DATOS Y METADATOS DE INVESTIGACIÓN EN CIENCIAS SOCIALES Y HUMANIDADES: UNA APROXIMACIÓN DESDE LOS REPOSITORIOS TEMÁTICOS DE DATOS

Social sciences and humanities research data and metadata: A perspective from thematic data repositories

Nancy-Diana Gómez, Eva Méndez y Tony Hernández-Pérez

Nota: This article can be read in its original English version on:
http://www.elprofesionaldelainformacion.com/contenidos/2016/jul/04.pdf

Nancy-Diana Gómez es alumna de doctorado del programa *Archivos y bibliotecas en el entorno digital* de la *Universidad Carlos III de Madrid*. Ha sido docente del *Departamento de Biblioteconomía y Documentación* de esta misma universidad (2009-2013) y co-coordinadora de la *Lista Latinoamericana de Acceso Abierto y Repositorios* (*Llaar*). Participa en proyectos de investigación nacionales e internacionales. Es bibliotecaria y licenciada en artes por la *Universidad de Buenos Aires*. Fue docente de la carrera de bibliotecología y documentación en la *Universidad de Buenos Aires* y directora de la *Biblioteca Central* de la *Facultad de Ciencias Exactas y Naturales* de la misma universidad (1994-2005).
http://orcid.org/0000-0002-6218-6248

ndgomez@bib.uc3m.es

Eva Méndez es profesora titular del *Departamento de Biblioteconomía y Documentación* de la *Universidad Carlos III de Madrid*, donde es actualmente vicerrectora adjunta de *Estrategia y educación digital*. Doctora en documentación, su docencia e investigación versa sobre metadatos, web semántica, bibliotecas digitales, acceso abierto, políticas de información y web social. Es miembro del *Advisory board* del *Dublin Core* (*DCMI*), desde 2015 pertenece también al comité asesor de *OpenAire* y al comité ejecutivo de *Rebiun*. Ha participado como experto independiente para la *Comisión Europea* en bibliotecas digitales y *open science*.
http://orcid.org/0000-0002-5337-4722

emendez@bib.uc3m.es

Artículo recibido el 18-04-2016
Aceptación definitiva: 08-06-2016

Tony Hernández-Pérez es doctor en ciencias de la información y profesor del *Departamento de Biblioteconomía y Documentación* de la *Universidad Carlos III de Madrid* en donde dirige el programa de doctorado en documentación. Su labor docente e investigadora está ligada al grupo *TecnoDoc* incluyendo asignaturas de web social, gestión de contenidos web, metadatos, búsqueda y recuperación de información, e-learning y documentación periodística y audiovisual.
http://orcid.org/0000-0001-8404-9247

tony@bib.uc3m.es

Universidad Carlos III de Madrid
Facultad de Humanidades, Comunicación y Documentación
C/ Madrid, 128. 28903 Getafe (Madrid), España

Resumen

Se estudian los repositorios de datos de investigación en ciencias sociales y humanidades (CSH), recogidos en el *Registro de repositorios de datos de investigación* (*re3data*), prestando especial atención a los modelos de metadatos que utilizan para describir los datasets incluidos en ellos. Se revisan a nivel global los 397 repositorios que, según *re3data*, recogen datos de investigación sobre esas disciplinas, incluidos, los de carácter multidisciplinar. Se discute y reflexiona sobre las particularidades de los datos de investigación en estas disciplinas y sobre la cobertura e información que recoge *re3data*. Se analizan los esquemas y estándares de metadatos más utilizados en los repositorios de CSH, con un análisis más pormenorizado de los seis repositorios de datos especializados más importantes.

Palabras clave

Repositorios; Datos de investigación; Metadatos; Ciencias sociales; Humanidades; *Re3data*.

Abstract

This paper studies research data repositories in the social sciences and humanities (SSH), from the *Registry of Research Data Repositories* (*re3data*), paying particular attention to metadata models used to describe the datasets included in them. 397 repositories are reviewed at the general level, including those of a multidisciplinary nature. We discuss and reflect on the special features of research data in these disciplines, and on coverage and information collected by *re3data*. The metadata schemas and standards most commonly used in SSH repositories are analyzed, with special emphasis on the six main repositories.

Keywords

Repositories; Research data; Metadata; Social sciences; Humanities; *Re3data*.

Gómez, Nancy-Diana; **Méndez, Eva**; **Hernández-Pérez, Tony** (2016). "Social sciences and humanities research data and metadata: A perspective from thematic data repositories". *El profesional de la información*, v. 25, n. 4, pp. 545-555.

http://dx.doi.org/10.3145/epi.2016.jul.04

1. Introducción

La gestión de datos de investigación está cobrando cada vez más importancia en todos los campos científicos. Una evolución lógica y necesaria debida, por un lado, al desarrollo tecnológico que permite cada vez más, una ciencia basada en datos, y por otro, al impulso político de la idea de *open science* que incluye, además del acceso abierto a las publicaciones, la apertura de los datos utilizados en las investigaciones.

Compartir datos de investigación se ha convertido en una práctica habitual en disciplinas en las que existe una cultura científica muy colaborativa, como la física, la astronomía (**Pepe** *et al.*, 2014) o la genética (**Paltoo** *et al.*, 2014). A esa cultura disciplinar se une además el hecho de que las instituciones públicas que financian la investigación han comenzado a exigir a los investigadores que hagan públicos sus resultados no sólo en forma de publicaciones sino también, abriendo los datos subyacentes utilizados. Lo recomienda la *OCDE* (2015) y lo exige el gobierno de EUA desde 2013 a través de las diversas agencias de financiación: *National Science Foundation* (*NSF*, 2014) y los *National Institutes of Health* (*NIH*, 2015), entre otros. En Europa, el acceso abierto a los datos de investigación ha sido, hasta ahora, sólo un piloto (*ORD Pilot*) para nueve áreas de proyectos financiados en el marco de *Horizon 2020*, invitando a otras áreas y programas a participar voluntariamente (*European Commission*, 2016). Sin embargo, el 19 de abril de 2016, la *Comisión* declaró que los datos de investigación abiertos serán la opción por defecto para todos los nuevos proyectos financiados en *H2020* a partir de 2017 (COM 2016, p. 8).

La tendencia a la apertura de datos está creciendo en todas las instituciones vinculadas a la investigación, tanto por parte de las agencias que la financian, como por las organizaciones que la llevan a cabo (ej. la *League of European Research Universities* (*LERU*, 2013) o por los editores de revistas que publican sus resultados (*PLoS*, 2014). A pesar de que esta tendencia varía de una disciplina a otra y entre investigadores particulares, en general son muchas las motivaciones (**Kim**; **Stanton**, 2016) y ventajas que trascienden a los mandatos o tendencias a la hora de compartir los datos (**Lyon**, 2016):

- aumenta las posibilidades de incrementar el impacto y la visibilidad de la investigación;
- potencia la reproducibilidad de la ciencia;
- ahorra costes a la hora de crear datos;
- fomenta la colaboración;
- contribuye a aumentar la credibilidad en el sistema.

Por supuesto, también son muchas las reticencias de los investigadores a la hora de compartir "sus" datos. El estudio de *Wiley* al que respondieron 2.886 investigadores (**Ferguson**, 2014) y que refleja **Alice Meadows** (2014) en su blog, pone de manifiesto algunas de ellas:

- miedo a las consecuencias negativas de compartir datos (mal uso, consecuencias legales o comerciales, etc.);
- falta de reconocimiento;
- carga de trabajo que supone preparar los datos para su publicación;
- desconocimiento de cómo y dónde compartir los datos.

> Para cada disciplina o dominio científico existe una interpretación de qué son conjuntos de datos o *datasets* de investigación, su naturaleza, cómo se recopilan y cómo se describen (metadatos)

1.1. Datos de investigación: un problema disciplinar visto desde las ciencias sociales y humanidades (CSH)

Para cada disciplina o dominio científico existe una interpretación de qué son datos o *datasets* de investigación, su naturaleza y cómo se recopilan. Y por supuesto, varía también la forma en que se describen esos datos (metadatos) y la problemática asociada al hecho de compartirlos. **Christine Borgman**, que ha tratado profusamente esta diferenciación y heterogeneidad disciplinar (**Borgman**, 2008; **Borgman**; **Wallis**; **Mayernik**, 2012) hace referencia al concepto de datos en ciencia y tecnología donde están más o menos claras las definiciones, e incluso el tipo de datos. Así, en este dominio se entiende por datos:

> "hechos, números, letras y símbolos que describen un objeto, idea, condición, situación u otros factores" a los que se unen "las manifestaciones digitales de literatura (incluyendo textos, sonidos, imágenes fijas, imágenes en movimiento, modelos, juegos o simulaciones)".

Por su parte, la *NSF* de EUA distingue: datos observacionales, datos computacionales y datos experimentales, pero partiendo de la naturaleza digital de todos ellos (**Borgman**; **Wallis**; **Mayernik**, 2012).

Sin embargo, en ciencias sociales y humanidades (CSH) no todos los datos se recopilan de forma digital, sino que adoptan otras muchas formas y formatos. Por ejemplo, en sociología los datos procedentes de encuestas y entrevistas pueden ser fácilmente captados digitalmente; no obstante, en arqueología el resultado de los datos observacionales estará más ligado al objeto y la información contextual sobre dicho objeto [coordenadas geográficas, muestras y dibujos del objeto sobre papel, o fotografías o vídeos (digitales)] (**Frank**; **Yakel**; **Faniel**, 2015).

Otro tema clave en CSH es el origen de los datos, ya que muchas investigaciones se basan en datos que no fueron producidos originalmente por o para la investigación. Este es el caso de los datos gubernamentales, documentos corporativos o estadísticas, etc., que, a su vez, son capaces de generar nuevos datos, es decir, son simplemente datos, pero se usan "para" la investigación y sirven para generar otros datos "de" investigación. Los humanistas dependen mucho más de fuentes de datos externos que los investigadores de otras disciplinas. Casi cada registro de la actividad humana puede considerarse "datos" (**Borgman**, 2008). Las CSH, en comparación con las ciencias más puras, generan muchos menos datos por observaciones puesto que tienden a utilizar datos procedentes, en general, de toda clase de fuentes: desde sonidos para estudios lingüísticos a películas para un análisis de objetos, vestidos o habla; hasta lo más clásico: libros, mapas, periódicos, diarios personales, fotografías, registros administrativos, etc., de tal forma que en muchas ocasiones los datos de investigación y las publicaciones pueden confundirse o entremezclarse.

La *National Endowment for the Humanities* de los EUA define los datos en este contexto disciplinar como los materiales generados o recopilados en el transcurso de una investigación, por ejemplo, citas, código de software, bases de datos, coordenadas geoespaciales, etc., informes y artículos, pero excluye expresamente borradores de artículos y las comunicaciones con colegas (*NEH*, 2015). Además, dentro del amplio espectro de materias y disciplinas que cubren las humanidades se pueden dar diversas definiciones de qué son datos, que complican aún más el panorama para su gestión y recuperación.

Quizá estas características particulares de las CSH es lo que lleva a que un 64% de investigadores de estas disciplinas, no compartan los datos en repositorios (**Meadows**, 2014). Pero se puede deber también al desconocimiento del dónde y cómo compartirlos, en unas fronteras tan difusas entre datos y publicaciones, y entre datos "de" investigación y "para" la investigación.

1.2. Metadatos o cómo hacer útiles los datos de investigación

A diferencia de lo que ocurre con las publicaciones donde, a pesar de diferentes estilos disciplinares, hay un núcleo común de propiedades formales en todas ellas, como ya hemos destacado, los datos científicos presentan una heterogeneidad que varía radicalmente entre disciplinas, campos temáticos e incluso hasta en grupos de investigación e investigadores.

La *NSF* en EUA pide que el *plan de gestión de datos* incluya los estándares de metadatos que se van a utilizar (**Bishoff**; **Johnston**, 2015). El piloto de datos abiertos (*ORD Pilot*) de la *European Commission* (2016) solicita además que se incluyan los metadatos asociados a los datos que son, en definitiva, los que hacen útiles esos datos. En el mundo de las bibliotecas digitales, los metadatos siempre han contribuido a hacer útiles los datos mediante la descripción de las publicaciones y otros objetos o activos digitales o digitalizados. Y en el mundo de los datos, son la única forma de hacerlos útiles: describiéndolos, dimensionándolos

y contextualizándolos para que se pueden encontrar, independientemente del silo disciplinar en el que se sitúen, y para que se puedan reutilizar a través de otros dominios. Sin los metadatos y las descripciones de los métodos de investigación y del contexto, los datos son sólo colecciones de números, listas de códigos, bonitas fotos o cajas de piedras (**Borgman**, 2008).

Las agencias de financiación están creando conciencia y haciendo presión sobre los investigadores para que gestionen sus datos, los compartan de una forma reutilizable, faciliten recuperación y además los preserven a largo plazo. Lo que implica (o debería de implicar) que los datos sean *FAIR* (localizables, accesibles, interoperables y reutilizables, por sus siglas en inglés: *findable*, *accessible*, *interoperable* y *reusable*). Crear datos *FAIR* y ciencia *FAIR* pone de manifiesto la necesidad de mejorar las e-infraestructuras para reutilizar la información científica (**Wilkinson** *et al.*, 2016), pero también la necesidad de fomentar la interoperabilidad a partir de los metadatos.

> Crear datos y ciencia *FAIR* pone de manifiesto la necesidad de mejorar las e-infraestructuras para reutilizar la información científica y la necesidad de fomentar la interoperabilidad a partir de los metadatos

Cuando los investigadores tienen que compartir sus metadatos, además de los datos, en un repositorio de datos, implica que deben traducir la metaiformación que utilizan en sus VREs (entornos virtuales de investigación), en sus servidores, o en sus ordenadores personales lo que **Tenopir** *et al.* (2015) llaman metadatos de laboratorio o metadatos específicos de una institución- al esquema/s de metadatos normalizados que utilice el repositorio en cuestión. **Tenopir** y su equipo de investigación encuestaron a más de 1.000 investigadores en cada uno de sus dos estudios, realizados en 2011 y 2015, sobre cómo gestionan sus datos (**Tenopir** *et al.*, 2011). Más del 50% declaró no utilizar ningún estándar de metadatos, el 14%, algún estándar dentro de su institución, y el 20% un estándar de su laboratorio (en el estudio de 2011); y más o menos en la misma proporción (47,9% ninguno y 16,7% estándar propio de laboratorio) en 2015. En nuestro estudio analizaremos los esquemas de metadatos que usan los repositorios, o al menos los que declaran usar sus administradores (en *re3data*) para describir los datos que depositan los investigadores en CSH.

2. Objetivos y metodología

De acuerdo con el contexto que reflejamos en el apartado anterior, este artículo se centra en dos dominios (ciencias sociales y humanidades) donde no hay tanta tradición como en otras disciplinas, ni a la colaboración, ni a la gestión de datos de investigación, ni (salvo algunas excepciones) a la utilización de esquemas de metadatos normalizados, ni de entornos virtuales de investigación u otras e-infraestructuras que requieran el uso de metadatos. Abordamos el problema de la gestión de datos científicos en CSH, a través del estudio de los repositorios de datos de estas disciplinas incluidos en *re3data* (un repositorio subvencionado por la *German Research Foundation*), para responder a las siguientes preguntas de investigación:

- ¿Qué tipo de datos son los que almacenan y gestionan los repositorios específicos de CSH?
- ¿Cómo es la distribución de los repositorios de datos de investigación en las diversas áreas de conocimiento dentro de CSH?
- ¿Cuáles son las áreas temáticas más representadas?
- ¿Qué esquemas de metadatos se utilizan en dichos repositorios para identificar y describir los distintos tipos de datos?
- ¿Hay algún modelo o esquema predominante en cada caso?

2.1. Objetivos

- Identificar los repositorios de datos de investigación en CSH.
- Estudiar qué tipos de datos resultan de las investigaciones en estas disciplinas, a través de los datos que almacenan los principales repositorios.
- Presentar los esquemas de metadatos más utilizados en dichos repositorios.

Se trata de un estudio exploratorio para identificar los repositorios especializados más representativos de CSH, investigar sus prácticas y verificar el tipo de datos que almacenan, así como los esquemas de metadatos que utilizan o declaran utilizar.

2.2. Metodología

Para el análisis hemos utilizado el citado *re3data* (*Registry of Research Data Repositories*) como fuente, ya que se trata del registro de repositorios de datos de referencia, recomendado tanto por la *Comisión Europea* (*European Commission*, 2016), como por diversas editoriales (*PeerJ*, *Springer*, *Nature's Scientific Data*, etc.). Ese registro permite la fácil identificación de repositorios de datos en donde depositar datos de investigación por temas o disciplinas.

Inicialmente se planteó una metodología cuantitativa y analítica que permitiera analizar los 397 repositorios incluidos en la categoría de CSH de *re3data*. Sin embargo, en el transcurso del trabajo se decidió cambiar y realizar un estudio pormenorizado de una muestra reducida de repositorios especializados, los más representativos: 3 de ciencias sociales y 3 de humanidades, para corroborar la correspondencia del uso que se hacía de los esquemas de metadatos declarados en el registro. Así pues, el trabajo se llevó a cabo en tres fases:

a) Extracción y tratamiento de los registros de *re3data*

En esta fase se realizaron varias tareas:

a.1. Consulta y extracción, a través de la API que ofrece *re3data*, de un total de 1.457 repositorios registrados en el momento de la toma de datos (18 de febrero de 2016). Nótese que, en abril de 2016, *re3data* anunció que ya ha llegado a 1.500 registros de repositorios de datos. Aunque la última versión del esquema descriptivo (metadatos) de *re3data* es la 3.0 (**Rücknagel** *et al.*, 2015), la API responde, en realidad, a la primera versión de dicho esquema, mucho más limitada que la última versión.

a.2. A partir del listado de repositorios, se descargan en formato xml los 1.457 registros que describen los repositorios, utilizando técnicas de *scraping* con *R*[1].

a.3. Tratamiento de los registros a través de xslt[2] para procesar la información que interesaba para este estudio, fundamentalmente: tipos de datos y de metadatos utilizados por los repositorios, así como sus esquemas de clasificación e identificación.

b) Selección de la muestra y análisis cuantitativo de los datos extraídos

El objetivo de esta fase era filtrar los repositorios de datos en los que queríamos centrar el estudio, aquellos con contenidos de CSH. Se seleccionaron sólo los que en su descripción contuvieran algún esquema de clasificación temática de humanidades o ciencias sociales, de acuerdo con la clasificación que utiliza *re3data* que se puede ver en la tabla 1.

En el caso de la clasificación temática, cabe aclarar que, para clasificar un repositorio por el esquema de metadatos, *re3data* proporciona la propiedad *SubjectScheme* como un atributo obligatorio que permite introducir cuantos valores se quieran, teniendo siempre en cuenta que los únicos valores permitidos son los procedentes de la clasificación temática de la *German Research Foundation* (*DFG Classification of subject area*). Esta clasificación contempla cuatro grandes áreas:

- *humanities and social sciences*
- *life sciences*
- *natural sciences*
- *engineering sciences.*

Hay que considerar que cada repositorio se describe a sí mismo con cuantos temas crea que cubre, por lo que, a la hora de representarlos, un mismo repositorio puede aparecer en más de un área temática y hasta en las cuatro grandes áreas temáticas, como es el caso de los multidisciplinares.

Tabla 1. Número de repositorios del área específica de CSH, incluyendo los multidisciplinares, de acuerdo con la clasificación *DFG*

SubjectScheme* de *DFG	**Número de repositorios**
1 Humanities and social sciences	397
101 Ancient cultures	15
102 History	34
103 Fine arts, music, theatre and media studies	26
104 Linguistics	47
105 Literary studies	10
106 Non-European languages and cultures, social and cultural anthropology, Jewish studies and religious studies	18
107 Theology	4
108 Philosophy	3
109 Education sciences	146
110 Psychology	14
111 Social sciences	155
112 Economics	114
113 Jurisprudence	27

En el filtrado obtuvimos 397 registros que podían ser específicos de un área concreta dentro de CSH o multidisciplinares, pero con un énfasis especial en CSH, y que constituirían nuestra muestra de estudio para analizar los tipos de datos y los esquemas de metadatos que declara utilizar el responsable de cada repositorio al completar los datos del registro.

> La idiosincrasia particular de los investigadores de ciencias sociales y humanidades es lo que lleva a que muy pocos de ellos compartan sus datos, pero esto se puede deber también al desconocimiento del dónde y cómo compartirlos

Con respecto al tipo de dato, para identificarlo se tomó la propiedad *ContentType* (no obligatoria) del esquema de *re3data*, que permite especificar todos los tipos de contenido disponibles en un repositorio. Los valores permitidos en este campo quedan restringidos a los tipos de contenido reconocidos e identificados en el proyecto *Parse.insight* (*Permanent Access to the Records of Science in Europe*). La clasificación *Parse* presenta 15 opciones:

- *Archived data*
- *Audiovisual data*
- *Configuration data*
- *Databases*
- *Images*
- *Network based data*
- *Plain text*
- *Raw data*
- *Scientific and statistical data formats*
- *Software applications*
- *Source code*
- *Standard office documents*
- *Structured graphics*
- *Structured text*
- *Other*.

Sin embargo, no es obligatorio seleccionar una de ellas a la hora de completar el registro sobre el repositorio.

Por último, para identificar el esquema de metadatos, tomamos la propiedad *MetadataStandardName* del esquema, que tampoco es obligatoria.

c) Identificación de un subconjunto de repositorios de datos clave en CSH

Identificado el subconjunto, se hizo un análisis cualitativo e individual de los esquemas de metadatos que realmente utilizan.

Esta última fase de la metodología, se incluyó al darnos cuenta de las limitaciones de *re3data* en dos sentidos:

- completar/declarar el esquema de metadatos que utiliza el repositorio no es obligatorio;

- la información sobre el estándar utilizado responde al momento en que se completó el registro y puede haber cambiado con el paso del tiempo.

Al acceder a los repositorios se añadieron otras dificultades, como:

- posibilidad de corroborar los esquemas de metadatos declarados en *re3data*;
- acceso restringido a usuarios autorizados, en algunos casos;
- ausencia de manuales o bibliografía, etc., en otros.

Así pues, para completar el estudio se decidió seleccionar 3 repositorios en ciencias sociales y 3 en humanidades en base a:

- cobertura o número de *datasets* que albergan;
- uso que hacen de ellos sus respectivas comunidades;
- representatividad para este estudio, que cubrieran, además, varios temas y países.

En el caso de las humanidades se seleccionó un repositorio de lingüística (*Clarin*), uno de arqueología, y otro de historia y arte (*Prometheus*) por ser las subdisciplinas con más repositorios de datos identificados en *re3data*. Los repositorios seleccionados se muestran en la tabla 2.

Tabla 2. Selección de repositorios representativos (CSH)

Ciencias sociales	
Inter university Consortium for Political and Social Research (*ICPSR*, EUA)	*http://www.icpsr.umich.edu*
UK Data Service (Reino Unido)	*https://www.ukdataservice.ac.uk*
Gesis Zacat (Alemania)	*http://zacat.gesis.org/webview*
Humanidades	
Common Language Resources and Technology Infrastructure (*Clarin*, EU):	*http://www.clarin.eu*
Archaeology Data Service (Reino Unido)	*http://archaeologydataservice.ac.uk*
Prometheus (Alemania)	*http://www.prometheus-bildarchiv.de*

3. Resultados y discusión

3.1. Repositorios de datos de investigación en CSH

Un primer panorama de los repositorios existentes en CSH registrados en *re3data*, se puede ver la figura 1: un *treemap* que representa el número/volumen de repositorios de las áreas estudiadas de acuerdo con la sub-clasificación de CSH de la tabla 1.

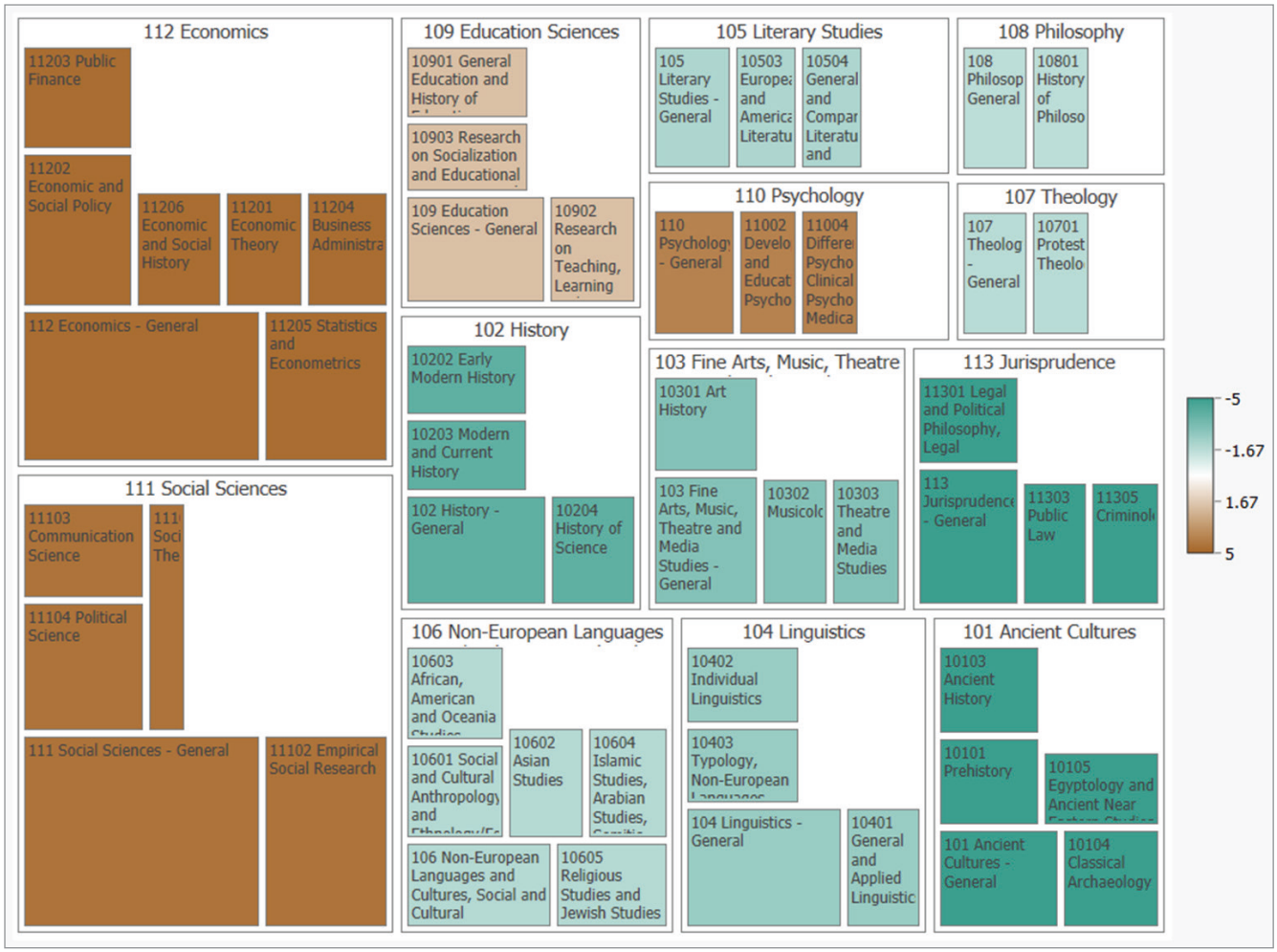


Figura 1. Representación proporcional de los repositorios de CSH de *re3data*, incluyendo los multidisciplinares.

Tabla 3. Esquemas de metadatos en los repositorios representativos de CSH

Repositorio	Esquema de metadatos
Ciencias sociales	
Inter university Consortium for Political and Social Research ICPSR (EUA) *http://www.icpsr.umich.edu*	DDI DC
UK Data Service (Reino Unido) *https://www.ukdataservice.ac.uk*	DDI, DC, ISO 19115, METS (*Metadata encoding and transmission standard*), ISAD (*International standard archival description*)
Gesis Zacat (Alemania) *http://zacat.gesis.org/webview*	DDI DC
Humanidades	
Common Language Resources and Technology Infrastructure (*Clarin*, EU) *http://www.clarin.eu*	IMDI (*ISLE meta data initiative*), TEI headers, DC, DCTerms, DC-OLAC (*Open language archive community*) (**Van-Uytvanck**; **Stehouwer**; **Lampen**, 2012)
Archaeology Data Service (Reino Unido) *http://archaeologydataservice.ac.uk*	ADS Schema DC MIDAS
Prometheus (Alemania) *http://www.prometheus-bildarchiv.de*	EDM (*Europeana data model*) METS DC

Con el objeto de dar una visión más precisa, se adjunta una tabla donde se indica el número de repositorios de acuerdo con la clasificación y el *SubjectScheme* de *DFG* que utiliza *re3data*, que contempla cuatro niveles. En este caso, en la tabla 1, se indica hasta el tercer nivel.

Ha de tenerse en cuenta que un repositorio multidisciplinar puede estar en más de una categoría, por lo que la suma de las partes es superior al total. De acuerdo con la clasificación temática de la *DFG*, las ciencias sociales (códigos del 109 al 113) cuentan con una representación mayoritaria, 456 repositorios frente a los 157 de humanidades.

3.2. Datos de investigación en CSH

Los tipos de contenido disponibles en *re3data* fueron representados según los tipos reconocidos e identificados en el proyecto *Parse.insight*. El tipo de datos científicos y estadísticos (formatos como spss, fits, gis, etc.) junto con los documentos tipo *Office* (*Word*, *Excel* o formatos parecidos de *OpenOffice*) y de imágenes (jpeg, jpeg2000, gif, tif, png, svg, etc.) son los más utilizados en proyectos de digitalización en humanidades. En la figura 2 se puede advertir que la proporción de tipos de contenidos es relativamente equilibrada en todas las áreas científicas respecto del uso que se hace en humanidades y ciencias sociales. Para cada tipo de dato (datos científicos, imágenes, texto plano, datos brutos, etc.) el uso que se hace en CSH está casi siempre alrededor del 27% (un mínimo de 20% y un máximo de 32%). No importa el tipo de contenido, siempre está alrededor de ese porcentaje. Resulta curioso porque cabía pensar que en esas materias habría más documentos tipo "standard office documents" que "scientific and statistical data formats" o "raw data", por ejemplo. Para todos los tipos de documento se da más o menos la misma proporción, aproximadamente (73% en otras disciplinas y 27% en CSH). Incluso en «audiovisual data», casi al final del gráfico, que son muchos menos que

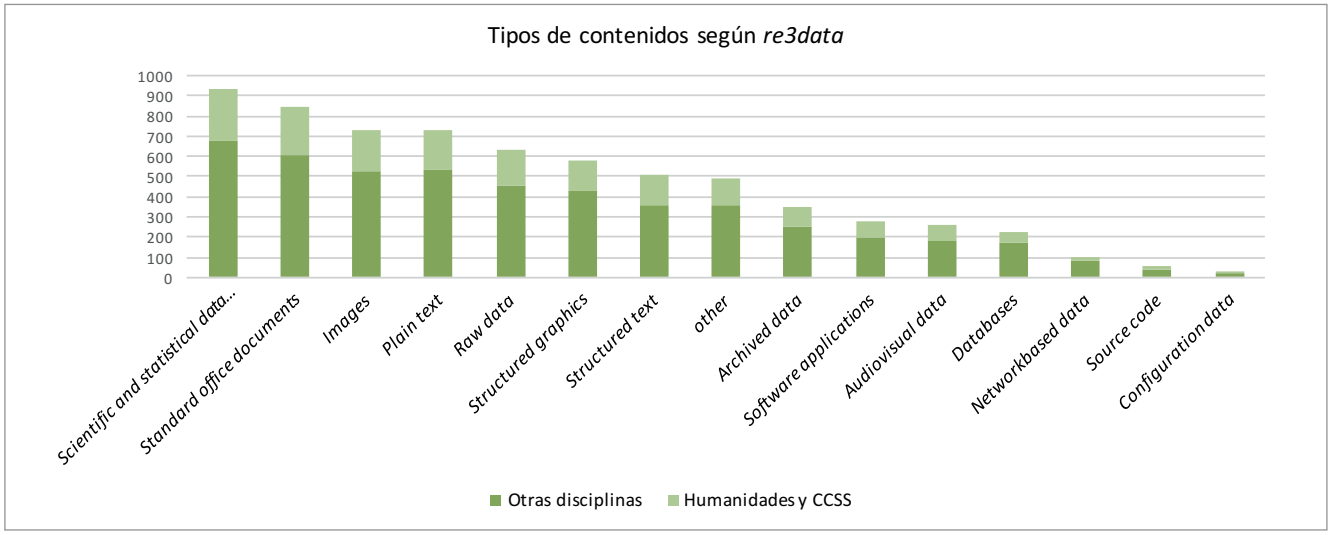


Figura 2. Tipos de contenidos declarados en *re3data*. En el eje de ordenadas consta el número de repositorios en que se encuentra cada tipo de contenido.

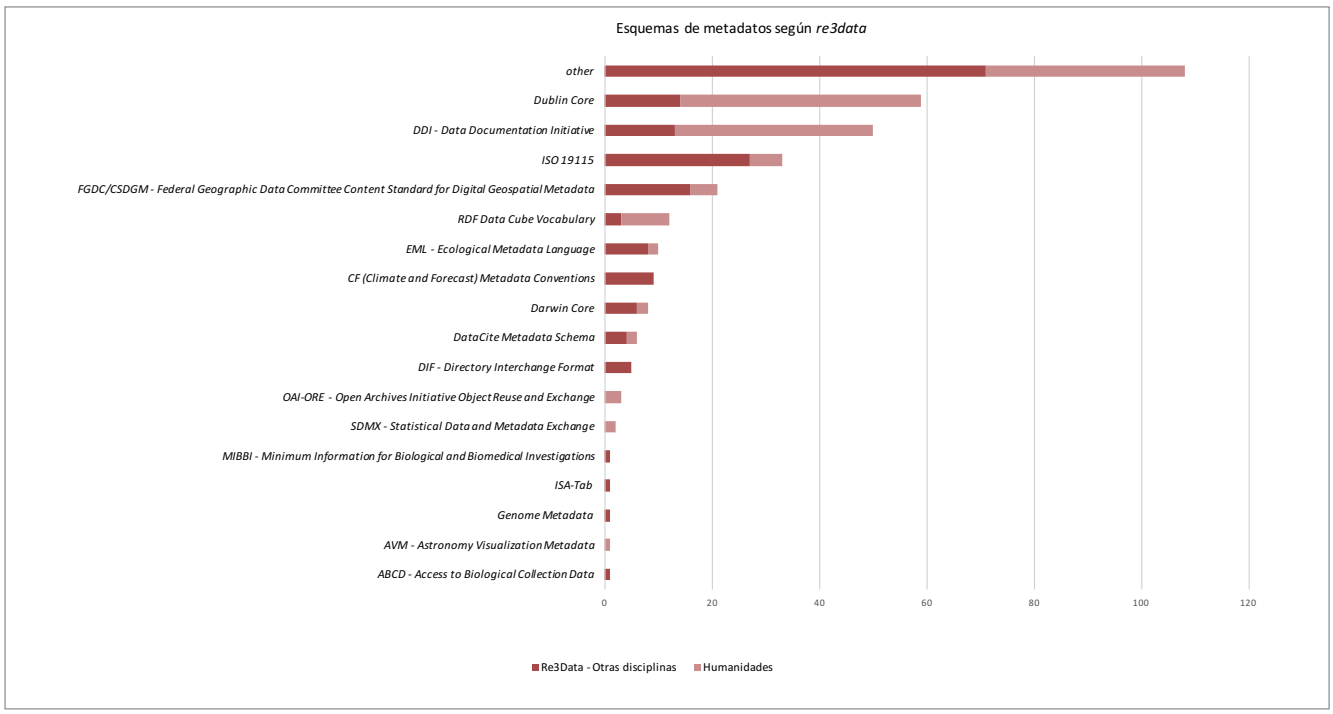


Figura 3. Esquemas de metadatos que declaran los repositorios de datos incluidos en *re3data*

los «standard office documents», se mantiene la proporción (69,3% en otras disciplinas y 30,7% en CSH).

3.3. Esquemas de metadatos utilizados en los repositorios de datos de investigación de CSH

Un 22,8% (332) del total de repositorios *re3data* especifican el/los esquema/s de metadatos que utilizan. Como se observa en la figura 3, en el campo de las CSH (en color más claro) *Dublin Core* y DDI (*Data documentation initiative*) son, con diferencia, los más utilizados. La razón que explica que "otros" sea el valor más destacado es que se trata de un campo no obligatorio en todas las versiones del esquema *re3data*, e indica la gran variedad de metadatos que se utilizan en todas las disciplinas, con unos pocos esquemas dominantes en ciertas áreas, y muchas variaciones específicas en aquellas disciplinas en las que ningún esquema se erige como dominante.

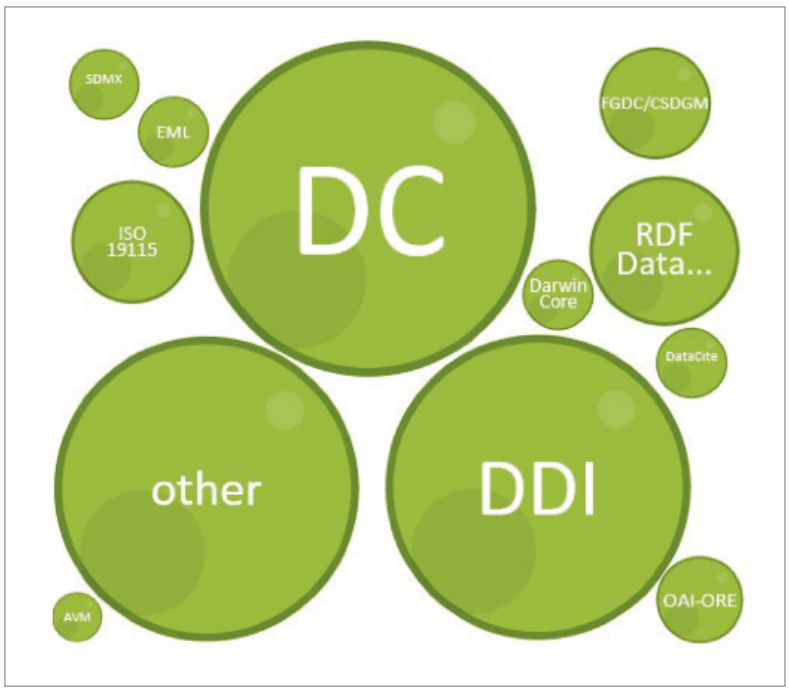
Por otra parte, un 25,2% de los repositorios de CSH declaran trabajar con algún tipo de esquema de metadatos. De ellos, un 45% utiliza *Dublin Core*, el modelo de metadatos más frecuente, en 45 repositorios. DDI y otros, ocupan el segundo lugar con un 37% para cada uno, utilizados en 37 archivos. Cabe aclarar, que "otros" se refiere a esquemas de metadatos propios (de la institución o de laboratorio). Tanto la representación gráfica de la situación como los nombres de los esquemas de metadatos y el número de repositorios que los utilizan se puede ver en la figura 4. Es de destacar que un 74,8% de los repositorios no suministran este dato, y que algún repositorio multidisciplinar utiliza más de un esquema, por lo que es posible encontrar esquemas propios de otras áreas científicas.

Con el objetivo de revisar los esquemas de metadatos que se utilizan en CSH, se seleccionaron 6 repositorios representativos (tabla 2). Estudiamos cada uno identificando el esquema de metadatos utilizado, bien analizando el repositorio, bien buscando en el sitio del repositorio la guía o instrucciones para el depósito, o buscando trabajos sobre los repositorios estudiados donde se declara esta información.

> La heterogeneidad y complejidad de los repositorios de datos de investigación se manifiesta sobre todo en los esquemas de metadatos que se eligen para describirlos, que es aún más evidente en los de humanidades

El análisis pormenorizado de los repositorios particulares seleccionados confirma la tendencia que nos ofrece *re3data* en ciencias sociales: el esquema de metadatos preponderante es DDI (*Data documentation initiative*), un estándar internacional para describir los datos estadísticos y datos de ciencias sociales con mucha tradición. DDI describe los datos que resultan de los métodos de observación en las ciencias sociales, conductuales, económicos y de la salud. Contempla los procesos de recolección de los datos, niveles variables de descripción y métodos. Es un esquema que podríamos denominar clásico, ya que se originó en 1995, cuando el *Dublin Core*, pero dentro de la comunidad de las ciencias sociales, y con el objetivo de describir datos. Desde entonces ha ido evolucionando de manera sostenida, mantenido por la *DDI Aliance* (**Vardigan**, 2013).
http://www.ddialliance.org

En el caso de las humanidades, los esquemas de metadatos utilizados son más diversos y particulares, según muestran los repositorios seleccionados analizados (tabla 3); tanto que la mayoría de los esquemas encontrados no están dentro de



Esquema	Nº
Dublin Core	45
DDI	37
Other	37
RDF Data cube vocabulary	9
ISO 19115	6
FGDC/CSDGM	5
OAI-ORE	3
EML	2
Darwin Core	2
DataCite	2
SDMX	2
AVM	1

Figura 4. Esquemas de metadatos utilizados en CSH según *re3data*

los registrados en *re3data*. Quizá allí encontramos la justificación del alto porcentaje para la categoría "otros" (37%) en los repositorios de humanidades, ya que hasta la versión 3.0 del esquema de *re3data* sólo se reconocían como valores permitidos los estándares de metadatos recogidos por el *Digital Curation Centre*.
http://www.dcc.ac.uk/resources/metadata-standards

> *Re3data* (*Registry of Research Data Repositories*) es la fuente de referencia para identificar repositorios donde depositar datos de investigación por temas o disciplinas

Esta diversidad de metadatos, o la falta de estándares comunes o regulares en humanidades, se justifica en la propia heterogeneidad de los repositorios de datos y a lo que se considera "datos" dentro del conjunto de las humanidades, tal y como discutíamos en la introducción.

Una constante dentro de las CSH es el uso de *Dublin Core* (DC) (figura 4). Esta predominancia se debe a nuestro juicio:

- a la vinculación con los repositorios de documentos/publicaciones, y a la falta de distinción en muchos casos aún entre éstos y los repositorios de datos, y
- al nivel de normalización que ha adquirido el DC, al ser el DC simple, la base para la interoperabilidad OAI-PMH entre repositorios.

A esto además se une el argumento que compartimos de **Willis**, **Greenberg** y **White** (2012): los creadores de esquemas de metadatos son más proclives a modificar y adaptar o enriquecer un esquema existente que crear uno nuevo. El DC se ha instalado y es más sencillo adaptarlo que modificarlo.

4. Conclusiones

La principal conclusión que extraemos de este estudio es la corroboración de la heterogeneidad y complejidad de los repositorios de datos de investigación, que aún es más palmaria en disciplinas pertenecientes a las humanidades. Esta heterogeneidad se manifiesta, sobre todo, en los esquemas de metadatos que los investigadores eligen para describirlos. No obstante, señalamos a continuación las conclusiones particulares a las que hemos llegado en el transcurso de este trabajo.

1) Tras la fusión entre *Databib* y *re3data* en un mismo registro a finales de 2015, *re3data* se ha convertido, sin duda, en el registro por antonomasia en donde encontrar los repositorios de datos de investigación en todas las disciplinas, donde hemos identificado y analizado 397 repositorios en CSH. Su mayor deficiencia, por ahora, es que carece de mecanismos que permitan saber cuándo se modifica el registro de un repositorio de datos o cambian las características inicialmente declaradas. La información sobre los repositorios no se puede actualizar online. Desde febrero de 2016 este problema intenta paliarse remitiendo a *re3data* un formulario de petición de cambios. Un mecanismo manual que esperemos que sea provisional.

El esquema de metadatos que utiliza *re3data* en su versión actual (v. 3.0) para describir los repositorios, incorpora algunas características sobre reutilización, métricas y políticas. Este modelo parece evolucionar en la buena dirección si no tienden a aumentar mucho más las propiedades ya existentes, y logran los mecanismos de automatización, como ocurre ahora con los repositorios de publicaciones y los agregadores.

La clasificación temática de la DFG que utiliza *re3data* resulta demasiado genérica, por lo que no es sencillo acotar la temática de cada repositorio ya que la gran mayoría se declaran multidisciplinares cuando muchas veces no lo son, o lo son en una proporción muy pequeña.

2) Teniendo en cuenta las limitaciones de *re3data*, para describir los repositorios, los datos y los esquemas de metadatos son menos homogéneos en humanidades que en ciencias sociales. A pesar del escaso número de repositorios de datos que declaran el estándar de metadatos que utilizan para describirlo, en ciencias sociales se confirma la tendencia que muestra *re3data* de utilización del esquema de metadatos DDI. Esto se puede deber a la madurez del estándar, a su nivel de implementación y a que fue un esquema originalmente creado para describir datos, no documentos. Al igual que ocurre con los sistemas de información digital geoespacial, donde desde mediados de los años 90 se utilizan los estándares *FGDC* (*Federal Geographic Data Committee*) y la *ISO 19115* para describir infraestructuras de datos geoespaciales.

La adopción de DDI por algunos de los repositorios más importantes como *ICPSR*, *Gesis* y la red de repositorios de datos de *Dataverse* le auguran un futuro destacado entre los estándares de metadatos en ciencias sociales, donde los datos "de" (y "para") investigación son compatibles con estadísticas, encuestas, sondeos de opinión, etc., a los que se dirige desde su creación el estándar DDI.

3) En el campo de las humanidades la situación se presenta más compleja y diversa. *Dublin Core* (DC) parece ser muy utilizado según los datos genéricos extraídos de *re3data*, pero si descendemos al detalle de repositorios de datos de campos específicos como la lingüística o la arqueología, se muestra el uso de esquemas propios o la adaptación, en mayor o menor medida, del DC. Es necesario destacar también que muchos proyectos de humanidades, especialmente los de digitalización de textos, utilizan *TEI Header*, vinculado al estándar TEI (*Text encoding initiative*), mientras que en otros casos carecen de esquemas de descripción. La exposición de sus datos de investigación se hace simplemente a través de gestores de contenidos sin apenas uso de metadatos. No sorprende tampoco que aparezcan como esquemas de metadatos para describir datos humanísticos en sus repositorios de datos, estándares de creación de bibliotecas digitales y de descripción de publicaciones textuales, de imágenes o audiovisuales (no sólo DC, sino también EDM, METS, MIDAS), teniendo en cuenta la tenue diferenciación en algunas de estas disciplinas, entre datos y documentos y también entre datos "de" y "para" la investigación.

4) El *Dublin Core* (DC) es el estándar por defecto en los repositorios de publicaciones, y esta tendencia se arrastra a los repositorios de datos, al menos en una primera instancia o aproximación. Aunque el DC tiene perfectamente establecido los mecanismos para crear perfiles de aplicación que se adapten a la descripción de un tipo de información o colección particular, aún es pronto para corroborar si este estándar se puede adaptar a la idiosincrasia disciplinar de los datos de investigación.

Notas

1. *Web scraping* (recolección en la web o extracción de datos) es una técnica informática para obtener información de webs.

R es un lenguaje de programación y software para computación y gráficos estadísticos mantenido por la *R Foundation for Statistical Computing*.

2. Xslt (*extensible stylesheet language transformations*) es un lenguaje para transformar documentos xml en otros documentos xml u otros formatos como html para páginas web, texto plano o en objetos de formato xsl, que posteriormente se pueden convertir a otros formatos, como pdf, postscript o png.

Agradecimientos

Este trabajo forma parte del proyecto *Curator-e: custodia y gestión digital para la reutilización de datos abiertos de investigación en humanidades y ciencias sociales*, financiado por el *Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad* (*Mineco*, España) (CSO2013-46754-R).

5. Bibliografía

Bishoff, Carolyn; **Johnston, Lisa** (2015). "Approaches to data sharing: An analysis of NSF data management plans from a large research university". *Journal of librarianship and scholarly communication*, v. 3, n. 2, p. eP1231.
http://dx.doi.org/10.7710/2162-3309.1231

Borgman, Christine L. (2008). "Data, disciplines, and scholarly publishing". *Learned publishing*, v. 21, n. 1, pp. 29-38.
http://dx.doi.org/10.1087/095315108X254476

Borgman, Christine L.; **Wallis, Jillian C.**; **Mayernik, Matthew S.** (2012). "Who's got the data? Interdependencies in science and technology collaborations". *Computer supported cooperative work* (*CSCW*), v. 21, n. 6, pp. 485-523.
http://nldr.library.ucar.edu/repository/assets/osgc/OSGC-000-000-012-014.pdf
http://dx.doi.org/10.1007/s10606-012-9169-z

COM (2016) 178 final. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: European cloud initiative - Building a competitive data and knowledge economy in Europe.
http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=15266

European Commission (2016). *Guidelines on open access to scientific publications and research data in Horizon 2020, v. 2.1*. European Commission. Directorate General for Research and Innovation.
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Ferguson, Liz (2014). "How and why researchers share data (and why they don't)". *Wiley Exchanges*. *Discover the future of research*, 3 November.
https://hub.wiley.com/community/exchanges/discover/blog/2014/11/03/how-and-why-researchers-share-data-and-why-they-dont?referrer=exchanges

Frank, Rebecca D.; **Yakel, Elizabeth**; **Faniel, Ixchel M.** (2015). "Destruction/reconstruction: preservation of archaeological and zoological research data". *Archival science*, v. 15, n. 2, pp. 141-167.
http://dx.doi.org/10.1007/s10502-014-9238-9

Kim, Youngseek; **Stanton, Jeffrey M.** (2016). "Institutional and individual factors affecting scientists' data-sharing be-

haviors: A multilevel analysis". *Journal of the Association for Information Science and Technology*, v. 67, n. 4, pp. 776-799.
https://www.asis.org/asist2013/proceedings/submissions/papers/123paper.pdf
http://dx.doi.org/10.1002/asi.23424

LERU (2013). *LERU roadmap for research data*. Advice paper n. 14.
http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

Lyon, Liz (2016). "Transparency: the emerging third dimension of open science and open data". *Liber quarterly*, v. 25, n. 4.
http://dx.doi.org/10.18352/lq.10113

Meadows, Alice (2014). "To share or not to share? That is the (research data) question…". *The scholarly kitchen*, 11 November.
http://scholarlykitchen.sspnet.org/2014/11/11/to-share-or-not-to-share-that-is-the-research-data-question

NEH (2015). *Data management plans for NEH Office of Digital Humanities. Proposals and awards*.
http://www.neh.gov/files/grants/data_management_plans_2015.pdf

NIH (2015). "NIH sharing policies and related guidance on NIH-funded research resources". *National Institutes of Health.*
https://grants.nih.gov/policy/sharing.htm

NSF (2014). "Chapter II. Proposal preparation instructions". *Grant proposal guide.* National Science Foundation. Where discoveries begin.
http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp

OECD (2015). *Making open science a reality*. Organisation for Economic Co-operation and Development.
https://www.innovationpolicyplatform.org/sites/default/files/DSTI-STP-TIP%282014%299-REV2_0_0_0_0.pdf

Paltoo, Dina N.; **Rodriguez, Laura-Lyman**; **Feolo, Michael**; **Gillanders, Elizabeth**; **Ramos, Erin M.**; **Rutter, Joni L.**; **Sherry, Stephen**; **Wang, Vivian-Ota**; **Bailey, Alice**; **Baker, Rebecca**; **Caulder, Mark**; **Harris, Emily L.**; **Langlais, Kristofor**; **Leeds, Hilary**; **Luetkemeier, Erin**; **Paine, Taunton**; **Roomian, Tamar**; **Tryka, Kimberly**; **Patterson, Amy**; **Green, Eric D.** (2014). "Data use under the NIH GWAS data sharing policy and future directions". *Nature genetics*, v. 46, n. 9, pp. 934-938.
http://dx.doi.org/10.1038/ng.3062

Pepe, Alberto; **Goodman, Alyssa**; **Muench, August**; **Crosas, Merce**; **Erdmann, Christopher** (2014). "How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers". *PLoS one*, v. 9, n. 8, p. e104798.
http://dx.doi.org/10.1371/journal.pone.0104798

PLoS (2014). "PLoS data policy prior to March 3, 2014". *PLoS*.
http://goo.gl/QIRIab

Rücknagel, Jessika; **Vierkant, Paul**; **Ulrich, Robert**; **Kloska, Gabriele**; **Schnepf, Edeltraud**; **Fichtmüller, David**; **Reuter, Evelyn**; **Semrau, Angelika**; **Kindling, Maxi**; **Pampel, H.**; **Witt, Michael**; **Fritze, Florian**; **Van-de-Sandt, Stephanie**; **Klump, Jens**; **Goebelbecker, Hans-Jürgen**; **Skarupianski, Michael**; **Bertelmann, Roland**; **Schirmbacher, Peter**; **Scholze, Frank**; **Kramer, Claudia**; **Fuchs, Claudio**; **Spier, Shaked**; **Kirchhoff, Agnes** (2015). *Metadata schema for the description of research data repositories*, v. 3.0.
http://dx.doi.org/10.2312/re3.008

Tenopir, Carol; **Allard, Suzie**; **Douglass, Kimberly**; **Aydinoglu, Arsev-Umur**; **Wu, Lei**; **Read, Eleanor**; **Manoff, Maribeth**; **Frame, Mike** (2011). "Data sharing by scientists: Practices and perceptions". *PLoS one*, v. 6, n. 6.
http://dx.doi.org/10.1371/journal.pone.0021101

Tenopir, Carol; **Dalton, Elizabeth D.**; **Allard, Suzie**; **Frame, Mike**; **Pjesivac, Ivanka**; **Birch, Ben**; **Pollock, Danielle**; **Dorsett, Kristina** (2015). "Changes in data sharing and data reuse practices and perceptions among scientists worldwide". *PLoS one*, v. 10, n. 8, p. e0134826.
http://dx.doi.org/10.1371/journal.pone.0134826

Van-Uytvanck, Dieter; **Stehouwer, Herman**; **Lampen, Lari** (2012). "Semantic metadata mapping in practice: the virtual language observatory". En: *LREC 2012: 8th Intl conf. on language resources and evaluation. European Language Resources Association* (*ELRA*), pp. 1029-1034.
http://goo.gl/IMgP4c

Vardigan, Mary (2013). "Timeline DDI". *Iassist quarterly*, v. 37, pp. 51-55.
http://www.iassistdata.org/sites/default/files/iq/iqvol371_4_vardigan2.pdf

Wilkinson, Mark D.; **Dumontier, Michel**; **Aalbersberg, Ijsbrand-Jan**; **Appleton, Gabrielle**; **Axton, Myles**; **Baak, Arie**; **Blomberg, Niklas**; **Boiten, Jan-Willem**; **Da-Silva-Santos, Luiz-Bonino**; **Bourne, Philip E.**; **Bouwman, Jildau**; **Brookes, Anthony J.**; **Clark, Tim**; **Crosas, Mercè**; **Dillo, Ingrid**; **Dumon, Olivier**; **Edmunds, Scott**; **Evelo, Chris T.**; **Finkers, Richard**; **González-Beltrán, Alejandra**; **Gray, Alasdair J.G.**; **Groth, Paul**; **Goble, Carole**; **Grethe, Jeffrey S.**; **Heringa, Jaap**; **Hoen, Peter A.C't**; **Hooft, Rob**; **Kuhn, Tobias**; **Kok, Ruben**; **Kok, Joost**; **Lusher, Scott J.**; **Martone, Maryann E.**; **Mons, Albert**; **Packer, Abel L.**; **Person, Bengt**; **Rocca-Serra, Philippe**; **Roos, Marco**; **Van-Schaik, Rene**; **Sansone, Susanna-Assunta**; **Schultes, Erik**; **Sengstag, Thierry**; **Slater, Ted**; **Strawn, George**; **Swertz, Morris A.**; **Thompson, Mark**; **Van-der-Lei, Johan**; **Van-Mulligen, Erik**; **Velterop, Jan**; **Waagmeester, Andra**; **Wittenburg, Peter**; **Wolstencroft, Katherine**; **Zhao, Jun**; **Mons, Barend** (2016). "The FAIR guiding principles for scientific data management and stewardship". *Scientific data*, v. 3, p. 160018.
http://dx.doi.org/10.1038/sdata.2016.18

Willis, Craig; **Greenberg, Jane**; **White, Hollie** (2012). "Analysis and synthesis of metadata goals for scientific data". *Journal of the American Society for Information Science and Technology*, v. 63, n. 8, pp. 1505-1520.
http://dx.doi.org/10.1002/asi.22683