Hacia una normalización de los datos para un buen funcionamiento de los repositorios universitarios en el Perú

Juanita Jara de Súmar. McGill University Library (r). Montreal, Canadá
juanita.jaradesumar@mcgill.ca

Ana María Talavera Ibarra. Pontificia Universidad Católica del Perú. Lima, Perú
atalave@pucp.edu.pe

Resumen

Para que el contenido de los repositorios universitarios peruanos tenga visibilidad y sea incluido en el repositorio nacional y los repositorios internacionales es preciso asegurar el empleo de normas internacionales de descripción. Este estudio analiza el empleo de elementos del Dublin Core para la descripción de documentos en los repositorios de cuatro universidades limeñas y dos de provincias, nacionales y privadas, y examina los aspectos en que es preciso mejorar el control de calidad de los datos proporcionados. Se proponen recomendaciones a partir de las conclusiones encontradas.

Palabras clave: Perú. Repositorios institucionales. Repositorios Universitarios, Dublin Core. Normalización

Introducción

Gracias al desarrollo de la tecnología y las comunicaciones, y teniendo en cuenta el incremento creciente del precio de las suscripciones a publicaciones periódicas de contenido académico, desde principios de este siglo las universidades y sus bibliotecas han venido estudiando alternativas de publicación que les permitan, de un lado, tener acceso a los resultados de las investigaciones nacionales y extranjeras y, de otro, dar visibilidad a las publicaciones de sus propios investigadores, ofreciendo acceso gratuito, siguiendo las iniciativas mundiales de acceso abierto (*Open Access*).

En el Perú, el movimiento empezó con la digitalización de tesis de grado y ha seguido con las revistas publicadas por nuestras universidades. Las estadísticas de consulta del repositorio Cybertesis de la Universidad Nacional Mayor de San Marcos (UNMSM)[1] empiezan en octubre de 2004.

Las obras digitalizadas se almacenan en los llamados repositorios institucionales, archivos digitales que pueden ser consultados utilizando Internet gratuitamente y desde cualquier parte del mundo.

Considerando la importancia de estos recursos, el gobierno peruano creó en mayo de 2013, por ley No. 30035, el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto, cuyo reglamento fue aprobado hace escasamente un año.[2] Este reglamento establece que la implementación y gestión del repositorio nacional, denominado ALICIA, es responsabilidad del Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC), quien además debe implementar la Red Nacional de Repositorios Digitales de Ciencia, Tecnología e Innovación de Acceso Abierto (RENARE). Integran esta red los repositorios institucionales de las entidades del sector público, aunque también pueden participar en ella repositorios del sector privado. Al 31 de enero 2016, RENARE tiene un total de 46 participantes: 29

[1] http://sisbib.unmsm.edu.pe/cgi-bin/awstats.pl?month=12&year=2004&output=main&config=cyber&framename=index
[2] https://portal.concytec.gob.pe/images/stories/images2013/portal/areas-institucion/dsic/reglamento_repositorio_nacional_alicia.pdf

universidades nacionales y privadas, 9 institutos de investigación y 8 repositorios gubernamentales.

El reglamento de la ley define los conceptos empleados para describir los repositorios y establecen criterios técnicos y académicos para pertenecer a RENARE. Entre ellos podemos mencionar: compilación de estadísticas de monitoreo (art. 2.5); interoperabilidad que permita el intercambio y transferencia de datos, metadatos e información (art. 2.7); empleo de metadatos normalizados, (art. 2.9); contar con un responsable en cada repositorio, así como regulación interna para la administración de los metadatos; y, el contenido y tipos de documentos que puede ser aceptados en el repositorio (art. 3.3) ("Reglamento de la ley N° 30035, ley que regula el repositorio nacional digital de ciencia, tecnología e innovación de acceso abierto : Decreto Supremo N° 006-2015-PCM," 2015)

Hay también *Directrices* emitidas por CONCYTEC para el procesamiento de información en los repositorios institucionales, ahora en su versión 05.15.[3] Estas directrices proporcionan un cuadro con 71 metadatos de acuerdo al estándar *Dublin Core*. De ellos, 12 son obligatorios, 8 recomendados y 51 opcionales. Los metadatos obligatorios y recomendados incluyen descripción, consideraciones técnicas y ejemplos.

Este es el marco legal del repositorio Nacional ALICIA,[4] que cosecha los datos de los repositorios individuales que son el objeto de nuestro estudio. Estas directrices deben favorecer la posibilidad de comparar datos, dado que todos los repositorios que participan en la red emplean las mismas normas y protocolos.

Metadatos, normalización e interoperabilidad

Los metadatos, o “datos sobre datos”, son elementos que permiten describir los objetos digitales. Además de los elementos típicos utilizados en la descripción de cualquier recurso bibliográfico, se incluyen elementos necesarios para la accesibilidad, conservación, garantía de los derechos de autor, etc.

[3] http://portal.concytec.gob.pe/images/documentos/alicia/directrices_repositorio_06-04-2015.pdf
[4] http://alicia.concytec.gob.pe/vufind/

La norma internacional de la American National Standard (ANSI) para la creación de metadatos, ANSI/NISO Z39.85-2012 (revisada febrero 20, 2013), más conocida como Dublin Core (DC), establece un conjunto mínimo de 15 elementos sin calificar, *The Dublin Core Metadata Element Set* (DCMES).

Además, con el fin de intercambiar metadatos y hacer la información accesible a la comunidad internacional, se han establecido normas de interoperabilidad entre repositorios, como el protocolo OAI-PMH (*Open Archives Initiative, Protocol for Metadata Harvesting*). Estas normas permiten a los "cosechadores" reunir los metadatos de diferentes fuentes, para dar acceso abierto a los repositorios.

Sin embargo, cada una de los repositorios en el ámbito nacional e internacional interpreta las normas de diferente manera, aunque el elemento básico para el intercambio de información es justamente la normalización. En este sentido Pons, Hilera y Pagés (2013) mencionan que es indispensable que los metadatos sean de calidad para ser útiles, y que de su buena definición dependerá su correcta creación. Ellos indican especialmente que "La calidad en los metadatos refleja el grado con el que los metadatos realizan sus funciones esenciales bibliográficas de búsqueda, localización, uso, procedencia, autenticación y administración" Pons, Hilera &Pagés, 2013, p. 336).

Estos autores hacen referencia a diferentes expertos que coinciden en los siete criterios más usados en la medición de la calidad de los metadatos: precisión, completitud, procedencia, conformidad con las expectativas, coherencia, seguimiento temporal y accesibilidad. De éstos, los más comunes son la precisión, completitud y la consistencia.

Por otro lado, Medrano, Figuerola y Berrocal, en su evaluación de los metadatos de los repositorios digitales españoles, mencionan que la calidad es especialmente medida por la normalización de los contenidos en los diferentes campos de descripción. Para ellos es particularmente importante la normalización de los campos de fechas (*date*), códigos de idiomas (*language*), palabras-clave (*subject*), tipo de publicación (*type*) y la normalización de los nombres personales (Medrano, Figuerola & Berrocal, 2012, p. 112).

De acuerdo con Stvilia y sus colaboradores (Stvilia, Gasser, Twidale, Shreeves, & Cole, 2004), los problemas en la calidad de los metadatos pueden surgir tanto

a nivel del esquema (nivel macro), como de los componentes (nivel micro). En su estudio aplicaron un modelo de nueve pasos a 150 registros provenientes de 16 participantes en un repositorio colectivo, elegidos aleatoriamente. Los investigadores encontraron 6 tipos principales de problemas cualitativos: (1) los datos estaban incompletos –ninguno utilizó los 15 elementos básicos; (2) tenían metadatos superfluos –94% tenían elementos duplicados; (3) faltaba claridad –esto se dio particularmente al no poder determinar a qué se referían las fechas; (4) uso incorrecto de los elementos de DC o inconsistencias semánticas, como colocar información en un campo que no corresponde; (5) estructura inconsistente, particularmente en el formato de las fechas, y en el uso de vocabularios controlados; (6) representación errónea, como enlaces con errores o falsa representación del contenido.

Además de aplicar el modelo, dichos autores sugieren que debe tomarse en cuenta el uso que se hace de la información. Una idea clave para ellos es que la calidad de la información se vuelve crítica sólo en el grado en que afecta la calidad de los *resultados* de la actividad. Y para ello deben considerarse tanto los componentes individuales como el resultado global del registro completo. También sugieren una medida de bajo costo mediante la cual los usuarios pueden contribuir a mejorar la calidad de los repositorios múltiples, como por ejemplo, ofrecer la posibilidad de enviar comentarios y comunicar errores a los administradores. Los comentarios de los usuarios tienden a provenir de la experiencia y referirse a problemas relacionados con la usabilidad, y pueden ser de mucha ayuda para mejorar los metadatos y el diseño del sistema. (Stvilia et al., p. [12])

La Confederación de Repositorios de Acceso Abierto (COAR) considera que el elemento clave para obtener un cuerpo unificado de documentos académicos es la interoperabilidad "specifically, that repositories follow consistent guidelines, protocols, and standards for interoperability which allow them to communicate with each other; connect with other systems; and transfer information, metadata, and digital objects between each other." [5] (Confederation of Open Access

[5] específicamente, que los repositorios sigan lineamientos, protocolos y normas para la interoperabilidad coherentes que les permitan comunicarse a unos con otros; conectarse con otros sistemas; y transferir información, metadatos y objetos digitales entre ellos. [traducción nuestra]

Repositories, 2012, p. 4). Entre las siete áreas principales que se vienen trabajando, se mencionan los sistemas de estadísticas de uso y los de registro e identificación de autores.

Hay que mencionar también las investigaciones y publicaciones del *Dublin Core Metadata Initiative* (DCMI), como los *DCMES* y los *DCMI Metadata Terms*, que permiten conocer el contenido y sintaxis de los datos a ingresar en cada uno de los elementos del DC. Además, la Library of Congress (LC) ha elaborado interfaces (crosswalk) para convertir los datos de MARC21 al DC y viceversa, incluyendo tanto los 15 elementos de DC sin calificar como los elementos calificados. Entre estos últimos se propone calificar los datos utilizando listas o componentes ya normalizados, como por ejemplo LCSH, MESH, u otros, en el elemento *dc.subject;* códigos ISO en *dc.language*; y los MIMEtype para *dc.type*. (Library of Congress. Development and MARC Standards Office, 2008).

Un documento sumamente relevante para la normalización de los datos, es el publicado por OCLC (Online Computer Library Center), que reúne un conjunto de buenas prácticas denominado "*Best Practices for CONTENTdm and other OAI-PMH compliant repositories: creating sharable metadata"* versión 3.1 (OCLC, 2013). En este documento, para cada elemento del DC se incluyen comentarios y se citan ejemplos y recomendaciones de diversos usuarios de los DCMES. Una valiosa herramienta a seguir para normalizar los metadatos.

El buscador de un repositorio nacional actúa de una manera similar a las búsquedas federadas. No combina los datos de todos los repositorios, sino que combina los resultados. Y para que los resultados correspondan a la búsqueda efectuada deben utilizar una estructura estandarizada. En el caso de ALICIA, los datos de los repositorios individuales deben corresponder a los mismos elementos de DC para poder ser cosechados adecuadamente utilizando el OAI-MHP.

Pero pese a contar con las directrices de CONCYTEC, vemos que sus definiciones y ejemplos no son suficientemente claros en el caso de algunos elementos cruciales para garantizar su normalización. Ya, en un estudio de los metadatos empleados en repositorios de tesis de 10 universidades canadienses, se concluyó que la falta de normalización encontrada parece deberse a que cada

repositorio decide cómo definir los elementos, no respeta la descripción proporcionada por quienes elaboraron las normas, o prefieren agregar calificadores que les parecen más aparentes. Las diferencias e inconsistencias se encontraron sobre todo en la interpretación del significado de ciertos elementos, como por ejemplo, los varios tipos de fechas y los roles de autores, así como en la forma de describir los grados académicos. (Park & Richard, 2011). Estos autores piensan que aunque los consorcios de repositorios proponen normas para todos sus participantes, sistemas como el cosechador OAI-PMH están diseñados de manera que no constituyan una barrera muy elevada para los proveedores de datos. Esta circunstancia proporciona flexibilidad pero, al mismo tiempo, permite inconsistencias en la elección de los elementos y la descripción del contenido.

Uno de los elementos en que la normalización es muy importante, es la materia. Las universidades que utilizan en este campo el mismo vocabulario controlado que en el catálogo de su biblioteca contribuyen a que sus usuarios obtengan resultados de mejor calidad al buscar en el repositorio. Pero cuando se crea un repositorio nacional, en el que cada miembro participa con su propia lista de materias o utiliza simplemente palabras-clave, como es el caso de ALICIA, se pierden las ventajas del vocabulario controlado y el usuario requiere un mayor conocimiento del tema para encontrar lo que necesita. Una sugerencia interesante para mejorar los resultados de la búsqueda por materias en los repositorios múltiples incluye el mapeo a tesauros y ontologías que reduzcan la complejidad de los metadatos. (Stvilia et al., 2004, p. [11])

Selección de la muestra

Para este estudio decidimos que en la muestra estuvieran representadas universidades públicas y privadas de Lima y de provincias con un contenido adecuado para efectuar el análisis. Basándonos en los datos de ALICIA respecto al tamaño de los repositorios, y teniendo en cuenta los tipos de documentos registrados en ellos, seleccionamos cuatro universidades en la capital y dos en provincias (en orden de fecha de fundación):

- Universidad Nacional Mayor de San Marcos (UNMSM)

- Universidad Nacional de Ingeniería (UNI)
- Universidad Nacional de Trujillo (UNITRU)
- Pontificia Universidad Católica del Perú (PUCP)
- Universidad De Piura (UDEP)
- Universidad Peruana de Ciencias Aplicadas (UPC)

Además, revisamos otros repositorios universitarios, no incluidos en este estudio, para constatar que la muestra elegida fuera representativa.

Datos a analizar y comparar

Al iniciar la búsqueda de los datos sobre los repositorios de cada universidad advertimos que no todas las universidades dan visibilidad adecuada a sus recursos ni proporcionan información mínima acerca de ellos. Decidimos entonces iniciar el análisis a partir de los puntos de acceso al repositorio encontrados en la página principal de cada universidad y en la de su biblioteca y luego, examinar las características generales de la página inicial del repositorio.

Tabla 1: Características de la página inicial de los repositorios

	UNMSM[6]	UNI[7]	UNITRU[8]	PUCP[9]	UDEP[10]	UPC[11]
Enlace desde la página principal de la Universidad	no	no	no	si	si	si
Enlace desde la página de biblioteca	si	no	si	si	si	si
Lineamientos/ política/jerarquía	no	no	no	no	si	si
Historia: fecha inicio	no	no	no	no	no	no
a) Contenido/ b) Estadística global	b si	no	no	no	no	Wikipedia b) si
Búsqueda a)simple	a	a	a	a	a	a

[6] http://cybertesis.unmsm.edu.pe/
[7] http://cybertesis.uni.edu.pe/
[8] http://dspace.unitru.edu.pe/xmlui/
[9] http://tesis.pucp.edu.pe/repositorio
[10] http://pirhua.udep.edu.pe/
[11] http://repositorioacademico.upc.edu.pe/upc/

b)avanzada	b	b	b	b		b
Listados por a) Autor; b) título; c) comunidad /colección d) fecha e) tema	todos	todos	todos	todos	todos	todos
Software	DSpace	DSpace	DSpace	DSpace	DSpace	OpenRep
Ayuda	DSpace	DSpace	no	FAQ	no	FAQ
Contacto	si	no	si	si	no	si
Idioma interfaz	spa	spa	eng	eng	bilingüe	bilingüe

Varias universidades tienen un repositorio diferente para las revistas y para las tesis. Es el caso de la PUCP y la UNMSM que tienen un Repositorio Institucional general, además de repositorios separados para tesis y revistas. Los datos de la PUCP y la UNMSM en la tabla 1, corresponden únicamente a sus repositorios de tesis.

Acceso

Las tres universidades privadas tienen enlaces al repositorio desde la página principal de la universidad y desde la página de la biblioteca, expresando así la importancia asignada a este recurso y su relación con los recursos de información. La UDEP emplea el nombre del repositorio, *Pirhua*, pero no el término repositorio. La UNI no tiene enlaces en ninguna de las dos páginas.

Únicamente la UPC enlaza directamente a la información sobre su contenido, (en Wikipedia), y a las políticas que rigen el repositorio, emitidas por la Dirección de Gestión de Conocimiento. La UDEP tiene políticas emitidas por el Vicerrectorado de Investigación. El enlace está en la página de *guías y tutoriales de la biblioteca*, bajo el nombre *Políticas de publicación del RI Pirhua*, aunque cubre aspectos administrativos del repositorio y no sólo de publicación. Ambas universidades asignan la responsabilidad sobre la calidad del texto a los autores y los departamentos académicos a los que pertenecen; y señalan a la biblioteca como responsable de la calidad de los registros y los metadatos.

Es de suponer que las demás universidades cuentan también con un documento de política y asignación de responsabilidades, tal como lo exige el reglamento de la ley peruana, pero esa información no parece estar disponible públicamente.

Desarrollo histórico

Otro aspecto muy importante es la falta total de información sobre la fecha de implementación de los repositorios, así como una línea de tiempo que indique el inicio y fin de las digitalizaciones retrospectivas. Estas fechas son muy importantes para efectuar análisis de visibilidad e impacto del contenido incorporado al repositorio con posterioridad a la fecha en que fue originalmente publicado. Asimismo, para juzgar las estadísticas de consulta y visualización total de cada documento, es necesario saber con precisión qué período abarcan.

Estadísticas globales

Las estadísticas de uso disponibles, tanto globales como para documentos individuales, proporcionan más preguntas que respuestas. Únicamente UPC y UNMSM tienen un enlace a sus estadísticas globales de uso, aunque como utilizan diferente software, sus estadísticas no son totalmente comparables. Los enlaces de ambas se encuentran en la parte inferior de la página inicial, no muy visibles para los usuarios.

La UNMSM ofrece estadísticas exhaustivas, mes a mes, desde octubre 2004[12]. Estas estadísticas son una fuente riquísima de información sobre *cuándo* se efectúan las consultas (mes, día, hora), *quienes* consultan (países, servidores, robots), *cómo* navegan, *enlaces* de acceso al repositorio y de búsquedas, etc.

Las estadísticas de la UPC[13] ofrecen totales globales de: visitas y descargas, documentos más vistos y descargados, países y ciudades desde donde se consulta el repositorio (incluyendo un mapamundi interactivo). Pero no se sabe qué período cubren estas estadísticas, pues no hay fecha de inicio. Se presentan también estadísticas para cada uno de los últimos seis meses. Estas estadísticas despiertan preguntas sobre la terminología y el significado de las cifras. Se habla de elementos (items) y descargas (downloads) cuyas cifras globales son bastante similares: 414,418 elementos vistos y 404,096 descargas. Pero cuando vemos el detalle de la consulta de los últimos seis meses (tabla 2), encontramos que, en algunos meses, la cifra de descargas casi duplica la de elementos y en otros meses es menos de la mitad, lo que hace dudar del significado y conteo de cada uno de estos parámetros.

[12] http://sisbib.unmsm.edu.pe/cgi-bin/awstats.pl?config=cyber
[13] http://repositorioacademico.upc.edu.pe/upc/displaygastats

Tabla 2: Muestra de estadísticas de la UPC

	September 2015	October 2015	November 2015	December 2015	January 2016	February 2016
Items	26739	26071	23576	11723	15986	613
Downloads	44003	33039	5504	4807	8190	351

Estructura de las comunidades y colecciones

La estructura de los repositorios por comunidades (categorías de primer nivel) y colecciones es muy variada y no siempre es visible desde la página inicial. UDEP y UPC, cuyas tesis están listadas a nivel de colección y subcolección, junto con otros tipos de documentos en un mismo repositorio, requieren tres o cuatro “clics” para encontrar las tesis y la cantidad de ellas existente en cada colección del repositorio.

Los repositorios especializados de UNMSM, UNITRU, UNI y PUCP sólo contienen tesis y tienen un listado de sus comunidades en la página inicial. La estructura de comunidades y colecciones de cada repositorio se basa en un principio diferente y no permite compararlos. Sólo UNMSM y PUCP tienen sus repositorios desarrollados; los otros tienen varias comunidades listadas pero sin contenido.

Listados y barras de búsqueda

Excepto UPC, que utiliza otro software, y UDEP, que ha diseñado una página inicial diferente, los demás repositorios mantienen la funcionalidad original de DSpace para la barra de búsqueda – búsqueda simple y búsqueda avanzada con algunos límites—y el listado de los mismos índices -- comunidades, autores, títulos, temas y fecha de ingreso al sistema. La herramienta de búsqueda es sumamente limitada en sus opciones y su algoritmo no ayuda a obtener resultados precisos, particularmente en el área temática.

La UDEP ofrece sólo búsqueda simple en la página inicial y opciones diversas en cada una de las páginas. Además de la búsqueda simple, UPC ofrece una búsqueda avanzada en 23 campos del registro.

Idioma de la interfaz

UNMSM y UNI utilizan la misma versión y traducción de DSpace. Tienen únicamente dos páginas que no han sido traducidas al castellano: el archivo de ayuda y las etiquetas de la página de estadísticas. UNITRU utiliza la interfaz original en inglés del DSpace y PUCP también utiliza la versión original, pero ha incluido tres elementos con etiquetas en castellano: compartir (para las redes sociales), datos numéricos (total de tesis por tipo de grado) y preguntas frecuentes.

UDEP ofrece la posibilidad de convertir la página original al castellano. La opción se encuentra en las páginas principales, pero es algo inestable y regresa al inglés inesperadamente. Por último, UPC, que también ofrece interfaces en los dos idiomas, tiene una traducción descuidada que produce etiquetas a medio traducir (ej.: listar items by, buscar: start with).

Metadatos en Repositorios de Tesis

Como indicamos en la introducción (p. 3), las *Directrices* de CONCYTEC señalan el uso obligatorio de 12 elementos DC en las descripciones de documentos, 11 de ellos aplicables a la descripción de tesis (2015, pp. 4-10). Estos elementos son la base para elegir la información que ven los usuarios en las interfaces cuando realizan búsquedas. Todos los repositorios tienen una interfaz de resultados con datos mínimos, y una interfaz que describe el documento (registro sencillo) e incluye el enlace al archivo del texto completo, la lista de metadatos y las estadísticas de uso (cuando éstas están disponibles).

La Tabla 3 muestra cuántos elementos DC utiliza cada uno de los repositorios en la descripción completa de los documentos y cuántos se utilizan en el listado de resultados y en la página inicial, así como en el registro sencillo de cada documento.

Tabla 3: Elementos DC empleados por los repositorios estudiados

N° de elementos en	UNMSM	UNI	UNITRU	PUCP	UDEP	UPC

Descripción completa *	11, 1, 4	11, 0, 5	8, 0,3	7, 1, 7	9, 3, 10	8,1, 4
Listado	3	3	5	3	4-5	5 (+1)
Registro sencillo	7	7	5	5 - 6	11-14	11

*la primera cifra corresponde a elementos obligatorios, la segunda a elementos recomendados y la tercera a elementos opcionales

Podemos ver que UNMSM y UNI cumplen con describir todos los elementos considerados obligatorios en las *Directrices*, mientras los otros repositorios omiten algunos. Ninguno de los otros cuatro repositorios utiliza los elementos *dc.source* para registrar el nombre y acrónimo de la institución y el nombre del repositorio. PUCP omite también el elemento *dc.publisher*, de manera que el nombre de la universidad no figura en los elementos obligatorios. Otro elemento que es ignorado por UNITRU y PUCP es *dc.rights* que describe el nivel de acceso al documento; mientras que UDEP usa dos elementos calificados para *dc.rights*, aunque en ninguno de ellos proporciona la información que piden las *Directrices*. UPC sí emplea *dc.rights,* pero no incluye *dc.description.abstract*.

Entre los elementos recomendados, *dc.contributor.advisor* es utilizado por UNMSM, PUCP y UDEP. UDEP es el único que utiliza *dc.identifier.citation* y *dc.format.*

Todos los repositorios utilizan elementos opcionales o adaptados. De estos elementos, los más comunes en todos los repositorios son las diversas alternativas de *dc.date*, que causan gran confusión porque cada repositorio interpreta a su manera lo que significa cada opción. El elemento obligatorio *dc,date.issued*, no ha sido definido en las *Directrices*, situación que favorece la confusión. Un elemento que todos incluyen, pero que no causa ningún impacto en las descripciones es dc.*language*. Hay muy pocas tesis en lengua extranjera y el título las identifica.

De todos los elementos adaptados, los tres añadidos por la PUCP son los más útiles: *dc.thesis.degree.level* (grado), *dc.thesis.degree.grantor* (Universidad. Departamento o Escuela), y *dc.thesis.degree.discipline* (carrera o especialidad). El empleo del segundo de estos elementos proporciona la oportunidad para que figure el nombre de la PUCP en el registro.

No todos los elementos DC empleados en la descripción de los documentos se muestran en las interfaces que utilizan los usuarios. Varios elementos consignan

información repetida. Podemos ver en la Tabla 3 que la mayoría emplea entre 3 y 5 elementos en el listado de resultados --título, autor y fecha, y en algunos casos se agrega el resumen y el nombre de la universidad y el Departamento o Facultad.

El número de elementos que se muestran en el registro sencillo no es tan importante como lo son los elementos con datos erróneos y los que faltan. Alguna importancia tiene el diseño de las páginas, aunque todas las descripciones, excepto la de UPC empiezan por el título y el autor.

El error que más problemas causa es el empleo del elemento que registra la fecha de ingreso al repositorio en lugar de la fecha de creación del documento, sin indicación de qué representa exactamente esa fecha.

Otro problema importante es que algunos repositorios no distinguen al autor de su asesor y lo presentan como coautor, director o colaborador. Igualmente causa dificultad el que no se incluya ni resumen, ni materia.

Excepto en las páginas de la PUCP, es difícil encontrar información sobre el grado obtenido y en qué especialidad. Generalmente hay que buscar en qué colección se encuentra la tesis, o es necesario seguir la línea de ruta de ese registro.

Una de las posibilidades para conocer el impacto de los documentos contenidos en un repositorio, es el análisis de las estadísticas de visitas a las páginas del registro sencillo y al archivo de texto completo. Pero no todos los repositorios tienen un enlace a las estadísticas y cuando las tienen, no existe una guía para interpretar el significado de las etiquetas que definen los datos.

Ni UNITRU ni PUCP proporcionan estadísticas de consulta de sus repositorios. UNMSM, UNI y UDEP presentan las páginas en inglés utilizando los términos: item statistics, visits, views, file downloads. UPC, de cuyas estadística ya hablamos anteriormente, (p. 10) utiliza los mismos términos en sus páginas en inglés y los traduce por estadísticas de elemento, visitas, visualizaciones, descargas de archivos. Las cifras de descarga de archivos son, en muchos casos, exageradamente más elevadas que las visitas; en otros casos sucede justamente a la inversa. Nos preguntamos si el registro sencillo es el medio más

utilizado para acceder al archivo, o si hay otros puntos de acceso que no estamos considerando.

Metadatos en Portales de Revistas

Además de las tesis, la otra gran fuente de documentos de investigación en los repositorios universitarios son los artículos en revistas publicadas por la institución. Aunque hay actualmente una fuerte presión para que los profesores e investigadores depositen todas sus publicaciones en el repositorio institucional, sin tener en cuenta dónde fueron publicadas, el proceso es reciente y no hay datos suficientes en los repositorios para hacer un estudio. Por ello, sólo vamos a analizar los metadatos de artículos en revistas incluidas en los portales respectivos.

No es fácil encontrar todas las versiones digitalizadas de revistas universitarias peruanas, así como los puntos de acceso a ellas. Aquí revisaremos únicamente los portales de UNMSM,[14] UNITRU,[15] PUCP,[16] UPC,[17] quienes utilizan el sistema Open Journals System (OJS) para ofrecer sus revistas digitalizadas y son la fuente que utiliza ALICIA para cosechar los datos.

La única indicación respecto a artículos que encontramos en las *Directrices* emitidas por CONCYTEC se encuentra en el metadato *dc.relation*, de aplicación opcional. Se establece que cuando el *dc.type* sea artículo, en *dc.relation* "se completará este campo indicando la publicación a la que corresponde el artículo (título, número, páginas y año)." (2015, p. 8). Esto significa utilizar para los artículos las mismas indicaciones que para las tesis para ser cosechados para ALICIA, aunque el sistema OJS utiliza únicamente los 15 elementos sin calificar, diferentes a los elementos obligatorios.

Las diferencias fundamentales con la descripción de tesis ya analizada son:

– para los autores se utiliza el elemento *dc.creator*;
– hay un solo elemento para *dc.date*, que es la fecha de ingreso al repositorio;

[14] http://revistasinvestigacion.unmsm.edu.pe/index.php
[15] http://revistas.unitru.edu.pe/
[16] http://revistas.pucp.edu.pe/
[17] http://revistas.upc.edu.pe/

– *dc.subject* recibe únicamente palabras-clave: y,
– *dc.source* se utiliza para el título de la revista, volumen, número y año entre paréntesis, y no se registran las páginas.

Es imposible determinar a simple vista cuáles de las revistas listadas en cada portal han sido digitalizadas e indizadas, en parte o totalmente. Y algunos de los títulos, en un mismo repositorio, pueden ofrecer diferentes detalles y funcionalidades en las páginas de acceso al texto completo. Para describir cada repositorio sería necesario revisar cada título de revista y cada volumen digitalizado. Por lo tanto, únicamente queremos señalar que la página de acceso a los artículos no incluye el número de páginas y hay que volver a la tabla de contenido para encontrarlo. Además, para artículos cuya fecha de ingreso al repositorio no corresponde con la fecha de publicación, la herramienta de citas bibliográficas escoge la fecha de ingreso al repositorio y no la de publicación.

Exportación de datos al repositorio ALICIA

El Repositorio Nacional ALICIA registra los datos tal como aparecen en los repositorios individuales y explícitamente declara no asumir responsabilidad por el contenido ni por posibles errores en las citas bibliográficas. La lista de elementos del cosechador OAI muestra que la fecha de ingreso al repositorio individual es considerada como fecha de publicación y que el campo de autor incluye la mención de la filiación institucional. Estos dos hechos generan automáticamente errores en la creación de citas bibliográficas.

La interfaz es bilingüe y, en varios casos, se pudo observar que no solamente las etiquetas sino también los títulos de los artículos y los resúmenes fueron traducidos al idioma seleccionado para la interfaz, creando la consiguiente confusión.

Conclusiones

1. La correcta aplicación de normas de descripción de documentos, por ejemplo Dublin Core, es clave para el que el contenido de los repositorios individuales pueda ser recogido por los cosechadores de datos de los

consorcios de repositorios nacionales e internacionales, y por las bases de datos bibliográficas gratuitas y por suscripción. Pero únicamente dos de los repositorios analizados proporcionan todos los elementos obligatorios de acuerdo a las directrices peruanas. Además las *Directrices* emitidas se acomodan al registro de entrada de DSpace pero no considera todos los elementos básicos de DC que se utilizan para el registro de artículos de revistas.

2. El elemento *dc.date* que se utiliza con diversas calificaciones, es interpretado localmente, puesto que en las *Directrices* no existen descripciones ni ejemplos para cada una de las alternativas. Y la misma fecha es generalmente asignada a tres elementos diferentes.
3. Para el elemento *dc.subject,* cada repositorio determina si va a emplear el vocabulario controlado del catálogo, o frases, o palabras-clave, incluso números de clasificación. Las búsquedas por materia o tema dan, por lo tanto, resultados insatisfactorios, aun cuando se consulte cada repositorio individualmente.
4. Ciertos elementos opcionales importantes, como *dc.contributor.advisor*, se emplean muy poco, mientras que se emplean otros que duplican información ya presente en otro elemento, por ej. *dc:creator* y *dc:contributor.author*.
5. Ninguno de los repositorios estudiados ofrece en sus páginas suficiente información sobre su desarrollo, o sus políticas, o incluso para determinar la cantidad de consultas al repositorio y sus documentos. Sólo algunos repositorios tienen enlaces para ver las estadísticas de consulta global y de cada documento, pero la falta de explicación sobre la metodología de compilación, el significado de los datos y el período que abarcan, dificultan la interpretación.

Recomendaciones

– CONCYTEC debe revisar la selección de elementos obligatorios y definir con más precisión el significado de las opciones de autor y fecha y elegir los elementos calificados más apropiados en función del objetivo que

deben cumplir, tanto en el caso de la descripción de tesis, como de artículos de revistas.

- Es necesario que las instituciones que forman parte de RENARE cumplan con las exigencias de las directivas respecto a la gobernancia y publicación de estadísticas.
- Los repositorios universitarios deben dar visibilidad prominente a sus repositorios en su página principal. Además deben considerar la adopción de una estructura para sus comunidades y colecciones basada en los mismos principios.
- Los repositorios deben tomar medidas para que el tratamiento de los documentos producidos por digitalización retrospectiva de las colecciones, siga las mismas normas empleadas para los documentos recientes.
- Se debe establecer un mejor control de calidad en el caso de la revistas, para asegurar que *dc.source* y *dc.relation* sean entendidos adecuadamente y proporcionen la cita completa al artículo seleccionado.
- Debe establecerse un mejor control de calidad para asegurar que las normas se apliquen uniformemente y que los módulos de traducción de las interfaces utilicen los mismos términos.
- Se requiere tomar una decisión sobre los vocabularios controlados utilizados en el elemento *dc.subject*, o hacer que el texto completo en *dc.abstract* sea un elemento recuperable.

Referencias

CONCYTEC. Consejo Nacional de Ciencia Tecnología e Innovación Tecnológica. (2015). Directrices para el procesamiento de información en los repositorios institucionales. version 05.15. Disponible en http://portal.concytec.gob.pe/images/documentos/alicia/directrices_repositorio_06-04-2015.pdf

Confederation of Open Access Repositories. (2012). The Current State of Open Access Repository Interoperability (2012). Disponible en https://www.coar-repositories.org/files/COAR-Current-State-of-Open-Access-Repository-Interoperability-26-10-2012.pdf

Library of Congress. Development and MARC Standards Office. (2008). MARC to Dublin Core Crosswalk. Disponible en http://www.loc.gov/marc/marc2dc.html

Medrano, J. F., Figuerola, C. G., & Berrocal, J. L. A. (2012). Repositorios Digitales en España y calidad de Metadatos. *Scire: representación y organización del conocimiento, 18*(2), 109-121. Disponible en http://ibersid.eu/ojs/index.php/scire/article/view/3977 Retrieved from http://ibersid.eu/ojs/index.php/scire/article/view/3977

OCLC. (2013). *Best practices for CONTENTdm and other OAI-PMH compliant repositories, creating sharable metadata* Disponible en http://www.oclc.org/content/dam/support/wcdigitalcollectiongateway/MetadataBestPractices.pdf.

Park, E. G., & Richard, M. (2011). Metadata assessment in e-theses and dissertations of Canadian institutional repositories. *Electronic Library, 29*(3), 394-407. doi:http://dx.doi.org/10.1108/02640471111141124

Pons, D., Hilera, J. R., & Pagés, C. (2013). *La estandarización para la calidad en los metadatos de recursos educativos virtuales*. Paper presented at the IV Congresso Internacional sobre Qualidade e Acessibilidade da Formação Virtual (CAFVIR 2013), Lisboa.

Reglamento de la ley N° 30035, ley que regula el repositorio nacional digital de ciencia, tecnología e innovación de acceso abierto : Decreto Supremo N° 006-2015-PCM, (2015).

Stvilia, B., Gasser, L., Twidale, M. B., Shreeves, S. L., & Cole, T. W. (2004). *Metadata quality for federated collections: IQ Concepts, Models, Case Studies*. Paper presented at the International Conference on Informnation Quality. 9th, Cambridge, MA. Disponible en https://www.ideals.illinois.edu/bitstream/handle/2142/721/iciq_144_final_v1.pdf?sequence=2