Visibilidad y calidad de metadatos en repositorios de universidades peruanas: Propuesta indicadores para evaluación

Resumen

Los repositorios digitales, como plataformas de recolección, gestión, diseminación y preservación de documentos electrónicos, se han consolidado en la base fundamental dentro de esta movimiento Open Access; por lo que, esta investigación analiza el grado de calidad de metadatos (MQ ratio) y el nivel de visibilidad académica en Google Scholar (IGS ratio) asociados a la cobertura de cuatro tipos de documentos (tesis, artículos, libros y conferencias) en 48 repositorios institucionales de universidades peruanas integradas en el Repositorio Nacional Digital Peruano ALICIA. Por ello, se realiza un estudio descriptivo y correlacional de corte transversal con muestreo intencional no probabilística que emplea ALICIA (alicia.concytec.gob.pe) como fuente de datos para seleccionar 48 repositorios de universidades nacionales (n=10) privadas (n=38).

Palabras clave: Repositorios; visibilidad web; metadatos, universidades.

Introducción

Los repositorios digitales, como plataformas de recolección, gestión, diseminación y preservación de material académico producido por una institución, tienen como punto de origen el movimiento de Acceso Abierto (OA), siendo hitos importantes dentro de sus antecedentes las declaraciones de Berlín (2003) y de Budapest (Budapest Open Access Initiative, 2012), pertenecientes como tal a la denominada ruta verde (Green Vía).

Dentro de la tipología de repositorios existentes, se caracteriza la creación del tipo institucional, desarrollados a partir de servicios de gestión relacionados a la colecta, organización, diseminación y preservación de la producción académica de los miembros de una institución (Costa, 2014) los cuales proceden principalmente del apoyo de las universidades a la iniciativa del acceso abierto, dado su rol de espacios destinados a albergar la documentación (académica, docente, institucional, etc.) que produce la universidad (Serrano, 2014). Es así, como indica Chaves (2017), que estos sistemas garantizan una mayor visibilidad de la producción científica y la mejor gestión del conocimiento institucional o temático.

En relación a la tecnología utilizada, la mayoría de los repositorios institucionales peruanos están implementados bajo la plataforma DSpace, y solo uno en otra plataforma como OpenRepository (University of Nottingham, 2017). DSpace se concibió principalmente para alojar repositorios institucionales que utilizan el estándar Dublin Core para registrar metadatos que también se pueden importar utilizando un esquema XML (Barroso, Azevedo & Ribeiro, 2009).

Dentro de algunas iniciativas destacables en el pais tenemos al proyecto Cybertesis (Octubre, 2004) liderada por la Universidad Nacional Mayor de San Marcos (Perú), que permitía el acceso público al texto completo de tesis producidas por esa universidad en versión PDF, HTML y XML replicada en otras instituciones públicas e incluso privadas del ámbito local (Vílchez-Román, 2008). Es sobre estas iniciativas y/o buenas practicas que se promulga la Ley del Repositorio Nacional (Ley Nº 30035 , 2013), creándose el Recolector y Repositorio Nacional de Ciencia y Tecnología (denominado ALICIA).

El Repositorio Nacional Digital (ALICIA) entró en funcionamiento el 06 de mayo 2014 tras la promulgación de la Ley Nº 30035, Ley promulgada el 5 de junio del 2013, que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto (El

Peruano, 2013), quien se encarga de cosechar a los repositorios digitales de las instituciones de educación superior pública y privada, así como a organismos no gubernamentales y dependencias estatales.

A traves del Repositorio Nacional, Perú participa en el proyecto LA Referencia (Red Federada de Repositorios Institucionales de Publicaciones Científicas de Latino América), para lo cual el CONCYTEC (Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica) adoptó al modelo Dublin Core como estándar para el tratamiento de metadatos y DRIVER Guidelines 2.0. como directrices para exposición de recursos textuales con el protocolo OAI-PMH, mecanismo para generar archivos interoperables y en el intercambio de metainformación con otros sistemas y recolectores de metadatos. (Bueno-de-la-fuente, et al, 2009). Ambos lineamientos adoptados por CONCYTEC están dirigidos a gestores y administradores de repositorios digitales, quienes son los responsables del cumplimiento de estos y otros requisitos para la adhesión de un repositorio institucional al Repositorio Nacional Digital (ALICIA).

En ese sentido, las universidades con un repositorio institucional implementado y gestionado bajo los lineamientos de Dublin Core y Driver 2.0, y que solicitan su inclusión en ALICIA a CONCYTEC, pasan por un proceso de evaluación de tres aspectos:

a) **Configuración:** Revisión de instalación y configuración del DSpace según aspectos técnicos sugeridos.

b) **Metadatos:** Verificación de los metadatos registrados por cada documento publicado y de acuerdo al tipo de documento (ver Tabla 1).

c) **Contenido:** Se verifica que el contenido sea igual a los metadatos registrados

Tabla 1. Metadatos requeridos por CONCYTEC

N	Metadata name	Metadata type
1	dc.contributor.author	Obligatorio
2	dc.contributor.advisor	Obligatorio *
3	dc.title	Obligatorio
4	dc.date.issued	Obligatorio
5	dc.publisher	Obligatorio
6	dc.identifier.uri	Obligatorio
7	dc.type	Obligatorio

8	dc.languaje.iso	Obligatorio
9	dc.rights	Obligatorio
10	dc.rights.uri	Obligatorio **
11	dc.subject	Obligatorio
12	dc.source	Obligatorio
13	dc.source	Obligatorio
14	dc.description.abstract	Obligatorio
15	dc.description.uri	Obligatorio *
16	thesis.degree.level	Obligatorio *
17	thesis.degree.name	Obligatorio *
18	thesis.degree.grantor	Obligatorio *
19	thesis.degree.discipline	Obligatorio *
20	dc.format	Recomendado
21	dc.relation	Opcional / Obligatorio ***
22	dc.relation.uri	Opcional / Obligatorio ***
23	thesis.degree.program	Recommended*
24	dc.identifier.journal	Obligatorio ****
25	dc.description.peer-review	Obligatorio ****

* Si el tipo de documento es una tesis.
** Si el metadato dc.rights es de Acceso Abierto.
*** Si el tipo de documento es un dataset.
**** Si el tipo de documento es un artículo.

Para julio de 2018, ALICIA cuenta con más de 160 mil registros de un total de 100 de instituciones integradas de los cuales 69 pertenecen a universidades y 25 a instituciones no gubernamentales y estatales (CONCYTEC, 2018).

La problemática en relación a lo expuesto se gesta en función a los registros producidos por las instituciones y que finalmente son cosechados, considerando ciertos criterios ya señalados, en ALICIA y por otro lado el grado de indizacion que estos repositorios pueden obtener con los mismos registros en Google Scholar (GS), donde aplicando ciertas practicas se consigue una cobertura parcial o total. Se considera pertinente este último ya que es valorado como indicador en algunas rankings u observatorios dedicados al seguimiento de

la actividad cientifica y con ello la visibilidad que los documentos y sus instituciones productoras puedan alcanzar.

Finalmente, el artículo pretende analizar el grado de calidad de metadatos (que llamamos MQ ratio) y el nivel de visibilidad académica en Google Scholar (que denominamos IGS ratio) asociados a la cobertura de cuatro tipos de documentos (tesis, artículos, libros y conferencias) en 48 repositorios institucionales de universidades públicas y privadas peruanas integradas en el Repositorio Nacional Digital peruano ALICIA.

Materiales y metodología

Estudio descriptivo y correlacional de corte transversal con muestreo intencional no probabilística.

Unidades de análisis: La población de análisis se constituye por 69 repositorios (Tabla 2) de universidades públicas y privadas integrados en ALICIA (alicia.concytec.gob.pe).

A partir del total, se seleccionaron 48 repositorios de universidades (Nacionales = 10; Privadas = 38) que cumplieron con los criterios de inclusión:

- Universidad con repositorio institucional o de tesis en ALICIA.
- Repositorio disponible al día de la colecta de datos.

Table 2. Análisis descriptivo por tipo de universidad.

University type	Statistic	Metadata quality	Indexation ratio	Thesis coverage	Article coverage	Book coverage	Conference coverage
Nacional	Average	0.686	0.430	0.673	0.001	0	0
	Median	0.667	0.317	0.667	0	0	0
	Minimum	0.281	0.160	0.279	0	0	0
	Maximum	0.910	0.908	0.908	0.004	0.003	0
	Rank	0.629	0.748	0.629	0.004	0.003	0
	SD	0.209	0.253	0.223	0.001	0.001	0
	Variance	0.044	0.064	0.050	0	0	0
Privada	Average	0.608	0.599	0.425	0.060	0.040	0.029

Median	0.651	0.617	0.421	0	0	0
Minimum	0.047	0.062	0	0	0	0
Maximum	0.974	0.938	0.943	0.853	0.242	0.682
Rank	0.927	0.875	0.943	0.853	0.242	0.682
SD	0.279	0.220	0.309	0.166	0.075	0.115
Variance	0.078	0.048	0.095	0.028	0.006	0.013

Recolección de datos.

Varios estudios, como Heather (2016) que analiza la calidad de los metadatos en los repositorios de Estados Unidos o Bijan (2016) quien estudia los global repositories in LIS domain, enfocados al análisis y evaluación repositorios emplean a OpenDOAR, un directorio autorizado de repositorios académicos de acceso abierto, como fuente de datos para muestreo, porque cada repositorio de OpenDOAR ha sido visitado por el personal del proyecto para verificar la información que se registra aquí (University of Nottingham, 2014). Sin embargo, observamos que el portal aparece descontinuado desde abril de 2014, por el cual no fue considerado como fuente de datos para el estudio.

Por otro lado, Webometrics, el Ranking Mundial de Repositorios, iniciativa del Laboratorio de Cibermetría, un grupo de investigación perteneciente al Consejo Superior de Investigaciones Científicas (CSIC), el organismo de investigación público más grande de España (Webometrics, 2017), a pesar de ser una fuente que emplea indicadores rigurosos para la evaluación de repositorios y presenta una actualización semestral (enero y julio), para el caso peruano, los repositorios de universidades no se encuentran totalmente representados.

Por estas razones, y al tratarse de un estudio local que evalúa a los repositorios institucionales de universidades peruanas, empleamos al cosechador nacional de repositorios ALICIA como fuente de datos; además, a razón de que los repositorios incluidos en este portal pasaron por un proceso de evaluación de calidad de los metadatos, que es un indicador que se pretende medir en este estudio.

La colecta de datos se realizó en junio de 2018.

Procesamiento y análisis.

Se utilizaron como criterios de evaluación tres indicadores: calidad de metadatos, indización en Google Scholar, y cobertura del tipo de documento tesis. Cada uno de los tres indicadores de análisis derivó de un cálculo inicial de división entre dos datos:

mqr = #items en ALICIA/ #items en el Repositorio... (1)
igsr = #items en GS / #items en el Repositorio... (2)
tcr = #items de tesis en ALICIA/ #items en el Repositorio... (3)

Donde:

mqr : Metadata quality ratio.
igsr : Indexation in Google Scholar ratio.
tcr : Thesis coverage ratio.

Para determinar el supuesto de normalidad en los indicadores (1), (2), y (3), empleamos la prueba de Shapiro-Wilk para muestras < 50 casos. La prueba de normalidad arrojó que los valores de los indicadores de análisis no se distribuían de manera normal, por lo que se usó U de Mann-Whitney, el equivalente no paramétrico de la Prueba T para la diferencia de dos medias en muestras independientes con supuestos de no normalidad en las poblaciones.

Finalmente, para analizar la asociación entre variables de estudio se empleó el coeficiente de correlación rho de Spearman al identificar la no distribución normal de las tasas de análisis.

Resultados

Análisis descriptivo

Calidad de los metadatos.

El indicador metadata quality ratio (MQ ratio) arrojó una mediana de 0,67 [RIC: 0,552 - 0,891] para las universidades nacionales y una mediana de 0,65 [RIC: 0,407 - 0,838] para las universidades privadas; con una diferencia estadísticamente no significativa (U de Mann Whitney Test: p-value=0,542).

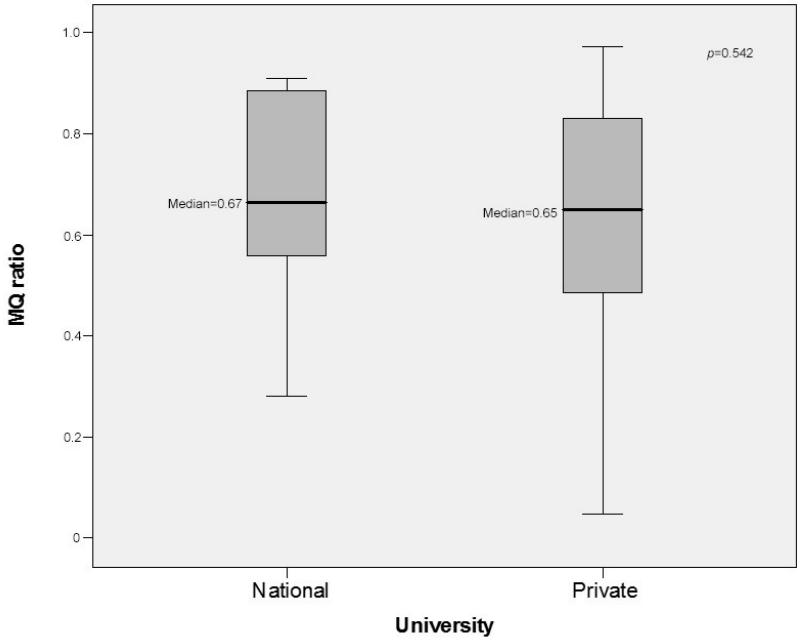

Figura 1. Proporción de calidad de metadatos por universidades peruanas.

Visibilidad web.

Para el indicador metadata quality ratio (MQ ratio) arrojó una mediana de 0,32 [RIC: 0,241 - 0,596] para las universidades nacionales y una mediana de 0,62 [RIC: 0,464 - 0,749] para las universidades privadas; con una diferencia estadísticamente no significativa (U de Mann Whitney Test: p-value=0,054).

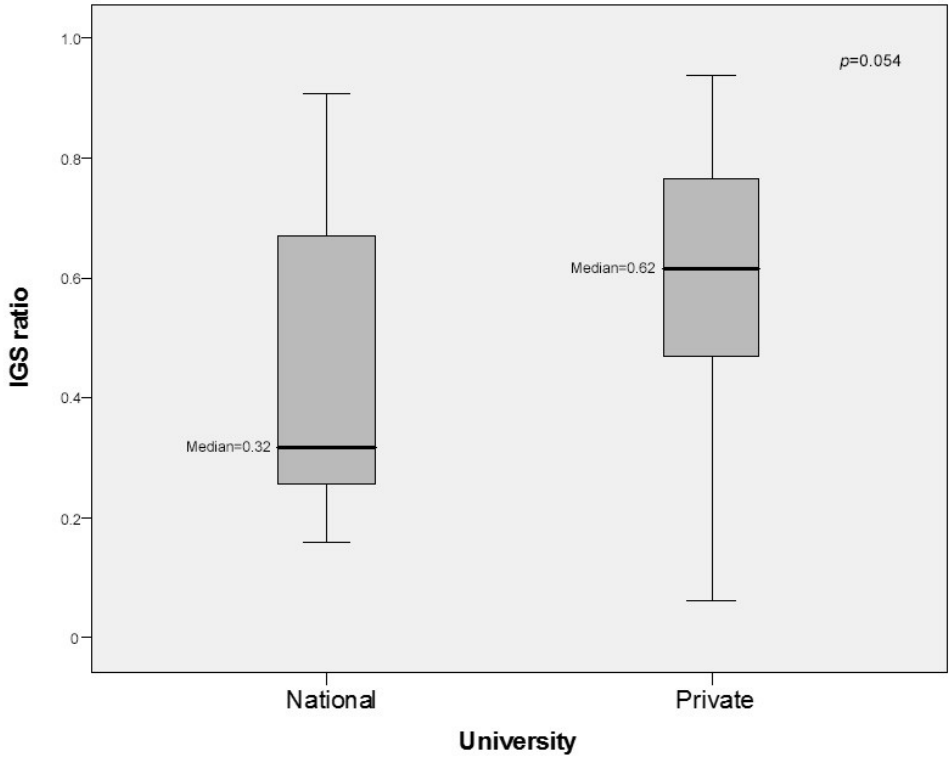

Figura 2. Índice de visibilidad web académica por universidades peruanas.

Cobertura de tipo de documento: tesis

El indicador thesis coverage ratio (TC ratio) arrojó una mediana de 0,67 [RIC: 0,494- 0,891] para las universidades nacionales y una mediana de 0,42 [RIC: 0,115 - 0,679] para las universidades privadas; sin embargo, para este indicador sí existe una diferencia estadísticamente significativa (U de Mann Whitney Test: p-value=0,021).

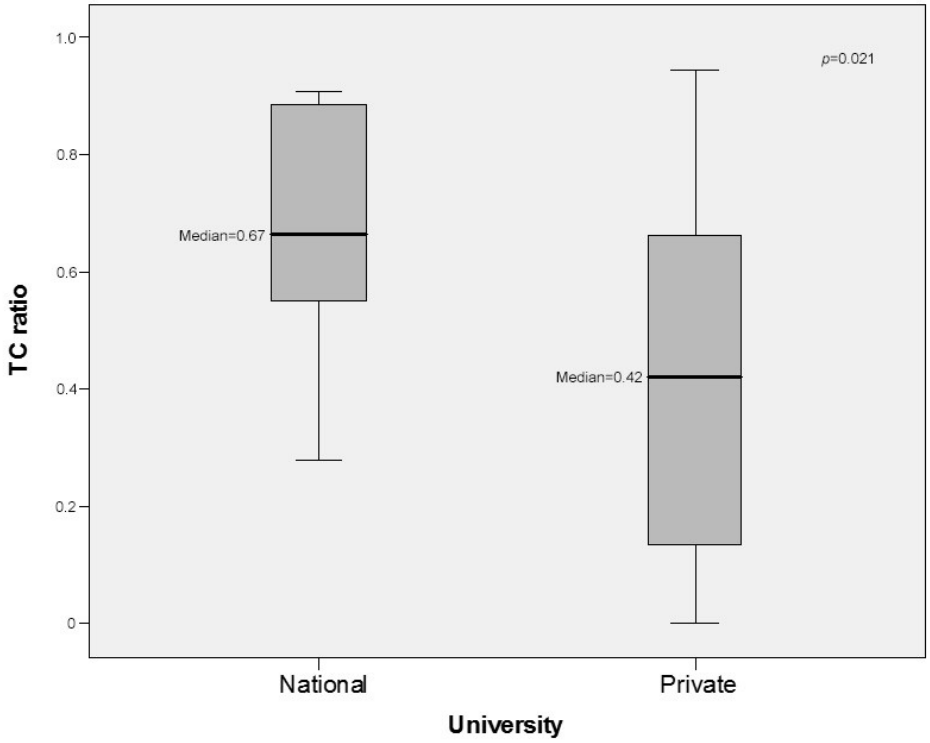

Figura 3. Cobertura del índice de tesis por universidades peruanas.

Análisis correlacional

El análisis de correlación ρ de Spearman entre los indicadores MQ and IGS ratio y tres indicadores de cobertura por tipo de documento (thesis, article, book, conference), para todos los repositorios de la muestra, arroja una correlación moderada (ρ=0,594; P<0,01) a nivel de Metadata quality y solo con un indicador de tipo de documento (Thesis). Además, como muestra la Tabla 3, existe baja correlación (ρ=0,157) entre la proporción de indización de documentos en Google Scholar y la proporción de documentos cosechados en ALICIA con respecto a la cantidad de documentos en el repositorio de origen.

Table 3. Análisis de correlación de Spearman por ratios de estudio

Variable	1	2	3	4	5	6
1. Calidad de Metadata	1	0.157	**0.594****	0.080	0.100	0.41
2. Indexacion en Google Academico		1	0.109	-0.340*	-0.020	0.73
3. Thesis coverage			1	-0.245	-0.457**	-0.220
4. Article coverage				1	**0.432****	0.301

5. Book coverage	1	**0.411**
6. Conference coverage		1

*Significant at the 0.05 level (2-tailed)
**Significant at the 0.01 level (2-tailed)

Por otro lado, los resultados muestran que existe moderada correlación entre varias tasas de cobertura por tipos de documentos como entre artículo y libro (ρ=0,432; P<0,05) y libro y conferencia (ρ=0,411). Sin embargo, se observa nula asociación entre cobertura por tipo de documento y la tasa de indexación en Google Scholar.

Propuesta de plataforma de evaluación

Dada la pertinencia del monitoreo constante de la producción académica o científica resultante de las actividades de universidades o instituciones dedicadas a la investigación, se plantea la implementacion del observatorio digital, denominado Media Lab.

Infraestructura tecnológica.

La plataforma se albergará en servidores de la Facultad de Letras y Ciencias Humanas (UNMSM) bajo el sistema de gestión de contenidos (CMS) WordPress. El proyecto de observatorio que mostrará los datos resultantes de la investigación, forma parte del proyecto Media Lab UNMSM, cuya dirección web se encuentra aparcada en http://medialab.letras.unmsm.edu.pe/

Figura 4. Imagen y texto de ingreso desde al Home del Media Lab UNMSM hacia el portal de propuesta.

Arquitectura de información.

La información que se liberará, a través del portal de evaluación de repositorios universitarios peruanos propuesto, tendrá una distribución de información bajo la siguiente estructura jerárquica:

Header.- Titular del portal.
Description.- Detalle general del portal.
Hiperlink.- Enlace hacia una página que detalla la metodología.
Images.- Tres iconos que indican los aspectos a evaluar.

Figura 5. Imagen y texto de cabecera del portal observatorio propuesto.

Figura 6. Cobertura de analisis

Conclusiones

La mayor proporción de indización en Google Scholar (GS) para las universidades privadas (60%) con respecto a las universidades públicas (43%) supondría que existe un mayor trabajo a favor de la visibilidad de los contenidos en el sector privado; sin embargo, la cantidad de items depositados en los repositorios de universidades públicas supera en gran medida a la cantidad de documentos albergados en un repositorio de universidad particular.

En ambos casos, cerca de la mitad de los repositorios analizados presentan escasa tasa de indización en GS lo que indica un desconocimiento o subestimación de la visibilidad web que puede obtener de GS o debido a la restricción de acceso (embargo y restringido) a los documentos de tesis, casos presentes sobretodo en universidades privadas en el área de las ciencias empresariales.

La proporción de la cantidad de ítems integrados en el repositorio nacional ALICIA con respecto a la totalidad de items de repositorio tanto en universidades públicas (69%) como privadas (61%) indican que alrededor de la mitad de documentos depositados en los repositorios no cumplen con los requerimientos de calidad de los metadatos por CONCYTEC o que se tratan de tipos de documentos no contemplados para la cosecha en ALICIA.

La mayor cobertura de documentos tipo tesis en ALICIA por parte universidades públicas y privadas frente a otros tipos de documentos como artículos, libros y conferencias; indicaría que la implementación de un repositorio institucional por parte de varias universidades peruanas, responde a un sentido de cumplimiento formal en respuesta a los requerimientos de organismos nacionales como SUNEDU o CONCYTEC, mas no al sentido trascendental de poner gestionar y difundir las investigaciones de la institución y que esta es a su vez son una oportunidad para la gesta de colaboraciones, búsqueda de fondos y mejoras en el prestigio institucional.

Bibliografía

Barroso, I., Azevedo, M., & Ribeiro, C. (2009, September). Thematic digital libraries at the University of Porto: Metadata integration over a repository infrastructure. In International Conference on Theory and Practice of Digital Libraries (pp. 392-395). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-04346-8_42

Becerril, A., Espinosa, R.L., Espinosa, J.M.M. (2016). Semantic approach to context-aware resource discovery over scholarly content structured with OAI-PMH | Enfoque semántico para el descubrimiento de recursos sensible al contexto sobre contenidos académicos estructurados con OAI-PMH. Computacion y Sistemas, 20 (1), pp. 127-142. Recuperado de http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2189/2100

Bueno-de-la-Fuente, G., Hernández-Pérez, T., Rodríguez-Mateos, D., Méndez-Rodríguez, E. M., & Martín-Galán, B. (2009). Study on the use of metadata for digital learning objects in university institutional repositories (MODERI). Cataloging & Classification Quarterly, 47(3-4), 262-285. doi: 10.1080/01639370902737315

Chaves, L., & Kafure, I. (2017). Evaluación de la Usabilidad del Repositorio Institucional de la Universidad de Brasília. Revista General De Información y documentación, 27(1), 87-106. doi:10.5209/RGID.56563

CONCYTEC. (2017). Alicia Concytec. http://alicia.concytec.gob.pe

Costa, M. P. da. (2014). Características e contribuições da Via Verde para o acesso aberto à informação científica na América Latina. Brasília, Universidade de Brasília.

DCMI. (2017). Dublin Core® Metadata Initiative (DCMI). http://dublincore.org/

Heather, S., Dykas, F. (2016). High-quality metadata and repository staffing: Perceptions of United States-Based OpenDOAR participants. Cataloging and Classification Quarterly, 54 (2), pp. 101-116.

Bijan, R., Biswas, S.C., Mukhopadhyay, P. (2016). Global repository movement in the domain of library and information science discipline. International Journal of Information Science and Management, 14 (2), pp. 15-32. Cited 1 time.

Ley Nº 30035 (2013). Ley que regula el Repositorio. Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto. El Peruano, Normas Legales, 05 de junio.

Ley Nº 30220 (2014). Ley universitaria. El Peruano, Normas Legales, 08 de julio.

Otto, J. J. (2014). Administrative Metadata for Long-Term Preservation and Management of Resources. Library Resources & Technical Services, 58(1), 4-32. doi: 10.5860/lrts.58n1.4

Serrano, R., Melero, R., Abadal, E. (2014). Indicators for the evaluation of open access institutional repositories | Indicadores para la evaluación de repositorios institucionales de acceso abierto. Anales de Documentacion, 17 (2). Recuperado de http://revistas.um.es/analesdoc/article/view/190821/165851 |

University of Nottingham. (2017). OpenDOAR: Directory of Open Access Repositories. http:// www.opendoar.org/

Vílchez-Román, C., Shimabukuro, D.N. (2008). Usabilidad de un sistema de recuperación de información a texto completo: El caso del portal Cybertesis Perú. ACIMED, 17 (3). Recuperado de http://scielo.sld.cu/pdf/aci/v17n3/aci03308.pdf

Walsh, M. P. (2011). Repurposing MARC Metadata for an Institutional Repository: Working with Special Collections and University Press Monographs. Library Resources & Technical Services, 55(1), 33. http://dx.doi.org/10.5860/lrts.55n1.33

Webometrics (2017). Ranking web of repositories.Recuperado de http://repositories.webometrics.info

Zavalina O.L., Kizhakkethil P. & Shakeri S. (2015). Metadata change in traditional library collections and digital repositories: Exploratory comparative analysis. Proceedings of the Association for Information Science and Technology, 52 (1) , pp. 1-5.