Repositorios digitales en España y calidad de los metadatos

Digital repositories in Spain and metadata quality

José Federico Medrano (1), Carlos G. Figuerola (2), José Luis Alonso Berrocal (3)

(1) Universidad de Salamanca, Departamento de Informática y Automática, Francisco de Vitoria, 6-16, 37008 Salamanca (España), fmedrano@usal.es. (2, 3) Universidad de Salamanca. Instituto Universitario de Estudios de la Ciencia y Tecnología
(2) figue@usal.es. (3) berrocal@usal.es

Resumen

Los repositorios digitales han experimentado un notable desarrollo en los últimos años, en especial a partir de la formulación de la Open Archives Initiative (OAI) y de la aparición de instrumentos como el protocolo Open Archives Initiative Protocol for Metada Harvesting (OAI-PMH). En este artículo se analizan los metadatos de los 69 repositorios españoles accesibles mediante dicho protocolo y se realiza un estudio cuantitativo y cualitativo de los datos contenidos en dichos repositorios. En especial, se analizan factores que inciden directamente en la calidad de dichos datos, identificando los problemas más frecuentes. Los ocho campos Dublin Core más utilizados fueron *title, identifier, date, language, format, description, type* y *subject* (difiriendo en *language* y *type*). *Creator* solo es cumplimentado en un 56%. *Subject* también es descuidado. También lo es el control de la forma de los metadatos, que resulta clave para mejorar el acceso. Los repositorios más grandes suelen ofrecer más problemas.

Palabras clave: Repositorios digitales. OAI. PMH. Metadatos. Harvesting. Calidad.

Abstract

Digital repositories have experimented a significant development in recent years, from the raising of the Open Archives Initiative (OAI) to the emergence of tools as the Open Archives Initiative Protocol for Metada Harvesting (OAI-PMH). In this paper we analyze metadata from the 69 Spanish repositories available through such protocol. Also, a qualitative and quantitative analysis is performed on the data contained in those repositories. Specifically, we analyze factors that directly affect the quality of the data, identifying the most common problems. The eight most used Dublic Core fields were *title, identifier, date, language, format, description, type* and *subject*. *Creator* was filled only in 56% and *subject* is frequently missed, and also vocabulary control in general, a key aspect to offer good search results. Biggest repositories usually offer worse quality.

Keywords: Digital repositories. OAI. PMH. Metadata. Harvesting. Quality.

1. Introducción

Los repositorios digitales han experimentado un notable desarrollo en los últimos años; ya sea como instrumento de preservación digital, ya sea como herramienta para la difusión de los resultados de la investigación científica, el número de repositorios y la cantidad de registros (documentos) de éstos ha crecido considerablemente. El origen de este movimiento puede situarse en la Convención de Santa Fe, en 1999, y en la formulación de la *Open Archives Initiative* (OAI) (Ginsparg, Luce y Van de Sompel, 1999; Van de Sompel y Lagoze, 2000). La OAI tiene como misión desarrollar y promover estándares de interoperabilidad para facilitar la difusión eficiente de contenidos en internet y el descubrimiento de contenido en archivos distribuidos. Por ello, el OAI-PMH está diseñado para facilitar las transferencias incrementales de metadatos de una manera simple y general en contraposición a la búsqueda remota de metadatos (Webley, Chipeperekwa y Suleman, 2011).

En los años siguientes, el ámbito de la OAI fue extendido para incorporar no sólo lo relativo a publicaciones científicas, sino también lo relativo a cualquier material almacenado en soporte electrónico (Lagoze y Van de Sompel, 2003; Warner, 2003). En el año 2001, el OAI-PMH vio la luz (Lagoze y Van de Sompel, 2001), pero no fue hasta Junio del año 2002, luego de un periodo largo de evaluaciones, definiciones y modificaciones, cuando se puso en producción la versión 2.0 del OAI-PMH. Dicha versión es estable y no es experimental como su predecesora, y a día de hoy no ha sufrido modificaciones.

OAI-PMH difiere de otros protocolos de interoperabilidad anteriores tales como Z39.50 (Lynch, 1997), Dienst (Lagoze y Fielding, 1998) o SDLIP (Paepcke *et al.*, 2000), en la facilidad de implementación, en el uso de estándares ampliamente conocidos y utilizados, y en el

mecanismo subyacente de recolección de metadatos a implementar por los proveedores de servicios.

Si bien el protocolo OAI-PMH ha sido fundamental para permitir la interoperabilidad entre los repositorios, se ha diseñado como un método de descubrimiento, no existe un protocolo de importación correspondiente a la interfaz de programación de aplicaciones (API). Hasta ahora no ha habido forma estándar de abstraer un objeto completo tal que pueda ser transportado de la plataforma de un repositorio a otra sin alguna pérdida de datos. Aquí es donde el protocolo Open Archives Initiative Object Reuse and Exchange (OAI-ORE) (Lagoze *et al.*, 2008) entra en juego. El protocolo OAI-ORE especifica interfaces de importación y exportación para permitir la reutilización y el intercambio de objetos digitales. Desde la perspectiva de la preservación digital, esto permite la migración futura de los objetos a una nueva plataforma preservando la funcionalidad que se espera desde el repositorio digital (Tarrant *et al.*, 2009).

Siempre que se intenta medir la calidad de algo se debe tener en cuenta qué es lo que se define como calidad y cómo se puede obtener una medida de ésta. A veces resulta complicado separar la subjetividad en medidas cualitativas, por ello (Guy, Powell y Day, 2004) definen la calidad dentro del contexto de los metadatos como: "los metadatos de alta calidad son los que respaldan los requerimientos funcionales del sistema que esté diseñado a soportar", lo que puede resumirse como "la calidad está relacionada con la aptitud para el propósito".

Son numerosos y variados los estudios realizados para medir la calidad en los metadatos; por su parte (Park, 2009) realiza un análisis del estado de la investigación y práctica sobre los metadatos basándose en la funcionalidad de éstos, la medición y criterios de evaluación, junto con mecanismos para mejorar la calidad de los mismos. En Hillmann (2008) se presenta un esquema de evaluación de la calidad basada en siete características: integridad, exactitud, procedencia, conformidad con las expectativas, la consistencia lógica y coherencia, oportunidad y accesibilidad. Cada una de ellas cuenta con una serie de criterios e indicadores de cumplimiento que el responsable debe tenerlos en cuenta al momento de realizar la evaluación.

Por otro lado Ochoa y Duval (2006) proponen un conjunto de métricas basadas en los mismos parámetros de calidad utilizados por un humano al realizar una revisión de los metadatos. La mayoría de estos estudios analizan la calidad en términos de la generación de los metadatos haciendo revisiones de las técnicas y herramientas (automáticas o semi-automáticas) implicadas (Guy, Powell y Day, 2004; Elings y Waibel, 2007; Hillmann, 2008; Man *et al.*, 2010; Mendes, Mühleisen y Bizer, 2012), mientras que sólo algunos se centran en la calidad de los datos una vez realizada la recolección de los mismos (Ward, 2002; Shreeves *et al.*, 2005; Nichols *et al.*, 2008; Jackson *et al.* 2008; Cechinel, Sánchez Alonso y Sicilia, 2009).

Para el presente trabajo se realizó un análisis de los metadatos recolectados sin tener en cuenta el modo en que éstos fueron generados. Una vez que los datos son recolectados deben ser entendidos; por ello un análisis de metadatos (Tennant, 2004) debería ser capaz de dar respuesta a preguntas como qué campos de los metadatos están realmente presentes y en qué porcentaje del total de registros; el grado de normalización y consistencia entre los valores reales de esos campos, o si es posible detectar patrones en los contenidos de tales metadatos.

Además, la mayoría de los autores (Hillmann, 2008; Ochoa y Duval, 2006; Nichols *et al.*, 2008; Jackson *et al.* 2008; Elings y Waibel, 2007; Shreeves *et al.*, 2005; Guy, Powell y Day, 2004; Ward, 2002; Man *et al.*, 2010; Mendes, Mühleisen y Bizer, 2012) coinciden en que la exactitud, integridad y consistencia son los criterios más comúnmente utilizados a la hora de medir la calidad en los metadatos.

2. Recolección de datos

El *Registry of Open Access Repositories* (ROAR) posee alrededor de 2.707 repositorios registrados en todo el mundo que funcionan actualmente bajo el protocolo OAI-PMH. Utilizando el filtro por países, ofrece un listado de 108 repositorios españoles, los cuales eliminando los duplicados se reducen a 74. De estos 74 repositorios, 5 no funcionan actualmente bajo el protocolo, ya que ni por la URL del repositorio ni por el *proxy* provisto por ROAR es posible recuperar los registros. Entre éstos se encuentran: *DADUN-Depósito Académico Digital de la Universidad Navarra*; *RODIN - University of Cadiz*; *Biblioteca Virtual de Andalucía*, *RODERIC (Universitat de Valencia)* y *CALAIX (Generalitat de Catalunya, Departament de Cultura)*. En la Tabla XI, en el apéndice, se presenta el listado completo de repositorios recolectados, con el nombre, la URL correspondiente donde se encuentran los recursos disponibles y la cantidad de registros de cada repositorio.

De estos 69 repositorios funcionales se recolectaron un total de 1.233.436 registros entre el 2011-11-05 y el 2011-12-10. Entre los reposito-

rios no existe una distribución uniforme en cuanto al volumen de registros por cada uno de estos, más del 47% del total de registros está concentrado sólo en 4 repositorios con más de 100.000 registros en cada uno, casi el 40% restante se concentra en 18 repositorios de 10.000 a 100.000 registros, mientras que poco más del 13% de registros se concentra en 35 repositorios de 1.000 a 10.000 registros, por último, algo menos que el 0.4% de registros se concentra en 12 repositorios con menos de 1.000 registros. Esto indica claramente una alta concentración de registros en unos cuantos repositorios de gran tamaño y un alto número de repositorios de tamaño medio (entre 1.000 y 10.000 registros).

Los registros fueron recolectados haciendo uso del protocolo OAI-PMH v 2.0. Para ello se desarrolló un programa *harvester* en lenguaje *C#* bajo el f*ramework* de *.NET 4.0* utilizando la librería *.NET OAI Harvester* como medio de recolección de metadatos de los repositorios.

Los metadatos fueron almacenados en una base de datos *SQL Server 2008R2* siguiendo el esquema *Dublin Core*. En total, por cada registro, se almacenaron 18 campos, 3 correspondientes al encabezado y los 15 restantes a los metadatos.

Los proveedores de datos o repositorios son diferentes en el volumen de datos que manejan, en la calidad de la conexión de red o ancho de banda, en el esquema de recolección (existen repositorios que al ser tan grandes sólo permiten la recolección mediante conjuntos, otros en cambio entregan el paquete entero de registros sin tener en cuenta una recolección selectiva), o en la forma de particionar el conjunto de registros. Por estas razones, el proceso de recolección debe ser cuidadoso de no sobrecargar al proveedor de datos con las peticiones y además debe contar con mecanismos de parada y de reanudación para evitar recolectar todo el conjunto más de una vez.

Como indica (Jackson *et al.*, 2008), los errores de sintaxis XML y codificación de caracteres, son problemas comunes, que de acuerdo al esquema de recolección seleccionado pueden afectar a un único registro o a todo el conjunto. Nosotros hemos encontrado una tasa muy baja de error, sólo 264 registros pertenecientes a 16 repositorios distintos entregan el XML resultante mal formado. Si bien es un número insignificante dada la cantidad total de registros recolectados, la herramienta desarrollada recolecta todos los repositorios de forma automática, con lo cual se tuvo en cuenta este tipo de errores de manera que el proceso no se detuviera en ningún momento.

Como se ha mencionado, los repositorios varían tanto en la cantidad de registros que poseen como en la velocidad de respuesta, si bien hay repositorios muy rápidos capaces de retornar hasta 8 registros por segundo, los hay muy lentos que retornan 1 registro cada 25 segundos. En cualquier caso, el tiempo promedio de recolección fue de 1.62 segundos por registro.

3. Campos Dublin Core

Un estudio realizado por (Ward, 2002) sobre 82 repositorios institucionales, indica la baja utilización de los quince elementos del esquema *Dublin Core*. Un resultado interesante al que llegó el autor fue que existe una media de ocho campos cumplimentados por registro.

Más tarde (Shreeves *et al.*, 2005) y las buenas prácticas publicadas por el *Institute of Museum and Library Services Digital Collections and Content Project* (IMLS DCC), confirmarían esos resultados identificando además ocho de estos quince elementos como importantes para la integridad de un registro de metadatos y de utilidad para las búsquedas. Los elementos en cuestión son: *title, creator, subject, description, date, format, identifier y rights*. En (Jackson *et al.*, 2008) se puede ver un estudio realizado sobre el portal del IMLS DCC basándose en estos ocho campos identificados como importantes.

De los repositorios recolectados ninguno cumplimenta todos los campos al 100%. Los más cercanos son los repositorios *Tiempos Modernos - Revista Electrónica de Historia* (60 registros) y *Tesis Doctorals en Xarxa (TDX)* (373 registros), ambos con un 88%. Por otro lado, el repositorio con menos porcentaje es *Open Data Córdoba - Cordoba Portal for Open Linked* (252 registros) con un 25% (este repositorio sólo cumplimenta los campos *title, creator, date y description*). Los dos repositorios con más alto porcentaje de cumplimentación son dos de los doce repositorios más pequeños (repositorios con menos de 1000 registros).

Por el contrario, el repositorio más grande (*Biblioteca Virtual de Prensa Histórica* con 203.922 registros) cumplimenta todos los campos en un 52%; el segundo repositorio más grande (*Revistes Catalanes amb Accs Obert* - RACO) con 138.533 registros) cumplimenta todos los campos en un 82% y el tercero (*Proyecto Agrega: busca y comparte contenidos*, con 133613 registros) en un 64%.

En cuanto a los campos recolectados, el campo *title* (título de la publicación) se encuentra cumplimentado casi en su totalidad por la mayoría de los repositorios, aunque con excepciones ya que los repositorios *SciELO – Spain* (19.794 registros), *UPCommons - Treballs acadèmics UPC* (10.277 registros) y *UPCommons - Universitat Politécnica de Catalunya* (9.766 registros) cumplimentan dicho campo en un 96%, 87% y 70% respectivamente.

En la Tabla I se observa en qué porcentaje cada campo recolectado es cumplimentado por el total de los repositorios (se agregó a los quince elementos el campo *setspec*; si bien este campo pertenece al encabezado y no a los metadatos, resulta también de interés tenerlo en cuenta). En esta tabla, el campo *identifier* de los metadatos aparece bajo la etiqueta *identifier2* para no confundir con el campo *identifier* del encabezado del registro.

De esta tabla destacan los campos *title* con un 99%, *creator* (autores de la publicación) con 56%, *subject* (descriptores o palabras claves) con 65% y *description* (resumen de la publicación) con 68%. Éstos permiten una mayor descripción del recurso y posibilitan búsquedas más cualificadas y precisas, y deberían, al igual que title estar cubiertos casi un 100%. Del total de registros que cumplimentan el campo *description,* el 16.8% posee menos de 100 caracteres de longitud, el 53.8% entre 100 y 500 caracteres, el 13% entre 500 y 1.000 caracteres, el 15.2% entre 1.000 y 5.000 caracteres, y algo más del 1% posee más de 5.000 caracteres.

4. Normalización

Un elemento importante de calidad es la normalización de los datos contenidos en los diferentes campos, en especial: formato de fechas (campo *date*), códigos o abreviaturas de lenguas (campo *language*), normalización y utilidad de palabras claves (campo *subject*), tipo de publicación (campo *type*), el formato de las mismas (campo *format*) y normalización de nombres personales (campos *creator*, *contributor*).

La *Online Computer Library Center* (OCLC) propone un conjunto de buenas prácticas para cada uno de los elementos del esquema *Dublin Core* con el objeto de fomentar la normalización de los metadatos generados por las entidades encargadas de la generación de los mismos. Entre estas normas figuran por ejemplo: en el caso de elementos repetidos que estos sean incluidos bajo una misma etiqueta y separados por punto y coma, o evitar el uso de títulos, honoríficos o sobrenombres en los elementos donde se requiera el nombre de una persona (elementos *creator, autor, contributor* y *publisher*) a menos que sea necesario para desambiguar dos nombres, o evitar el uso de términos poco descriptivos como el uso de la palabra “desconocido”. Para un mayor detalle de estas buenas prácticas referirse a (Ingram, 2012).

4.1. Campo *Date*

Los formatos de fechas dentro de cualquier tipo de repositorios siempre han sido un problema debido a la falta de normalización. Sobre todo si dicho repositorio no cumple una norma, formato o esquema. Este problema se ve agravado al reunir datos de diversas fuentes como es el caso del presente trabajo. Como se observó en el apartado anterior, el 93% de los registros cumplimentan el campo date, es decir, poseen un valor en dicho campo, pero esto no asegura que se trate de una fecha. Por ello, para tener un número real de cuántos registros no cumplen con un formato válido de fecha se construyó una función basada en expresiones regulares para verificar una serie de formatos y dar por válida una fecha determinada.

Campo	*% cumplimentación*
setspec	62
title	99
creator	56
subject	65
description	68
publisher	54
contributor	21
date	93
type	67
format	87
source	28
language	92
relation	46
coverage	48
rights	56
identifier2	99

Tabla I. Porcentaje de cumplimentación absoluto de los campos

Se tomó como formato base el *YYYY-MM-DD* Este formato presenta un compromiso o confluencia entre las distintas variantes, de acuerdo a la especificación del formato *Dublin Core*; éste propone como buena práctica el uso del perfil W3CDTF de la norma ISO 860.

Entre las distintas variantes encontradas y aceptadas como fechas válidas se encontraron: *31 de julio de 1949*, *[abans 1458]*, *[1987-jun.]*, *[1913]-1915*, *[1910-1920?]*, *16-02-2001*, *17/01/1903*, *[1914]*, *1919-11*, entre otras.

En la Tabla II se pueden ver contrastados los porcentajes de registros que no cumplen el formato de fecha elegido y el porcentaje de los registros que no son fechas, es decir, que al menos no pudieron ser verificados contra algún formato válido. En este análisis se encontró que el 4,28% de los registros (52.829 registros) no cumplen el formato de fecha indicado, mientras que sólo el 0,64% del total (7.865 registros) no son considerados fechas. Como datos sobresalientes, los repositorios *Biblioteca Un Single en Madrid* (255 registros) y *Bibliotecas Virtuales FHL - Registros de Autoridad* (3.855 registros), ambos poseen el 100% de registros que no respetan el formato de fecha, y de estos el 98% y 100% respectivamente, no son realmente fechas.

Otro dato curioso es que los repositorios *JABLE - Universidad de Las Palmas* (362 registros), *Biblioteca Digital Universitat les Illes Balears* (15.917 registros) y *REDINED* (73.610 registros), poseen el 100%, 99,5% y 77,8% de registros, respectivamente, que no respetan el formato de fecha, pero todos los registros fueron verificados como fechas válidas para esos tres repositorios.

# registros	*% mal formato*	*% no es fecha*
0-5000	27,97	13,73
5000-10000	5,44	0,45
10000-50000	19,09	0,87
+ de 50000	19,72	0,04

Tabla II. *Porcentaje de registros que no cumplen el formato de fecha YYYY-MM-DD vs. porcentaje de registros que no son fechas, en relación al tamaño de cada repositorio.*

4.2. Campo *Language*

El campo *language* indica el idioma en el que está escrita la publicación. Dentro de toda la base de datos de registros existen 359 expresiones distintas utilizadas para referirse al lenguaje del registro. Sin embargo, como muestra la Tabla III, las 20 variantes más representativas cubren el 90,84% de los registros, mientras que el 7,83% son cadenas vacías y el 1,33% corresponde a otras variantes.

Algo que salta a la vista es la falta de normalización ya que *ca, cat y Catalán* son el mismo lenguaje; *es, Español, y spa* son también el mismo lenguaje, y lo mismo sucede con las variantes *en, eng* e *Inglés*.

A la hora de indicar el valor del elemento lenguaje de un recurso web, sería deseable seguir algunas normas o estándares, por ello, la RFC 5646 (Phillips y Davis, 2009) propone un conjunto de las prácticas más recomendables para definir la estructura, contenido, construcción y semántica de la etiqueta lenguaje para los casos dónde es deseable indicar el lenguaje utilizado en un elemento de información.

Variantes Language	*# registros*
ca	25.429
cat	93.650
Catalán	8.117
en	25.419
eng	52.692
es	251.142
Español	125.188
eu	2.212
fr	1.116
fre	4.241
gl	1.519
glg	1.370
Inglés	2.573
ita	1.442
la	3.618
lat	10.609
por	1.606
spa	492.341
und	14.767
va	1.409
Vacíos	96.601
otros	16.375

Tabla III. *Variantes más representativas del campo Language*

4.3. Campo *Type*

Del mismo modo, el campo *type* señala qué tipo de publicación se trata; entre los más comunes se encuentran: artículos, libros, fotografías, artículo revisado por pares y tesis, entre otros. Este campo posee 601 variantes distintas y muchas de ellas provienen de errores ortotipográficos; por ejemplo existen 20 variantes para referirse a los artículos (*Aricle, Artcle, Arti-*

cle, Artícle, Articles, Articulo, Artículo, Arttículo, etc.). La Tabla IV resume los 25 tipos más frecuentemente utilizados y englobados con la etiqueta *Otros* el resto de identificadores; no se incluye la frecuencia de los espacios en blanco o vacíos (408.337 registros, el 33%).

Variantes Type	*# registros*
Artículo revisado por pares; Peer-reviewed Article	163.831
Artículo; article	131.902
Article de publicaci peridica	3.841
Dataset	3.860
Diapositiva	6.326
Fotografía	11.136
illustration	11.122
Imagen;info:eu-repo/semantics/other	5.059
info:eu-repo/semantics/article	11.296
info:eu-repo/semantics/book	12.245
info:eu-repo/semantics/doctoralThesis	18.098
Informe interno;Internal Report	8.248
journal article	19.794
Libros; Book	40.860
Master thesis (pre-Bologna period)	6.673
Parte de libro;Texto;info:eu-repo/semantics/bookPart;Text	11.402
photograph	95.474
Plano	5.028
Proyecto educativo	6.553
Pub. periódicas	38.871
sound effect	11.546
Tesis;Doctoral;PeerReviewed; Thesis	14.700
text	6.309
Varios	3.856
Otros	180.910

Tabla IV. Variantes más representativas del campo Type

En (DCMI 2006) existe una lista de términos aprobados que pueden ser utilizados como valor del elemento *type* de los metadatos para identificar el género de un recurso.

4.4. Campo *Format*

El campo o elemento *format* se utiliza sólo para los recursos digitales e indica el formato del archivo, el medio físico o la dimensión (tamaño o duración) del recurso. Es preferible el uso de la lista de tipos MIME o en su defecto un identificador de dos partes, formado por el *tipo/subtipo*, por ejemplo: audio/mp3.

Variantes Format	*# registros*
application/pdf	258.858
image;application/pdf	200.298
Vacío, sin especificar	159.860
image/jpeg	146.930
image/tiff	103.616
text/html;application/pdf	64.979
text/html	52.332
Formato Fsico	51.041
JPEG	12.438
2400 ppp., TIFF sin compresin ; 24 bits (color)	6.914
PDF	6.761
audio/mpeg	4.699
application/ogg	4.668
audio/wav	4.668
mb;ZIP	3.975
image/gif	2.440
image/x-wmf	2.437
application/x-shockwave-flash	2.276
Otros	144.246

Tabla V. Variantes del campo Format

En cuanto al formato de los registros almacenados, existe una gran variedad donde el predominio lo tienen los formatos *application/pdf* (19%), *image* (35%), *text/html* (8%) y *audio* (1%). Es necesario destacar que una gran parte de los registros está disponible en más de un tipo de formato. Por ejemplo, *image* y *application/pdf*, o *text/html* y *application/pdf*, como se puede observar en la Tabla V, son los elementos que tienen más de un tipo y se encuentran separados por el caracter de ";".

Existe casi un 20% de registros catalogados como *Otros* en dicha tabla; éstos pertenecen principalmente al formato *application/pdf* e *image* en realidad, pero están expresados en formas diversas: hasta 65.000 etiquetas distintas para referirse a estos formatos. Ejemplos de ello son: *XXXIX, 245 p.;application/pdf, recurso en línea, cr, marccarrier;aplication/pdf, p. 16;89967 bytes;application/pdf, KODAK i660 ; 300 ppp., TIFF sin compresión; 8 bits (escala de grises) ; 24 bits (color);984746 Bytes, EPSON GT 2500 ; 300 ppp., TIFF sin compresión ; 1 bit (blanco y negro);8901714 Bytes y CRO983;B/n; positiu;image/jpeg*, por nombrar sólo algunas.

4.5. Campo *Subject*

Las palabras claves o descriptores (campo *subject*) resultan de utilidad en los procesos de búsqueda, con lo cual contar con buenos descriptores es importante. Una medida de la calidad de éstos puede estar dada por la frecuencia de uso o por el porcentaje de descriptores que se hallan en el título o en la descripción (resumen) del registro.

# descriptores	*%*	*% acumulado*
1	19,43	
2	18,55	
3	10,36	
4	14,23	
5	7,42	
6	6,46	90,82
7	4,35	
8	4,50	
9	2,93	
10	2,60	
11	1,82	
12	1,68	
13	1,12	
14	0,97	8,29
15	0,81	
16 a 20	1,90	
21 a 55	0,88	0,88

Tabla VI. Cantidad de descriptores en un mismo registro

Sabiendo que el 65% del total de registros (801.293 registros) posee uno o más descriptores pero sólo el 53% (660.820 registros) de este total es distinto de espacios en blanco, en la Tabla VI se puede observar la cantidad de descriptores que se emplean por registro. Casi el 90.82% de los registros posee entre 1 y 10 descriptores, el 8.29% tiene entre 11 y 20 descriptores y el 0.88% entre 21 y 55.

En este trabajo, la base de datos conformada posee un conjunto de 461.941 descriptores distintos. La frecuencia de uso de éstos es importante ya que puede ayudar a identificar los mejores descriptores, los más/menos utilizados y también puede servir para identificar los posibles errores en su empleo. En la Tabla VII se enumeran los 20 descriptores más frecuentes, la cantidad de registros que los utilizan y el porcentaje sobre el total de registros que poseen descriptores utilizables (660.820 registros).

Descriptor	*Cantidad*	*%*
Prensa	63.323	9,58
UPC	37.467	5,67
Temàtiques	37.438	5,67
Arees temátiques de la UPC	37.431	5,66
Gallega	30.871	4,67
España	29.352	4,44
Historia	26.401	4,00
Arquitectura	16.527	2,50
Lingüística	16.280	2,46
Literatura	15.591	2,36
investigación	15.546	2,35
Politica	13.863	2,10
Control	13.563	2,05
Derecho	12.064	1,83
Economia	11.938	1,81
Lexicografia	11.702	1,77
lingüıstica sincrónica	11.619	1,76
Universidad	11.509	1,74
Salamanca	10.835	1,64
Informática	10.696	1,62

Tabla VII. Los 20 descriptores más utilizados

Por otro lado, es necesario destacar que de todo el conjunto de descriptores existe una gran cantidad que son utilizados muy pocas veces; en la Tabla VIII se resume la cantidad de descriptores en relación a su frecuencia de uso.

Como se mencionó anteriormente, muchos registros poseen más de un descriptor, aunque existen registros que poseen más de 45 descriptores.

Cantidad	*Frecuencia de uso*
406.971	1 - 10
42.433	11 - 100
10.625	101 – 1.000
1.577	1.001 +

Tabla VIII. Cantidad de descriptores vs. Frecuencia de uso

Con cierta frecuencia alguno o varios de éstos descriptores aparecen incluidos en el título o en el resumen del registro (campo *description* del esquema *Dublin Core*). En la Figura 1, se puede observar la cantidad de descriptores que aparecen tanto en el título como en el resumen del registro y la cantidad de registros que poseen esa cantidad de descriptores.

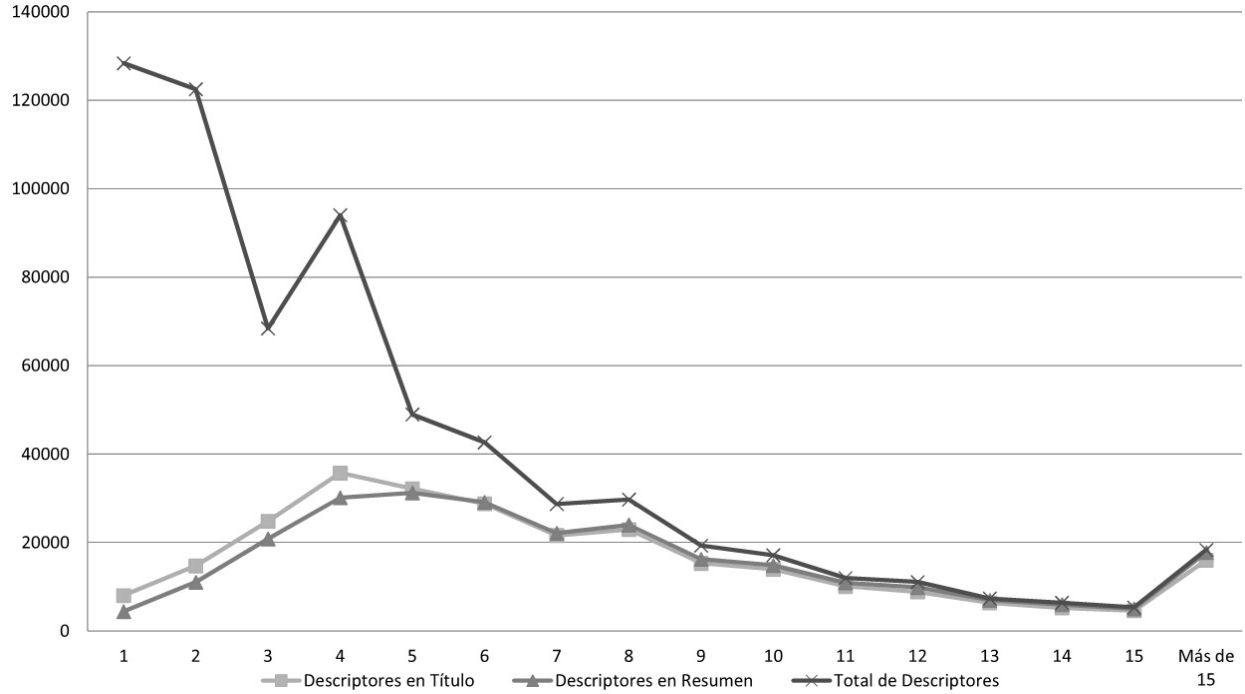

Figura 1. Cantidad de descriptores en títulos y resúmenes

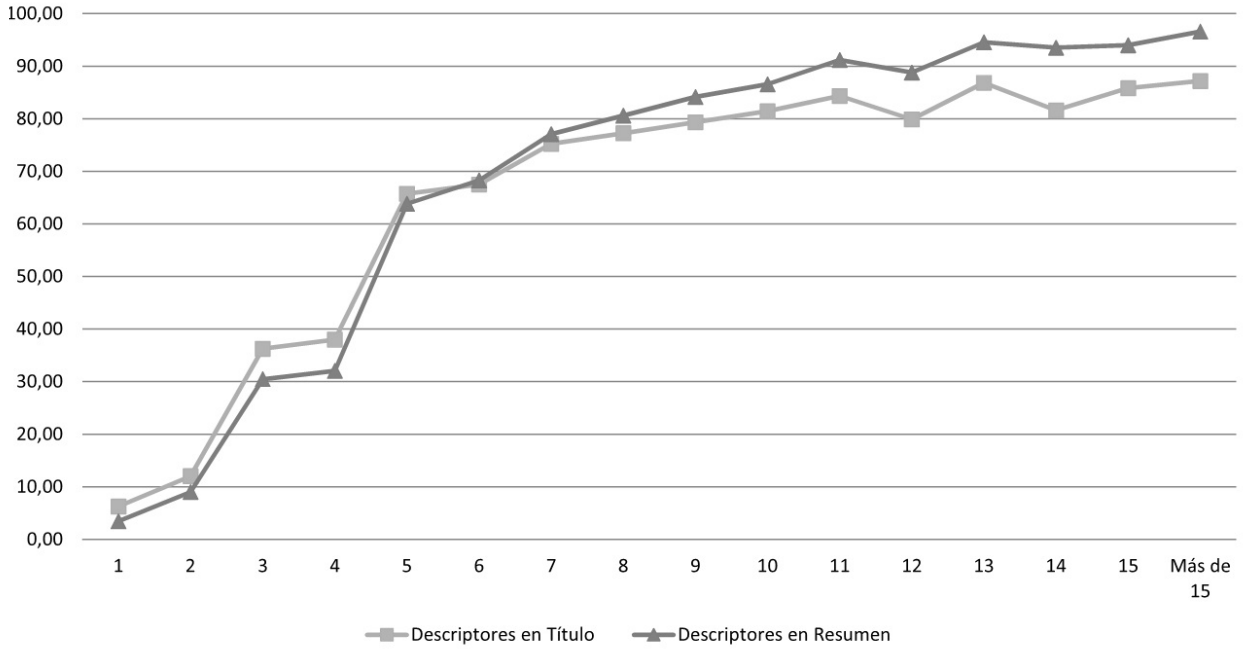

Figura 2. Porcentajes de registros que poseen descriptores en títulos o resúmenes

Para entender mejor este gráfico, en el eje *X* se encuentran las distintas cantidades de descriptores, como ejemplo, de los registros que poseen 4 descriptores (94.061 registros), existen 35.744 registros que incluyen al menos uno de los descriptores en el título y existen 30.148 registros que incluyen al menos uno de los descriptores en el resumen. En la Figura 2, se muestran los porcentajes de registros que poseen descriptores tanto en el título como en el resumen en relación a la cantidad de descriptores por registro.

4.6. Campo *Creator*

Del total de registros se extrajeron 456.652 nombres de autores distintos, correspondientes al campo *creator* del esquema de recolección.

El 93% (424.656 registros) se encuentra bajo la forma Apellidos, Nombres, mientras que el resto en la forma Nombres + Apellidos. La Tabla IX resume la cantidad de autores por publicación; bajo la etiqueta "Más de 10" se agruparon los registros que poseen entre 16 y 34 autores, que suman un total de 1.181 registros.

5. Esquema de metadatos

El hecho de que un repositorio pueda proveer los registros aplicando más de un esquema de metadatos o con esquemas más ricos o estructurados que el *Simple Dublin Core* permite un procesamiento de los metadatos más fácil y con mayores opciones (Bellini, Deussom y Nesi 2010).

Cantidad de autores	*# registros*
1	545.705
2	483.594
3	89.487
4	42.495
5	27.363
6	15.835
7	10.938
8	5.882
9	3.796
10	2.170
Más de 10	6.171

Tabla IX. Cantidad de autores por registros

En la Tabla X se resumen los esquemas de metadatos más utilizados y que se encuentran disponibles en los repositorios para poder recolectar los registros. Es obvio que el esquema OAI CD (*Dublin Core*) posee el porcentaje mayor, ya que es el único obligatorio que deben implementar los repositorios; el resto de esquemas son opcionales. En este listado sólo se ofrecen los 16 más frecuentes de las 24 variantes encontradas en este estudio; las demás variantes sólo son utilizadas por el 1% de los repositorios. En el estudio realizado por (Bellini, Deussom y Nesi 2010) se encontraron 153 esquemas distintos sobre 853 repositorios analizados y al igual que éste estudio, entre los más frecuentes se hallan: METS, RDF, DIDL, OAI MARC, CONTEXT OBJECT, MARCXML, RFC1807 Y UKETD DC, aunque en proporciones distintas.

%	*Prefijo (metadataPrefix)*
100	OAI_DC
43	METS
39	RDF
27	ESE
24	DID1
24	ORE
14	OAI_MARC
11	MARC21
11	QDC
10	MODS
8	EDM
4	MARC
4	Context_Object
4	MARCXML
4	RFC1807
4	UKETD_DC

Tabla X. Porcentaje de uso de los distintos esquemas de metadatos

6. Conclusiones

Este estudio se centró en el estado en que se encuentran el conjunto total de repositorios españoles que implementan el protocolo OAI-PMH. Analizando los metadatos recolectados de estos repositorios se logró identificar que el principal problema es producto de inconsistencias orto-tipográficas y la ausencia de mecanismos de control (títulos y descriptores muy extensos, diferentes formas de referirse al mismo idioma o al mismo tipo de archivo). Muchos proveedores de datos utilizan herramientas automáticas en la generación de los metadatos de los recursos que poseen, otros en cambio delegan la tarea de generación a personas con poca capacitación o a diferentes personas que aplican diferentes criterios en la carga de datos, y existen proveedores que sólo se dedican a adquirir los metadatos de fuentes externas. En este sentido sería de gran ayuda dotar de controles a dichas herramientas, supervisar periódicamente el proceso de generación de los metadatos, y en el caso de las personas implementar capacitaciones específicas y establecer criterios o normas para su creación.

Conviene destacar que si bien la especificación *Dublin Core* no exige que los campos se cumplimenten en un formato determinado, ya que

son simples cadenas de caracteres, sí propone normas y esquemas que los proveedores de datos pueden implementar para facilitar la interoperabilidad, tal es el caso de las RFC o los conjuntos de buenas prácticas de OCLC, para el uso normalizado de las fechas, lenguajes y tipos de archivo, y también para el resto de los campos como los títulos, resúmenes y nombres de personas. Luego de un extenso análisis, se puede afirmar de forma global que no existe un repositorio mejor que otro, pues si un repositorio cumplimenta la mayoría de los campos no lo hace de forma adecuada, y viceversa. Sobre todo los repositorios de gran tamaño son los que ofrecen una enorme falta de control y normalización, ya que no aplican las recomendaciones para la formación de los metadatos. Algo que también se debe tener en cuenta es que algunos de los repositorios fueron migrando de plataformas o tecnologías anteriores, con lo cual ciertos campos de información no existen o en su momento las especificaciones de formatos o reglas para su formación eran pobres o inexistentes, o simplemente existió una pérdida de información en el proceso de migración de las grandes bases de datos.

Por otro lado, coincidiendo con Ward (2002), se llegó a la conclusión de que no todos los campos son cumplimentados en su totalidad; y a diferencia de lo hallado por Shreeves *et al.* (2005), los ocho campos más utilizados fueron *title, identifier, date, language, format, description, type* y *subject*, difiriendo en *language* y *type*.

Además, uno de los campos más importantes como lo es *creator* solo es cumplimentado en un 56%, lo cual demuestra una enorme falta de control en la formación de los metadatos. Otro campo que ha sido descuidado ha sido *subject*; contar con descriptores o palabras claves bien formadas resulta útil en las tareas de búsquedas y recuperación. Mantener la calidad en los repositorios no es una tarea sencilla, además para que los proveedores de servicios puedan ofrecer resultados de calidad, dependen casi exclusivamente de los primeros. Por ello, siguiendo un conjunto de buenas prácticas como el uso de formatos o listas de control, o revisando los mecanismos de formación de los metadatos se puede dar un salto cualitativo en esta tarea.

En un trabajo posterior se intentará medir la calidad de estos repositorios analizando la duplicidad de registros e implementando técnicas de desambiguación que nos permita identificar de una forma más exacta que registros son duplicados cercanos, realizando búsquedas por aproximación dejando de lado las comparaciones exactas.

Como conclusión general, puede afirmarse que los repositorios digitales han experimentado un España un notable desarrollo, adquiriendo una masa crítica considerable. Pero este importante crecimiento en número y tamaño pone de manifiesto la necesidad de reforzar la calidad de los metadatos. Una vez plenamente asentados los repositorios, es tal vez el momento de centrar esfuerzos en ofrecer registros bien descritos, haciéndolos así más accesibles.

Agradecimientos

Este trabajo ha sido financiado parcialmente por el Ministerio de Ciencia e Innovación, a través de los proyectos de investigación FFI2011-27763 y CSO2009-07619.

Referencias

Bellini, Emanuele; Deussom, Marcel Aime; Nesi, Paolo (2010). Assessing Open Archive OAI-PMH implementations. // DMS. (2010) 153-158.

Cechinel, Cristian; Sánchez Alonso, Salvador; Sicilia, Miguel-Ángel (2009). Empirical Analysis of Errors on Human-Generated Learning Objects Metadata. // MTSR. (2009) 60-70.

DCMI Usage Board (2006). DCMI Type Vocabulary. // Technical report (August 2006). http://dublincore.org/documents/dcmi-type-vocabulary (2012-07-16).

Elings, Mary W.; Waibel, Gunter (2007). Metadata for All: Descriptive Standards and Metadata Sharing across Libraries, Archives and Museums. // First Monday. 12:3 (2007).

Ginsparg, Paul; Luce, Rick; Van de Sompel, Herbert (1999). The Open Archives initiative aimed at the further promotion of author self-archived solutions. // Universal PrePrint Service (UPS) Meeting, 1999.

Guy, Marieke; Powell, Andy; Day, Michael (2004). Improving the Quality of Metadata in Eprint Archives. // Ariadne. 38:38 (2004).

Hillmann, Diane I. (2008). Metadata Quality: From Evaluation to Augmentation. // Cataloging & Classification Quarterly. 46: 1 (2008) 65-80.

Ingram, Geri Bunker (2012). Best Practices for CONTENTdm and other OAI-PMH compliant repositories: creating sharable metadata // Technical report, Version 3.0 (2012-07-02). http://www.oclc.org/gateway/support/best_practices.pdf (2012-07-11).

Jackson, Amy S.; Han, Myung-Ja; Groetsch, Kurt; Mustafoff, Megan; Cole, Timothy W. (2008). Dublin Core Metadata Harvested Through OAI-PMH. // Journal of Library Metadata. 8:1 (2008) 5-21.

Lagoze, Carl; Fielding, David (1998). Defining Collections in Distributed Digital Libraries. // D-Lib Magazine. ISSN 1082-9873. 4:11 (November 1998).

Lagoze, Carl; Van de Sompel, Herbert (2003). The making of the Open Archives Initiative Protocol for Metadata Harvesting. // Library Hi Tech. ISSN 0737-8831. 21:2 (2003) 118-128.

Lagoze, Carl; Van de Sompel, Herbert (2001). The open archives initiative: building a low-barrier interoperability framework. // Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, 2001. 54-62.

Lagoze, Carl; Van de Sompel, Herbert; Johnston, Pete; Nelson, Michael; Sanderson, Robert; Warner, Simeon (2008). Open Archives Initative Object Reuse and Exchange (OAI-ORE). // Technical report, Open

Archives Initative (2008). http://www.openarchives.org/ore/1.0/toc (2012-07-11).

Lynch, Clifford A (1997). The Z39.50 Information Retrieval Standard Part I: A Strategic View of Its Past, Present and Future. // D-Lib Magazine. ISSN 1082-9873. 3:4 (April 1997).

Man, Yuan; Wei, Liu; Gang, Huang; Juntao, Gao (2010). A Noval Data Quality Controlling and Assessing Model Based on Rules. // Proceedings of the 2010 Third International Symposium on Electronic Commerce and Security, 2010. 29-32.

Mendes, Pablo; Mühleisen, Hannes; Bizer, Christian (2012). Sieve: Linked Data Quality Assessment and Fusion. // 1st International Workshop on Linked Web Data Management (LWDM 2012), 2012.

Nichols, David M.; Chan, Chu-Hsiang; Bainbridge, David; Mckay, Dana; Twidale, Michael B. (2008). A lightweight metadata quality tool. // JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, 2008. 385-388.

Ochoa, Xavier; Duval, Erik (2006). Quality Metrics for Learning Object Metadata. // Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, 2006. 1004-1011.

Paepcke, A.; Brandriff, R.; Janee, G.; R. Larson; Ludaescher, B.; Melnik, S.; Raghavan, S. (2000). Search Middleware and the Simple Digital Library Interoperability Protocol. // D-Lib Magazine. ISSN 1082-9873. 6:3 (March 2000) 5-8.

Park, Jung-Ran (2009). Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. // Cataloging & Classification Quarterly. 47:3 (2009) 213-228.

Phillips, A.; Davis, M. (2009). RFC 5646. Tags for Identifying Languages. // Technical report (September 2009). http://tools.ietf.org/html/rfc5646 (2012-07-16).

Shreeves, Sarah L.; Knutson, Ellen M.; Stvilia, Besiki; Palmer, Carole L.; Twidale, Michael B.; Cole, Timothy W. (2005). Is 'Quality' Metadata 'Shareable' Metadata? The Implications of Local Metadata Practices for Federated Collections. // Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, 2005.

Tarrant, David; O'Steen, Ben; Brody, Tim; Hitchcock, Steve; Jefferies, Neil; Carr, Les (2009). Using OAI-ORE to Transform Digital Repositories into Interoperable Storage and Services Applications. // The Code4Lib Journal. ISSN 1940-5758.6

Tennant, Roy (2004). Specifications for Metadata Processing Tools. // California Digital Library, 2004. http://roytennant.com/metadata_tools.pdf (2012-03-06).

Van de Sompel, Herbert; Lagoze, Carl (2000). The Santa Fe Convention of the Open Archives Initiative. // D-LIb Magazine. ISSN 1082-9873. 6:2 (February 2000).

Ward, J. (2002). A Quantitative Analysis of Dublin Core Metadata Element Set (DCMES) Usage in Data Providers Registered with the Open Archives Initiative (OAI). University of North Carolina at Chapel Hill, 2002. Tesis de Master.

Warner, Simeon (2003). Eprints and the open archives initiative. // Library Hi Tech. ISSN 0737-8831. 21:2 (July 2003) 151-158.

Webley, Lawrence; Chipeperekwa, Tatenda; Suleman, Hussein (2011). Creating a National Electronic Thesis and Dissertation Portal in. // 14th International Symposium on Electronic Theses and Dissertations, 2011.

7. Apéndice 1

Nombre	*URL para la recolección*	*registros*
Acceda. Universidad de Las Palmas de Gran Canaria	http://acceda.ulpgc.es/oai/request	4703
Archivo Abierto - Universidad Carlos III de Madrid	http://e-archivo.uc3m.es:8080/dspace-oai/request	10389
Archivo Digital de la Universidad Politécnica de M	http://oa.upm.es/perl/oai2/	7269
Arias Montano: Universidad de Huelva	http://rabida.uhu.es/dspace-oai/request	4887
Biblioteca Digital de Castilla y León	http://bibliotecadigital.jcyl.es/i18n/oai/oai.cmd	3159
Biblioteca Digital Universitat les Illes Balears	http://ibdigital.uib.es/greenstone/cgi-bin/oaiserver.cgi	15917
Biblioteca Un Single en Madrid	http://www.openarchives.es/singlemadrid/oai.asp	255
Bibliotecas Virtuales FHL. Registros bibliográfico	http://www.larramendi.es/i18n/oai/oai_larramendi.es.cmd	2116
Bibliotecas Virtuales FHL. Registros de Autoridad	http://www.larramendi.es/i18n/oai/oai_aut.larramendi.es.cmd	3855
Biblos-e Archivo - Universidad Autonoma de Madrid	http://digitool-uam.greendata.es/OAI-PUB	5297
DIALNET: Servicio de Alertas y Hemeroteca Virtual	http://dialnet.unirioja.es/oaites/OAIHandler	2297
DIGIBUG Universidad de Granada	http://digibug.ugr.es/oai/request	16042
Digital.CSIC	http://digital.csic.es/dspace-oai/request	38920
Digitum : Depósito de la Universidad de Murcia	http://roar.eprints.org/cgi/celestial/oai_proxy/1220	12317
Dipòsit de la Recerca de Catalunya (RECERCAT)	http://oai.recercat.net/request	28793
Diposit Digital de la Universitat de Barcelona	http://diposit.ub.edu/dspace-oai/request	4753
DSpace Universidad Alcala: Principal	http://dspace.uah.es/oai/request	6703
DUGiDocs – Universitat de Girona	http://dugi-doc.udg.edu/dspace-oai/request	2868

DUGiMedia – Universitat de Girona	http://diobma.udg.edu/dspace-oai/request	1660
EMD Fundacion Sancho el Sabio Fundazioa	http://www.memoriadigitalvasca.es/dspace-oai/request	3053
E-Prints Universidad Complutense Madrid	http://eprints.ucm.es/cgi/oai2	9953
e-Repositori UPF - Universitat Pompeu Fabra	http://oai-repositori.upf.edu/request	5219
Euskal Doktorego Tesien Bilduma -Tesis Doctorales	http://edtb.euskomedia.org/cgi/oai2	5517
Fotografía Sobre España en el Siglo XIX	http://roar.eprints.org/cgi/celestial/oai_proxy/2287	4079
Fundacion MAPFRE	http://www.mapfre.com/documentacion/publico/i18n/oai/oai.cmd	20126
GALICIANA: Biblioteca Dixital de Galicia	http://galiciana.bibliotecadegalicia.xunta.es/i18n/oai/oai.cmd	104875
Gestion del conocimiento e innovacion	http://www.invenia.es/oai/oai2.asp	518
GREDOS: Gestión del Repositorio Documental de la Universidad de Salamanca	http://gredos.usal.es/oai/request	90102
Helvia: Universidad de Cordoba	http://helvia.uco.es/oai/request	5513
JABLE. Universidad de Las Palmas	http://jable.ulpgc.es/jable/cgi-bin/oai	362
MDX (Materials Docents en Xarxa)	http://oai.mdx.cat/request	818
Memoria digital de Canarias (mdC)	http://bibmdc2.ulpgc.es/cgi-bin/oai.exe	32000
O2: The Oberta in open access - UOC Institutional	http://openaccess.uoc.edu/webapps/dspace_rei_oai/request	2761
Open Data Córdoba | Cordoba Portal for Open Linked	http://roar.eprints.org/cgi/celestial/oai_proxy/2663	252
Open Journal Systems	http://recyt.fecyt.es/index.php/index/oai/	6409
Portal de Revistas Científicas Universidad CompluM	http://revistas.ucm.es/index.php/oai/oai/	29785
Proyecto Agrega: busca y comparte contenidos	http://contenidos.proyectoagrega.es/oaipmh/OaiPmhRequest/OaiPmhRequest.do	133613
pub.bsalut.net	http://roar.eprints.org/cgi/celestial/oai_proxy/1334	165
Recursos de Investigación de la Alhambra	http://www.alhambra-patronato.es/oai	13313
REDINED	http://roar.eprints.org/cgi/celestial/oai_proxy/976	73610
Repositori Obert UdL - Universitat de Lleida	http://roar.eprints.org/cgi/celestial/oai_proxy/2258	8747
Repositori UJI - Universitat Jaume I	http://repodoc.uji.es/oai/request	5445
Repositorio da Universidade da Coruña	http://ruc.udc.es/oai/request	5344
Repositorio Institucional de Asturias (RIA)	http://ria.asturias.es/dspace-oai/request	597
Revistes Catalanes amb Accés Obert (RACO)	http://www.raco.cat/index.php/index/oai/	138533
RiuNet	http://riunet.upv.es/oai/request	7881
RODERIC. Universitat de Valencia	http://roderic.uv.es/oai/request	7890
RUA - Repositorio Universidad de Alicante	http://roar.eprints.org/cgi/celestial/oai_proxy/1013	16112
SAVIA - Conocimiento abierto. EOI Escuela de Organ	http://roar.eprints.org/cgi/celestial/oai_proxy/2548	1123
SciELO - Spain	http://scielo.isciii.es/oai/scielo-oai.php	19794
Temaria - Revistas digitales de biblioteconomía y	http://temaria.net/metadatos.php	2858
Tesis Doctorals en Xarxa	http://tdx.cat/oai/request	11623
Tesis Doctorals en Xarxa (TDX)	http://roar.eprints.org/cgi/celestial/oai_proxy/777	373
Tiempos Modernos: Revista Electrónica de Historia	http://roar.eprints.org/cgi/celestial/oai_proxy/300	60
Travesía	http://roar.eprints.org/cgi/celestial/oai_proxy/2242	1863
Universidad de Aragón	http://www2.cita-aragon.es/oai/request	1314
Universidad de Burgos - Repositorio Institucional.	http://dspace.ubu.es:8080/tesis-oai/request	106
Universidad de Málaga	http://riuma.uma.es/oai/request	1843
Universidad del País Vasco	http://addi.ehu.es/oai/request	288
Universidad Politécnica de Cartagena	http://repositorio.bib.upct.es/dspace-oai/request	1748

Universidad Rey Juan Carlos	http://eciencia.urjc.es/dspace-oai/request	1217
Universidade de Santiago de Compostela	http://dspace.usc.es/oai/request	3055
Universitat Autònoma de Barcelona	http://roar.eprints.org/cgi/celestial/oai_proxy/1239	17641
UPCommons - Revistes i congressos UPC	http://eprints.upc.es:8080/revistes-oai/request	8609
UPCommons - Treballs acadèmics UPC	http://eprints.upc.es:8080/pfc-oai/request	10277
UPCommons - Universitat Politècnica de Catalunya	http://eprints.upc.es:8080/dspace-oai/request	9766
UVaDOC - Universidad de Valladolid	http://uvadoc.uva.es/oai/request	721
Virtual Library of Bibliographical Heritage	http://bvpb.mcu.es/i18n/oai/oai.cmd	30443
Virtual Library of Historical Press	http://prensahistorica.mcu.es/prensahistorica/OAIHandler	203922

Tabla XI. Listado de repositorios recolectados

Enviado: 2012-05-04. Versión corregida: 2012-06-23.
Aceptado: 2012-08-21.