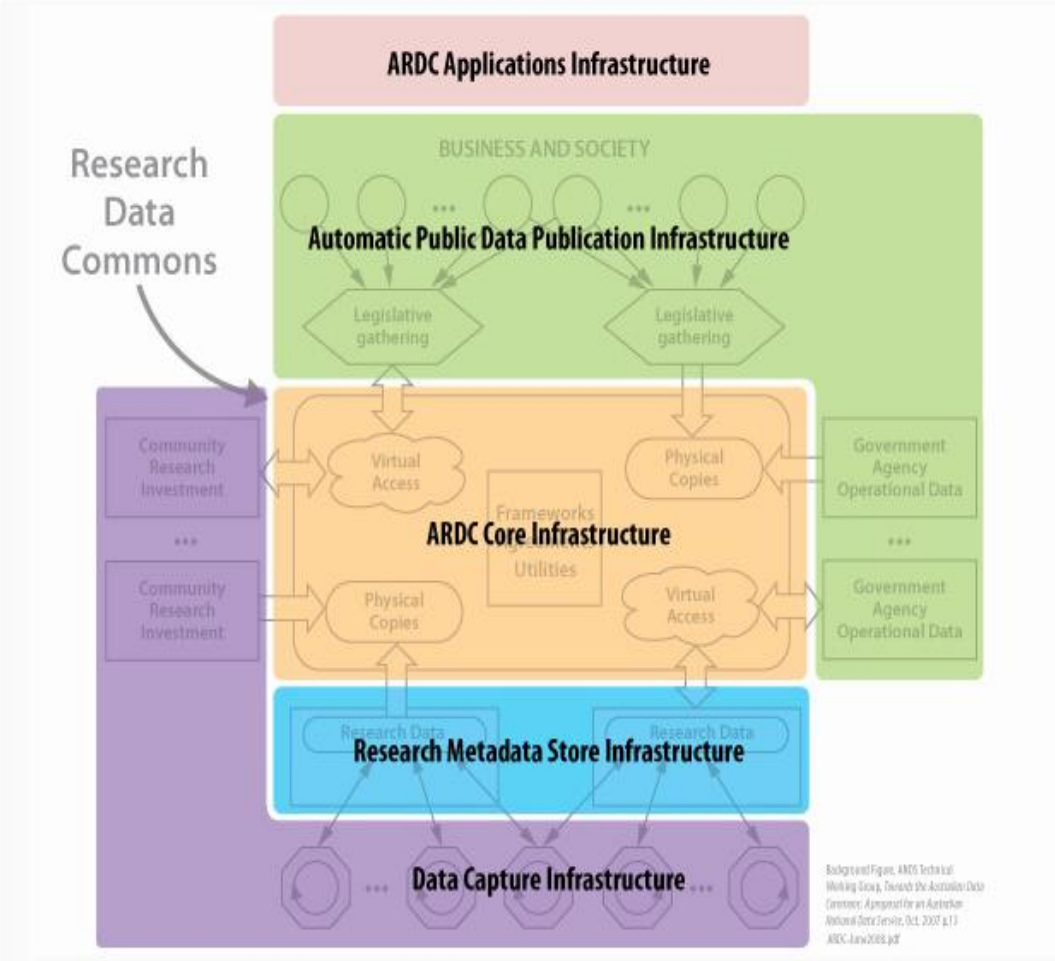
Además proporciona una serie de ventajas muy importantes para los investigadores australianos:

- Los datos de la investigación se publicarán de forma rutinaria mejorando la reputación de los investigadores australianos y el impacto de las distintas investigaciones.

- Gracias al entorno y a las infraestructuras creadas será más sencillo para los investigadores internacionales trabajar colaborativamente con los investigadores australianos.

- Facilitará que se realicen nuevas investigación a partir de los datos existentes.
- Si los datos se asocian correctamente será más fácil encontrar datos relevantes y emitir un juicio acerca de su valor.

Figura 7. Infraestructura del *Australian Research Data Commons*



Fuente: ANDS

La gestión de los datos es otra de las labores fundamentales del *Australian National Data Service*, para enfatizar su importancia trata de contestar a cuatro preguntas:

1. ¿Por qué es necesaria la gestión de datos?

- Para preservar la integridad de la investigación
- Para permitir que los datos estén disponibles y que otros puedan hacer uso de ellos
- Para ayudar a los investigadores a reducir el riesgo que supone la pérdida de datos
- Para asegurar el acceso continuo a los datos

2. ¿Por qué es necesario conectar los datos?

- Para conectar los datos con las personas, los proyectos y las publicaciones
- Para mejorar la capacidad de detección de los datos
- Para enlazar los datos con las investigaciones
- Para proporcionar un contexto más rico que de valor a los datos

3. ¿Por qué hay que hacer visibles los datos?

- Para mostrar la excelencia de la investigación
- Para que los investigadores puedan basarse en datos existentes y no tengan que volver a realizar las mismas investigaciones, ahorrando esfuerzo, tiempo y dinero
- Para fomentar la innovación
- Para proporcionar a los investigadores la capacidad de resolver grandes problemas en sus disciplinas

4. ¿Por qué es importante reutilizar los datos?

- Para poder verificar las distintas reclamaciones que pueden surgir en una investigación
- Para realizar nuevos descubrimientos a partir de los datos existentes
- Para integrar conjuntos de datos para realizar nuevos análisis
- Para volver a analizar investigaciones costosas, raras o irrepetibles
- Para tratar de reducir la duplicación de esfuerzos

La gestión de datos incluye todas aquellas actividades asociadas con los datos que no incluyan su uso directo, tales como:

- Organización de los datos
- Realización de copias de seguridad
- Archivo de los datos para la conservación a largo plazo
- Distribución o publicación de los datos
- Garantizar la seguridad de los datos confidenciales
- Sincronización de datos

Para que un plan de gestión de datos funcione correctamente se debe de definir perfectamente qué datos van a ser creados, qué políticas van a regular los datos, quién será el propietario y quién tendrá acceso a los datos, qué prácticas de gestión de datos se utilizarán, qué instalaciones y equipo se necesitará, y quién será el responsable de realizar cada una de estas actividades.

Una gestión inadecuada de los datos puede provocar la pérdida de datos o la violación de la intimidad de las personas.

Para llevar a cabo la gestión de datos de la investigación, el *Australian National Data Service,* desarrolla un programa de trabajo basado en 7 ejes:

- Formación y concienciación
- Elaboración de políticas para la gestión de datos de la investigación
- Planificación de la gestión de datos

- Gestión de datos personales
- Licencias, copyright y datos
- Captura de datos
- Almacenamiento de datos

1- Formación y concienciación:

La formación y la concienciación sobre los retos de la gestión de datos es fundamental.

Las buenas prácticas en su gestión permiten a los investigadores y a las instituciones mejorar la eficiencia de la investigación, al permitir que los datos estén disponibles para el intercambio, la validación y la reutilización.

Para conseguir una buena gestión de los datos desde el inicio se deben seguir las siguientes etapas:

- Planificación
- Recopilación
- Análisis
- Publicación
- Archivo
- Reutilización

2- Elaboración de políticas para la gestión de datos de la investigación:

Contar con políticas y procedimientos institucionales, tales como directrices, protocolos y normas, es fundamental para una buena gestión de los datos de la investigación.

Además es un requisito obligatorio establecido en el *Australian Code for the Responsible Conduct of Research.*

El ANDS ha elaborado una política de gestión de datos de la investigación que puede ser adaptada para su uso en diferentes universidades e instituciones australianas.

Esta policía debe de contar con los siguientes puntos:

- **Nombre:** debe ser claro, conciso e informativo, no debe de incluir siglas o abreviaturas.
- **Política de Uso:** breve declaración que describe lo que se quiere lograr con la política de gestión de datos.
- **Principios u objetivos clave:** hace referencia a los objetivos de la política de gestión.
- **Definiciones:** en este apartado se deben definir todos los términos que se están utilizando.
- **Excepciones:** se deben describir las situaciones en las que no se puede aplicar la política.
- **Aplicación y responsabilidades:** hay que definir las responsabilidades específicas de los investigadores, de la institución…
- **Periodos de conservación de los datos:** se debe definir el período durante el cual se conservarán los datos. Dependerá del tipo de investigación.
- **Almacenamiento:** se debe detallar dónde y cómo se van a almacenar los datos de investigación. Se pueden almacenar objetos digitales, por ejemplo, archivos, bases de datos, fotografías, grabaciones… u objetos físicos, como papeles, artefactos…
- **Seguridad y Protección:** las instituciones deben proporcionar facilidades para el almacenamiento seguro de los datos de la investigación.
- **Propiedad y acceso a los datos:** se debe determinar quién tiene acceso a los datos y de qué manera, la norma dice que los datos deben estar disponibles para su uso por parte de otros investigadores a menos que lo impidan las normas de privacidad o cuestiones de confidencialidad.
- **Eliminación o movimiento de datos y registros:** ¿Quién lo va a realizar? ¿Qué va a ser movido? ¿De dónde a dónde? ¿Cuándo y cómo?
- **Destrucción de registros:** ¿Quién lo hará? ¿Qué va a ser destruido? ¿Dónde, cuándo y cómo?
- **Propiedad de los datos:** cada institución debe tener una política clara sobre la propiedad de los materiales y los datos de la investigación tanto durante como después del proyecto de investigación. Esta política debe

cubrir todas las situaciones que se pueden plantear durante una investigación.

- **Proyectos de investigación colaborativa:** antes de comenzar proyectos de investigación colaborativa se deben de establecer acuerdos referentes a la propiedad de los datos, como mínimo, estos acuerdos deben cubrir la propiedad intelectual y la propiedad de los equipos y de los datos.

- **Requisitos especiales:** se deben establecer requisitos específicos para determinados datos y materiales tales como: cuadernos de laboratorio, patentes, datos etnográficos…

- **Registros de almacenamiento:** se han de conservar los datos de investigación de manera que sean duraderos y recuperables, hay que mantener un catálogo de los datos de la investigación que sea accesible.

- **Políticas y documentos relacionados:** hay que señalar qué documentos o políticas institucionales están relacionadas con nuestra política, el *Australian Code for the Responsible Conduct of Research*, políticas sobre los derechos de autor o de propiedad de investigación…

3- Planificación de la gestión de datos:

Los datos de alta calidad deben administrarse correctamente. La planificación es una parte fundamental de este proceso. La responsabilidad principal de la gestión de los datos de la investigación recae, por lo general, en el investigador. La mayoría de las instituciones que se dedican a la investigación tienen políticas integrales de gestión de datos, además de procedimientos para apoyar a sus investigadores.

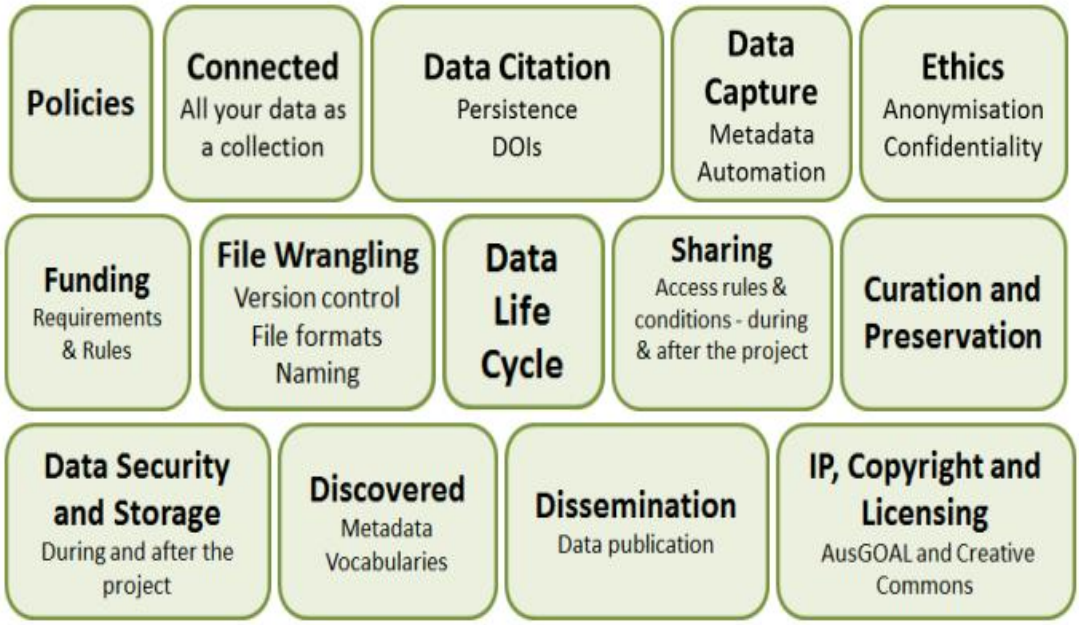
La planificación incluye entre sus objetivos mejorar la toma de decisiones con la meta de concretar un fin buscado. Por consiguiente, una estrategia de planificación debe tener en consideración la situación presente y todos aquellos factores ajenos y propios que pueden generar repercusiones para lograr ese fin.

Sólo es posible diseñar una planificación tras la identificación precisa del problema que se ha de abordar. Una vez conocida e interpretada esa problemática, se postula el desarrollo de las alternativas para su abordaje o solución.

Después de definir las ventajas y las desventajas de esos posibles enfoques, se opta por la planificación más conveniente y se decide su puesta en práctica.

En el siguiente diagrama se presentan los distintos elementos que investigadores e instituciones deben de considerar a la hora de planificar la gestión de datos.

Figura 8. Elementos de los planes de gestión de datos



Fuente: ANDS

También es importante que un plan de gestión de datos pueda contestar a las siguientes cuestiones:

- ¿Hay que convertir los datos existentes?
- ¿Qué datos se van a crear?
- ¿Quién es el dueño de los datos creados? ¿Quién podría estar interesado en ellos?
- ¿Qué formatos de archivo se utilizarán para los datos?
- ¿Qué tipo de metadatos se van a utilizar? ¿Qué formato o estándar se va a seguir?
- ¿Quién tendrá acceso a los datos? ¿Cómo se van a proteger los datos contra el acceso no autorizado?
- ¿Cómo se van a nombrar los archivos de datos? ¿Cómo se van a organizar los datos? ¿Cómo se van a gestionar las transferencias y la sincronización de datos entre diferentes equipos? ¿Se va a llevar un registro de las diferentes versiones de los archivos de datos y de los documentos?

- ¿Dónde se almacenarán los datos?
- ¿Cuál va a ser la estrategia a seguir en lo referente a las copias de seguridad? ¿Se van a conservar copias de seguridad fuera del edificio? ¿El proceso para realizar las copias de seguridad va a ser automático?
- ¿Qué herramientas de gestión de bibliografía se van a utilizar? ¿Cómo se van a compartir las referencias con los demás miembros del grupo?
- ¿Qué datos se van a compartir con otros? ¿Cómo se va a realizar?
- ¿Qué datos se van a destruir? ¿Cuándo? ¿Cómo?
- ¿Quién será el responsable de cada uno de los elementos de este plan?
- ¿Cuanto va a costar este plan?

4- Gestión de datos personales:

La información personal es aquella información que identifica o sirve para identificar a alguien, por ejemplo, el nombre o la dirección, los registros médicos, información de cuentas bancarias, fotos, videos…

Cuando la investigación se basa en datos personales, los investigadores deben respetar las normas éticas establecidas, pero eso no quiere decir que los datos no pueden ser compartidos ética y legalmente.

Los investigadores pueden obtener el consentimiento de las personas que participan en la investigación para hacer uso de los datos recopilados, para conseguirlo es importante informar a los participantes acerca de la manera en la que los datos van a ser almacenados, conservados y utilizados, también es importante definir cómo se mantendrá la confidencialidad de dichos datos.

5- Licencias, copyright y datos:

La propiedad intelectual o los derechos de autor han dejado de ser un tema de conversación exclusivo de los abogados o los juristas y se ha convertido en un tema de debate en el ámbito del acceso a los datos y a las publicaciones.

A la hora de utilizar los datos de la investigación es fundamental saber lo que se puede hacer con ellos.

No tener claros los permisos para reutilizar los datos puede tener el mismo resultado que prohibir su reutilización, la incertidumbre puede ser suficiente para desanimar a la reutilización.

ANDS ofrece una serie de servicios y proyectos para ayudar a las organizaciones a conocer las licencias existentes para trabajar a partir de ahí conociendo los permisos y las condiciones para el uso y la reutilización de los datos.

ANDS está promoviendo el uso de AusGOAL[13] que ofrece un marco informativo que sirve como guía para las personas y las instituciones a la hora de seleccionar licencias apropiadas para sus datos.

A través de varias preguntas sencillas en la propia página web de AusGOAL (www.ausgoal.gov.au) se puede determinar la licencia más adecuada para la información que se posee.

AusGOAL se ha utilizado con éxito en muchos países de la Commonwealth y en muchas agencias del gobierno australiano.

Incorpora licencias *Creative Commons*.

Estas licencias permiten que un autor pueda ceder algunos derechos sobre su creación en unas condiciones determinadas, y señala los derechos que se reserva si es el caso.

AusGOAL incluye las seis licencias estándares de *Creative Commons*, que son las más conocidas y las más utilizadas.

Estas licencias permiten la reproducción, la distribución y la comunicación de los datos siempre que se cumplan las condiciones establecidas por el titular de los derechos.

Las restricciones vienen determinadas por el tipo de licencia escogida:

- **Reconocimiento (by):** se permite el uso comercial de la obra y de las posibles obras derivadas, la generación y distribución de la cuales está permitida sin ninguna restricción.

[13] Disponible en: http://www.ausgoal.gov.au/

- **Reconocimiento-CompartirIgual (by-sa):** Se permite el uso comercial de la obra y de las posibles obras derivadas, la distribución de las cuales debe hacerse mediante una licencia igual que la sujeta a la obra original

- **Reconocimiento-SinObraDerivada (by-nd):** Se permite el uso comercial de la obra pero no la generación de obras derivadas.

- **Reconocimiento-NoComercial (by-nc):** se permite la generación de obras derivadas siempre que no se haga un uso comercial. Tampoco puede utilizarse la obra original con fines comerciales.

- **Reconocimiento-NoComercial-CompartirIgual (by-nc-sa):** no se permite un uso comercial de la obra original ni de las posibles obras derivadas, la distribución de las cuales debe hacerse mediante una licencia igual que la sujeta a la obra original.

- **Reconocimiento-NoComercial-SinObraDerivada (by-nc-nd):** No se permite un uso comercial de la obra original ni la generación de obras derivadas.

6- Captura de datos:

ANDS pretende simplificar el proceso de captura de datos mediante la construcción de infraestructuras que permitan la integración de todos los procesos relacionados con la creación o captura de datos para lograr de esta manera una ingesta efectiva de los datos fruto de la investigación y un correcto almacenamiento de los metadatos en la institución o en otro lugar.

Para ello ANDS desarrolla softwares destinados a permitir una mejor gestión y una correcta descripción de los datos de investigación y los metadatos asociados.

También se encarga de la construcción de infraestructuras e instalaciones de almacenamiento de metadatos.

Esta integración de procesos facilita a los investigadores la compartición de datos mediante el *Australian Research Data Commons*, explicado anteriormente, a través de su ventana: *Data Research Australia*[14]

7- Almacenamiento de datos:

El *Australian Code for the Responsible Conduct of Research* afirma que:

"Los materiales y los datos se almacenan y se conservan para poder justificar los resultados de la investigación y defenderlos si son desafiados. También se debe considerar el valor potencial del material para su posterior reutilización, especialmente cuando la investigación sea difícil o imposible de repetir".

"Las instituciones deben proporcionar instalaciones para el almacenamiento seguro de los datos de investigación".

"Los investigadores deben gestionar los datos de investigación de acuerdo con la política de la institución".

El almacenamiento de datos está fuera del alcance de las actividades que desarrolla el ANDS pero es esencial para que se cumpla su objetivo fundamental, que los datos de investigación estén disponibles con facilidad.

En cambio, el ANDS si que se encarga de la creación de metadatos.

[14] Disponible en: http://researchdata.ands.org.au

4 - Conclusiones

Los avances tecnológicos están produciendo enormes cambios en el seno de la actividad científica, la transformación es visible en las nuevas formas de publicación del conocimiento científico, y en la tendencia actual hacia una mayor publicación en acceso abierto.

Estos cambios no afectan exclusivamente a las publicaciones científicas, también se ha producido un cambio en la manera en la que se gestionan los datos de investigación.

Los datos son considerados, cada vez más, como una fuente de conocimiento independiente de las publicaciones científicas.

La gestión de los datos de investigación es un área de trabajo emergente y uno de los temas de actualidad en nuestro ámbito profesional.

La comunidad científica tiene una necesidad cada vez mayor de disponer de una infraestructura de gestión de datos.

Los proyectos desarrollados por organismos nacionales e internacionales muestran que los datos constituyen una realidad en ciernes, pese a que los niveles de compartición aún sean bajos y crezcan con lentitud.

Esto se debe a que cada vez hay más necesidad de comparar, preservar y gestionar grandes cantidades de datos por parte de instituciones relacionadas con el ámbito científico y técnico, bibliotecas, instituciones de carácter superior, públicas y privadas…

Realizar una correcta gestión de los datos, asegurar su buen almacenamiento, facilitar el acceso a ellos y su reutilización es una tarea fundamental.

La gestión de los datos de investigación debe llevarse a cabo durante todo el proceso de investigación: antes de la creación de los datos, durante su creación y uso y a lo largo de su ciclo de vida.

Algunos datos son únicos y no pueden ser reemplazados si se destruyen o se pierden.

Por el contrario, su correcta gestión y sostenibilidad es muy importante debido a que los datos son fundamentales para el fomento de la innovación científica y tecnológica, además, suponen un ahorro importantísimo para estas instituciones, ya que si se realiza una correcta custodia se pueden aprovechar los datos ya

existentes en proyectos de investigación futuros, es decir, se pueden re-utilizar y evitar la duplicación de los esfuerzos de investigación.

Hay que tener en cuenta que gran parte de las investigaciones se basan en trabajos e investigaciones anteriores, además, conservar los datos garantiza la trazabilidad y la repetibilidad de los experimentos.

Pero la publicación de los datos de la investigación sigue teniendo sus defensores y sus detractores sus ventajas y sus limitaciones.

Aquellos que abogan por su publicación afirman que compartir los datos proporciona una audiencia mucho mayor debido a que se aumenta la visibilidad del trabajo, provoca que la difusión internacional sea superior y que se puedan alcanzar estratos sociales más amplios, acercándose de este modo a uno de los objetivos subyacentes de la ciencia.

Además, publicar los datos es beneficioso para la entidad financiadora, ya que permite un mayor uso y explotación de la investigación por parte de un mayor número de personas, lo que facilita nuevas investigaciones.

También hay que destacar los beneficios que aporta a las instituciones académicas, hacer accesibles los datos aporta más publicidad para la investigación desarrollada en la institución.

Pero también hay una serie de limitaciones que pueden afectar a la compartición de datos:

- La falta de una política de datos global a nivel nacional e internacional.
- La falta de coordinación entre las iniciativas existentes.
- La fragmentación de depósitos de datos.
- La pérdida de conjuntos de datos que no están adecuadamente archivados ni documentados.
- La falta de interconexión entre los datos.
- La diversidad de formatos y la falta de interoperabilidad.
- La falta de una buena infraestructura de datos.
- Las dudas en torno a la calidad de los metadatos debido a que los autores no son catalogadores y los profesionales de la información, hasta el momento, han tenido una participación escasa en lo que a control se refiere.

Es un hecho que el acceso a largo plazo y de forma sostenible a los datos es una tarea enorme, pero los proyectos y las actuaciones realizadas por organismos como el *Digital Curation Centre* (DCC) y el *Australian National Data Service* (ANDS) demuestran que si que es posible realizar una correcta gestión de los datos de la investigación.

Para ello se necesitan políticas claras a nivel nacional o institucional que definan claramente planes de gestión de datos y repartan los roles y las responsabilidades entre los distintos actores.

También se necesitan infraestructuras y equipos que permitan el acceso, el uso y la reutilización de los datos.

Una gestión inadecuada de los datos puede provocar su pérdida o la violación de la intimidad de las personas.

5 - Bibliografía

- ANDS. <www.ands.org.au>

- AusGOAL. <http://www.ausgoal.gov.au/>

- Australian Code for the Responsible Conduct of Research <http://www.adelaide.edu.au/rb/code/#2>

- Borgman, C; Wallis, J; Enyedy, N. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries, 2007 7:17-30.

<http://dx.doi.org/10.1007/s00799-007-0022-9>

- Budapest Open Access Initiative (BOAI). Open Society Institute (OSI). <http://www.soros.org/openaccess/>

- Choudhury, S. Rethinking Scholarly Communication: Building Data Curation Infrastructure, 2009. <http://www.it.utah.edu/leadership/research/ciday/2009/notes/choudhury.pdf>

- Comisión Europea. *Agenda digital europea* 2003 <http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/1524&format=HTML&aged=0&language=ES&guiLanguage=en>

- Comisión Europea. Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones. Datos abiertos: Un motor para la innovación, el crecimiento y la gobernanza transparente, 2011. <http://ec.europa.eu/information_society/policy/psi/docs/pdfs/opendata2012/open_data_communication/es.pdf>

- *Data Asset Framework: Implementation guide. Londres: JISC, Octubre 2009.* <http://www.data-audit.eu/>

- DCC. <http://www.dcc.ac.uk/>

- *F. Giunchiglia, R. Chenu.* Scientific Knowledge Objects v.1, 2009. <http://wiki.liquidpub.org/mediawiki/upload/8/80/SKO_Main_v1_006.pdf>

- Franklin, Jack. (2003). Open access to scientific and technical information: the state of the art. *Information Services and Use* Vol 23, n 2,3 pp.67-86

- García y Rivera (2009). Open Access y web social: una mancuerna vital para la investigación científica. *Proceedings del 2º Congreso Mundial de Información y Conocimiento: Aspectos Tecnológicos*

- Gold, A. Data Curation and Libraries: Short-Term Developments, Long-Term Prospects. Office of the Dean (Library) 2010.

<http://digitalcommons.calpoly.edu/lib_dean/27>

- Harnad, S.; Brody, T. (2004). "Comparing the impact of open access (OA) vs. non-OA articles in the same journals". *D-lib magazine*. 10: 6.

- Harnad S.; Brody T, Vallières F. (2004). "The Access/Impact Problem and the Green and Gold Roads to Open Access". *Serials Review*. ty30: 310–314.

- Harvey, Ross. Digital Curation: A How-To-Do-It Manual® (Number 170). New York, NY: Neal-Schuman Publishers, 2010. 225 pp. $80.00 USD. ISBN-13: 978-1-55570-694-4.

- Hedstrom, M; Montgomery, S. Digital Preservation Needs and Requirements in RLG Member Institutions. Mountain View CA: RLG, 1998. p3. <www.oclc.org/programs/ourwork/past/digpresneeds/digpres.pdf>

- Higgins, S. Digital Curation: The Emergence of a New Discipline. *The International Journal of Digital Curation*, 2011; 6 (2).

- Higgins, S. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation* 2008; 3(1); 134-140.

- Hitchcock, S; Brody, T; Hey, J; and Carr, L. Laying the foundations for Repository Preservation Services: final report from the PRESERV PROJECT. 2007.
<http://www.jisc.ac.uk/media/documents/programmes/preservation/preserv-final-report1.0.pdf>

- Hitchcock, S; Brody, T; Hey, J; and Carr, L. Survey of repository preservation policy and activity. 2008 <http://preserv.eprints.org/papers/survey/survey-results.html>

- Kim, Y; Addom, BK; Stanton, JM. Education for eScience Professionals: Integrating Data Curation and Cyberinfrastructure. *International Journal of Digital Curation*, 2011, *6*(1) 125-138

- Kirkcz, J. (s/f). Scientific Communication as an object of science. http://www.portlandpress.com/pp/books/online/tiepac/session7/ch1.htm.

- Martinez-Uribe, L; Macdonald, S. User Engagement in Research Data Curation. 2009. <http://www.era.lib.ed.ac.uk/handle/1842/3206>

- Melero, R.; Abadal, E.; Abad, F., y Rodríguez-Gairín, J. M. (2009): Situación de los repositories institucionales en España: informe 2009 [s.I.]; Grupo de investigación Acceso Abierto a la Ciencia, 54. <http://hdl.handle.net/10261/11354>

- NESTA: National Endowment for Science Technology and the Arts, *Open Science Case Studies*. 2009 y 2010. <http://www.rin.ac.uk/our-work/data-management-and-curation/open-science-case-studies>

- Ogburn JL. The Imperative for Data Curation. Portal-libraries and the Academy. 2010; 10: 241-246. DOI: 10.1353/pla.0.0100.

- OpenDOAR.<http://www.opendoar.org/find.php>

- Open Archives Initiative. Object Reuse and Exchange. <http://www.openarchives.org/ore/>

- Palermo, Darwin. Open Access. Acceso libre al progreso. <http://www.ladinamo.org/ldnm/articulo.php?numero=27&id=706>

- Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library and Archives Journal*, Issue 1.Retrieved (preprint) June 18, 2008; <http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/libarch_curation.pdf>

- Pérez-González, L. E-ciencia y la información como bien público, algunas propuestas, 2011. <http://eprints.rclis.org/handle/10760/16527#.TyfRUMXbi68>

- Rojas, L. (2008) "¿Por qué publicar artículos científicos?". *Revista Orbis*, 10: 4, 120 – 137

- Shneiderman, B. Science 2.0. *Science*, 2008, v. *319*, n. 5868, pp. 1349-1350

- Suber, Peter: (2004). A Very Brief Introduction to Open Access. <http://www.earlham.edu/~peters/fos/brief.htm>

- Suber, Peter. (2007). Budapest Open Access Initiative: Frequently Asked Questions. <http://www.earlham.edu/~peters/fos/boaifaq.htm>

- Torres-Salinas, D. Compartir datos (data sharing) en ciencia: el contexto de una oportunidad, *Thinkepi* 2010

- Torres-Salinas, D. Primeros pasos hacia la gestión de datos de investigación en las universidades: la iniciativa DAF, *Thinkepi* 2010

- Torres-Salinas, D; Robinson-García, N; Cabezas-Clavijo, Á. Compartir los datos de investigación: introducción al *data sharing*". *El profesional de la información*, enero-febrero 2012, v. 21, n. 1

- W3C. *Publishing open government data*. W3C working draft, 2009. <http://www.w3.org/TR/2009/WD-gov-data-20090908>

-Wells Parham, S; Bodnar, J; Fuchs, S. Supporting tomorrow's research: Assessing faculty data curation needs at Georgia Tech. *Coll. res. libr. news January 2012, 73(1) 10-13*

- Whitfield, John. (2011). Open access come of age. *Nature*, June v.474, n.428, <http://www.nature.com/news/2011/110621/full/474428a.html>

- Wright, M; Sumner, T; Moore, R; and Koch, T. Connecting digital libraries to eScience: the future of scientific scholarship. International Journal of Digital Libraries, 2007. 7:1-4. <http://www.springerlink.com/content/832616v17076317m/>

- Yakel E. (2007). Digital curation. *OCLC Systems & Services*, 2007. 23 (4) pp.335-340. <http://www.emeraldinsight.com/journals.htm?articleid=1631436&show=abstract>

6 - Índice de figuras: