Metadata quality and academic visibility associated to document type coverage on institutional repositories in Peruvian universities

**Abstract:**
This article analyzes level of metadata quality (MQ ratio) and level of academic visibility in Google Scholar (IGS ratio) associated with coverage of four types of documents (theses, articles, books and conferences) in repositories of Peruvian universities. This research is a cross-sectional descriptive and correlational study with intentional non-probabilistic sampling that analyzes 48 repositories from national (n = 10) and private universities (n = 38) integrated in the Peruvian National Digital Repository Alicia (alicia.concytec.gob.pe). Regarding the MQ ratio, we found a median of 0.67 [RIC: 0.552-0.891] for national universities and a median of 0.65 [RIC: 0.407-0.838] for private universities (p = 0.542). Regarding the IGS ratio, we found a median of 0.32 [RIC: 0.241-0.596] for national universities and a median of 0.62 [RIC: 0.464-0.749] for private universities (p = 0.054). The p-value in Spearman's rank correlation shows a moderate correlation (ρ = 0.594; P <0.01) between MQ ratio and the thesis coverage indicator, and a low correlation (ρ = 0.157) between the index of document indexing in Google Scholar and the proportion of documents harvested in Alicia. We conclude that the highest proportion of academic visibility is concentrated in private universities and the metadata quality number of items integrated in Alicia favors public universities.

**Note:** The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.

- fig01.psd
- fig02.psd
FIGURES - Metadata quality and academic visibility associated to document type coverage on institutional repositories in Peruvian universities

Figure 1. *Metadata quality ratio by Peruvian universities.*

![Box plot showing metadata quality ratio for national and private universities in Peru. Median for national universities is 0.67, and 0.65 for private universities. The p-value is 0.542.]

Figure 2. *Academic web visibility ratio by Peruvian universities.*
Figure 3. Coverage of document type thesis ratio by Peruvian universities
Metadata quality and academic visibility associated to document type coverage on institutional repositories in Peruvian universities

Metadata package required by CONCYTEC for institutional repositories management

<table>
<thead>
<tr>
<th>Metadata name</th>
<th>Metadata type</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.contributor.author</td>
<td>Required</td>
</tr>
<tr>
<td>dc.contributor.advisor</td>
<td>Required *</td>
</tr>
<tr>
<td>dc.title</td>
<td>Required</td>
</tr>
<tr>
<td>dc.date.issued</td>
<td>Required</td>
</tr>
<tr>
<td>dc.publisher</td>
<td>Required</td>
</tr>
<tr>
<td>dc.identifier.uri</td>
<td>Required</td>
</tr>
<tr>
<td>dc.type</td>
<td>Required</td>
</tr>
<tr>
<td>dc.language.iso</td>
<td>Required</td>
</tr>
<tr>
<td>dc.rights</td>
<td>Required</td>
</tr>
<tr>
<td>dc.rights.uri</td>
<td>Required **</td>
</tr>
<tr>
<td>dc.subject</td>
<td>Required</td>
</tr>
<tr>
<td>dc.source</td>
<td>Required</td>
</tr>
<tr>
<td>dc.source</td>
<td>Required</td>
</tr>
<tr>
<td>dc.description.abstract</td>
<td>Required</td>
</tr>
<tr>
<td>dc.description.uri</td>
<td>Required *</td>
</tr>
<tr>
<td>thesis.degree.level</td>
<td>Required *</td>
</tr>
<tr>
<td>thesis.degree.name</td>
<td>Required *</td>
</tr>
<tr>
<td>thesis.degree.grantor</td>
<td>Required *</td>
</tr>
<tr>
<td>thesis.degree.discipline</td>
<td>Required *</td>
</tr>
<tr>
<td>dc.format</td>
<td>Recommended</td>
</tr>
<tr>
<td>dc.relation</td>
<td>Optional / Required ***</td>
</tr>
<tr>
<td>dc.relation.uri</td>
<td>Optional / Required ***</td>
</tr>
</tbody>
</table>
Descriptive analysis by university type.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Metadata quality</th>
<th>Indexation ratio</th>
<th>Thesis coverage</th>
<th>Article coverage</th>
<th>Book coverage</th>
<th>Conference coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>0.686</td>
<td>0.430</td>
<td>0.673</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Median</td>
<td>0.667</td>
<td>0.317</td>
<td>0.667</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.281</td>
<td>0.160</td>
<td>0.279</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.910</td>
<td>0.908</td>
<td>0.908</td>
<td>0.004</td>
<td>0.003</td>
<td>0</td>
</tr>
<tr>
<td>Rank</td>
<td>0.629</td>
<td>0.748</td>
<td>0.629</td>
<td>0.004</td>
<td>0.003</td>
<td>0</td>
</tr>
<tr>
<td>SD</td>
<td>0.209</td>
<td>0.253</td>
<td>0.223</td>
<td>0.001</td>
<td>0.001</td>
<td>0</td>
</tr>
<tr>
<td>Variance</td>
<td>0.044</td>
<td>0.064</td>
<td>0.050</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Average</td>
<td>0.608</td>
<td>0.599</td>
<td>0.425</td>
<td>0.060</td>
<td>0.040</td>
<td>0.029</td>
</tr>
<tr>
<td>Median</td>
<td>0.651</td>
<td>0.617</td>
<td>0.421</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.047</td>
<td>0.062</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.974</td>
<td>0.938</td>
<td>0.943</td>
<td>0.853</td>
<td>0.242</td>
<td>0.682</td>
</tr>
<tr>
<td>Rank</td>
<td>0.927</td>
<td>0.875</td>
<td>0.943</td>
<td>0.853</td>
<td>0.242</td>
<td>0.682</td>
</tr>
<tr>
<td>SD</td>
<td>0.279</td>
<td>0.220</td>
<td>0.309</td>
<td>0.166</td>
<td>0.075</td>
<td>0.115</td>
</tr>
<tr>
<td>Variance</td>
<td>0.078</td>
<td>0.048</td>
<td>0.095</td>
<td>0.028</td>
<td>0.006</td>
<td>0.013</td>
</tr>
</tbody>
</table>
Spearman correlation analysis between study rates

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata quality</td>
<td></td>
<td>0.157</td>
<td>0.594**</td>
<td>0.080</td>
<td>0.100</td>
<td>0.41</td>
</tr>
<tr>
<td>on Google Scholar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coverage</td>
<td></td>
<td>1</td>
<td>0.109</td>
<td>-0.340*</td>
<td>-0.020</td>
<td>0.73</td>
</tr>
<tr>
<td>Thesis coverage</td>
<td></td>
<td></td>
<td>-0.245</td>
<td>0.457**</td>
<td>-0.220</td>
<td></td>
</tr>
<tr>
<td>Article coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.432**</td>
</tr>
<tr>
<td>Book coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.411</td>
</tr>
<tr>
<td>Conference coverage</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

*Significant at the 0.05 level (2-tailed)  **Significant at the 0.01 level (2-tailed)
Metadata Quality and Academic Visibility Associated with Document Type Coverage in Institutional Repositories of Peruvian Universities

ALHUAY-QUISPE, Joel*
Affiliation: Universidad San Ignacio de Loyola, Lima, Perú.
ORCID: http://orcid.org/0000-0002-1903-4687

QUISPE-RIVEROS, David
Affiliation: Universidad Nacional Mayor de San Marcos, Lima, Perú.
ORCID: https://orcid.org/0000-0002-0518-1512

BAUTISTA-YNOFUENTE, Lourdes
Affiliation: Universidad ESAN, Lima, Perú.
ORCID: http://orcid.org/0000-0002-0277-489X

PACHECO-MENDOZA, Josmel
Affiliation: Universidad San Ignacio de Loyola, Lima, Perú.
ORCID: https://orcid.org/0000-0002-2251-8092

*Corresponding author:
Email: joel.alhuayq@gmail.com
Address: El Cubo building, 2nd floor. Universidad San Ignacio de Loyola, La Fontana 750 Avenue, La Molina, Lima - Peru.

Abstract
This article analyzes level of metadata quality (MQ ratio) and level of academic visibility in Google Scholar (IGS ratio) associated with coverage of four types of documents (theses, articles, books and conferences) in repositories of Peruvian universities. This research is a cross-sectional descriptive and correlational study with intentional non-probabilistic sampling that analyzes 48 repositories from national (n = 10) and private universities (n = 38) integrated in the Peruvian National Digital Repository Alicia (alicia.concytec.gob.pe). Regarding the MQ ratio, we found a median of 0.67 [RIC: 0.552-0.891] for national universities and a median of 0.65 [RIC: 0.407-0.838] for private universities (p = 0.542). Regarding the IGS ratio, we found a median of 0.32 [RIC: 0.241-0.596] for national universities and a median of 0.62 [RIC: 0.464-0.749] for private universities (p = 0.054). The p-value in Spearman's rank correlation shows a moderate correlation (ρ = 0.594; P <0.01) between MQ ratio and the thesis coverage indicator, and a low correlation (ρ = 0.157) between the index of document indexing in Google Scholar and the proportion of documents harvested in Alicia. We conclude that the highest proportion of academic visibility is concentrated in private universities and the metadata quality number of items integrated in Alicia favors public universities.

Keywords: Digital repositories; web visibility; metadata; Google Scholar; thesis; articles; books; Peruvian universities.

Main theme: Evaluation of search engine indexing on metadata or digital file formats

Introduction
The creation of institutional digital repositories is based on their fundamental role as a place to keep documentation (e.g. academic, teaching, institutional, etc.) by an institution that supports the Open Access initiative (Serrano Vicente, Melero Melero, and Abadal 2014), and they are developed from management services related to collection, organization, dissemination and preservation of academic publications in an institution (Costa, 2014), foundations already set out in Berlin (2003) and Budapest statements (BOAI, 2012) of the Open Access movement (OA). In consequence, Chaves (2017) refers to these systems ensuring greater visibility of scientific production and better management of institutional or thematic knowledge.

For the Peruvian case, a pioneering initiative that fostered the creation of institutional repositories was launched, called the Cybertesis Project (October, 2004). It restricted the publication of one type of document: theses. As Cerda (2016) says, Central Library "Pedro Zulen" presented the Cybertesis Project to Universidad Nacional Mayor de San Marcos (UNMSM) in 2001, and it was approved in 2002 under a rectoral resolution. This project was intended to host electronic theses produced by the UNMSM, the oldest university in South America and the most important Peruvian public university.

Cybertesis Project started with the cooperation among the University of Montreal (Canada), University of Lyon (France), University of Chile, and was replicated in Peru by the UNMSM, which in turn led the project and allowed its development in other Peruvian universities (Vílchez-Román and Nakamura Shimabukuro 2008). The application of the Cybertesis methodology allowed access to full text in two file formats: PDF and HTML. Also, it enabled automatic generation of XML files to ensure the conservation and preservation of digital theses.

Then, given evolution platforms and the adoption of policies and good practices, and other notable cases were influenced by the Cybertesis Project. Repository of the Pontificia Universidad Católica del Peru (PUCP) was launched in 2013 (Yañez, 2014), an ambitious project under the responsibility of the Technical Advisory Rector’s Office that started in 2010.

In 2013, the Academic Repository of the Universidad Peruana de Ciencias Aplicadas (UPC) was released (Eléspuru 2016). This repository uses Open Repository (OR) as a platform, which is a modification of DSpace (version 5), the Handle system as a persistent identifier, the protocol for content transmission (OAI-PMH) and Dublin Core metadata (qualified).

After this period of initiatives and good practices in several universities, the National Repository Law (Congress of the Republic, Peru 2013) is published. This law regulates the National Digital Repository of Science, Technology and Technological Innovation.

Metadata Management in Digital Repositories
There are several standards for metadata management in digital repositories. Otto (2014) identifies those 54 repositories from the Association of Research Libraries (ARL) using Qualified Dublin Core, Simple Dublin Core, Metadata Object Description Schema (MODS), PREMIS, NISO MIX (Z39.87) and MARC as standardized instruments to describe their digital objects deposited.

Dublin Core (DC) is a set of elements or properties that allows the semantic description of a wide range of information resources in the web, which allows registering and retrieving books, theses, research projects, conference presentations, multimedia documents, etc. (Dublin Core Metadata Initiative 2017). The so-called simple Dublin Core (unqualified) is the one originally used to describe resources and digital objects using 15 metadata: dc: title, dc: creator, dc: subject, dc: description, dc: publisher, dc: contributor, dc: date, dc: type, dc: format, dc: identifier, dc: source, dc: language, dc: relation, dc: coverage, dc: rights (Becerril García, Lozano Espinosa, and Molina Espinosa 2016)

Metadata management with Dublin Core in digital repositories implemented with platforms such as Dspace allows the interoperability between systems, migration (import and export) from traditional bibliographic description systems such as MARC (Machine Readable Cataloging) (Walsh 2011).

The adoption of a standard in metadata management ensures interoperability between technological platforms, through information transfer protocols such as Open Archives Initiative (OAI-PMH) for Metadata Harvesting. This protocol appears with the Open Archives Initiative published in 1999, whose first public version was in 2001, due to the need to convert files into interoperable files and build information retrieval services from many repositories (Becerril García, Lozano Espinosa, and Molina Espinosa 2016). In that sense, Open Access digital repositories have implemented the OAI protocol as a mechanism to achieve interoperability in the exchange of meta-information with other systems such as the metadata harvesters (Bueno-De-La-Fuente et al. 2009).

**Interoperability between Peruvian Institutional Repositories**

Most Peruvian institutional repositories were implemented under the DSpace platform, and only one in the Open Repository (University of Nottingham 2017). DSpace was created mainly to accommodate institutional repositories that use the Dublin Core standard to register metadata that can also be imported using an XML Scheme (Barroso, Azevedo, and Ribeiro 2009).

The Peruvian National Repository Alicia (Free Access to Scientific Information) is a National Digital Repository of Science, Technology and Innovation of Peru, and it was operational on May 6, 2014 by means of Law No. 30035 dated June 5, 2013. This law regulates the National Digital Repository of Science, Technology and Technological Innovation of Open Access
It integrates public and private higher education institutions, as well as non-governmental organizations and state agencies.

It was initially implemented in Open Knowledge Harvester Systems (https://pkp.sfu.ca/ohs) of the Public Knowledge Project (May 2014 - February 2015), and then it adopted the technology of VuFind (https://vufind.org/), developed and maintained by Memorial Falvey Library of the University of Villanova (the first two months of 2015). VuFind is based on Apache Solr search and on Java, using an index instead of a relational database, thereby enabling faster data retrieval (Katz and Nagy 2014). Alicia is the national node, which together with other national harvesters in Latin America, joins the regional collector project called "La Referencia" (Federated Network of Institutional Repositories of Scientific Publications of Latin America).

Peru, for its participation in the La Referencia Project, through CONCYTEC (National Committee in Science, Technology and Technological Innovation) adopted Dublin Core as a standard for metadata management and DRIVER Guidelines 2.0 as guidelines for exposing textual resources with OAI-PMH protocol. Both guidelines adopted by CONCYTEC are aimed at managers and administrators of digital repositories, who are responsible for the fulfillment of the requirements for the incorporation of institutional repositories into the Digital National Repository (Alicia).

In that sense, universities with an institutional repository implemented and managed under Dublin Core and Driver 2.0 guidelines, and which ask CONCYTEC their inclusion in Alicia, undergo an evaluation process of three points:

a) Configuration: Installation and configuration review of DSpace according to suggested technical aspects.

b) Metadata: Verification of metadata quality registered by each item published according to document type (Table 1).

c) Content: It verifies that document content is equal to the metadata registered.

By March 2017, Alicia included more than 61,000 records from 94 integrated institutions, of which 69 belong to universities and 25 to non-governmental and state institutions (National Committee in Science, Technology and Technological Innovation. Peru, 2017).

Google Scholar as an Academic Visibility Tool

For purposes of research, the coverage of theses and other types of documents such as articles, books and conferences will be taken into account. In the specific case of research works (thesis), they should be available, according to the New University Law 30220, in the repository of the National Registry of Research (RENAITI), which uses this information from Alicia (Congress of the Republic. Peru. 2014).

According to different works developed in Google Scholar (appeared since 2004), it is one of...
the first options and main tools used by a great group of researchers who look for information in the web, as stated by Orduña-Malea (2014). It should be borne in mind that Google and Google Scholar have become the gateway for users searching for academic information.

Google Scholar is often used in the same proportion as classic catalogs of libraries or other specialized search engines. It is important to point out that for the institutional repository to have presence in Google Scholar, it depends on the metadata scheme used, the technological platform and the criteria and strategies applied for that purpose (Orduña-Malea, 2014).

Google Scholar is often used in the same proportion as classic catalogs of libraries or other specialized search engines. But it is important to point out that for the institutional repository to have presence in Google Scholar, it depends on the metadata scheme used, the technological platform and the criteria and strategies applied (Orduña-Malea, 2014); although, in some cases Google Scholar discover few or nothing PDF files published in digital collections (Yang, 2016).

Regarding information retrieval, Google Scholar crawlers systematically analyze the entire academic web, without making distinctions between thematic areas, languages or countries, allowing the estimate of impact metrics for a wider collection of documents (Martin-Martin 2016). Under these considerations, GS is considered as the means of data analysis.

An important reference in the research on the IR in GS of the publications managed in digital repositories is of Arlitsch and O’Brien (2012), who among other subjects, had obtained as a result with the indexation in GS that: whether the IR has provided crawlers an efficient method to access its scholarly papers; and whether acceptable metadata schemas are provided that offer precise bibliographic information within the HTML page header tags.

Taking into account these experiences, and the current context in Peru regarding publications in institutional repositories and the popularity and validity of GS as tool used by researchers, and as a reference of web presence of scientific production of an academic institution, it is which takes into account the importance of retrieval analysis in GS of a particular publication type: theses.

Objectives

This article aims to analyze metadata quality degree (which we call the MQ ratio) and academic visibility level in Google Scholar (which we call IGS ratio) associated with the coverage by four document types (theses, articles, books and conferences) from 48 institutional repositories of Peruvian public and private universities integrated in the Peruvian National Digital Repository called Alicia.

The study aims to approximate that proportion of metadata quality is associated with only one document type (thesis) over others (articles, books). That is, the number of documents incorporated in repositories and that are harvested into Alicia are predominantly theses. The national situation of the education quality evaluation led to the creation of National Repository of Research Works - RENATI.

As a general focus, this research aims to answer:
• Is the relationship between MQ ratio and TC ratio significant?
• Is the relationship between IGS ratio and MQ ratio significant?
• Is the difference between private and public universities for MQ ratio significant?
• Is the difference between private and public universities for IGS ratio significant?
• Is the difference between private and public universities for TC ratio significant?

Methodology

Methods.
Cross-sectional descriptive and correlational study with non-probabilistic intentional sampling.

Units of Analysis
According to the analysis, there are 69 repositories (supplementary material) from public and private universities integrated in Alicia (alicia.concytec.gob.pe). Of them, only 48 repositories from national (n = 10) and private universities (n = 38) were selected according to two inclusion criteria:
- University with institutional or digital theses repository in Alicia.
- Repository site available at the time of collection data.

Data Collection
Several studies focused on analysis and evaluation repositories, such as the study made by Sandy and Dykas (2016) that analyzes the metadata quality in repositories from US, or such as the study conducted by Roy, Biswas, and Mukhopadhyay (2016) that analyzes global repositories in LIS domain, taking OpenDOAR as a data source for sampling. OpenDOAR is an authoritative directory of academic Open Access repositories, because each repository listed has been revised by OpenDOAR project staff in order to check the information recorded (University of Nottingham, 2014). However, we note that the portal has been discontinued since April 2014, for that reason, we do not considered it as a data source.

The World Repositories Ranking Webometrics, initiative of Cybermetrics Lab, a research group belonging to the CSIC (Spanish National Research Council, 2017), despite being a source that uses rigorous indicators of repository evaluation and presents semi-annual update (January and July), we note that many Peruvian university repositories are not fully represented.

For these reasons, and because this research evaluates institutional repositories of Peruvian universities, we use National Harvester Alicia as a data source. We also use it because repositories included in Alicia were evaluated at metadata quality level, which is an indicator of measure in this article.

Data was collected on March 26, 2017.

Data Processing and Analysis
Three indicators were used as evaluation criteria: metadata quality, Google Scholar indexing,
and document type coverage. Each indicator of analysis was derived from dividing two data:

\[ MQ_r = \frac{\text{#items in Alicia}}{\text{#items in Repository}} \quad (1) \]

\[ IGS_r = \frac{\text{#items in GS}}{\text{#items in Repository}} \quad (2) \]

\[ DC_r = \frac{\text{#items in Alicia}}{\text{#items in Repository}} \quad (3) \]

Where:

MQ: Metadata quality ratio.
IGS: Indexing in Google Scholar ratio.
DC: Document thesis coverage ratio.

Records counts in every repository (#items in Repository) correspond to the number of documents found at query time of data collection. The query was performed using an advanced search interface of the system.

The number of records harvested in Alicia (#items in Alicia) corresponds to the number of documents displayed in the National Repository portal for each institution and only by the institutional repository, excluding journal repositories.

From records counted as items in Alicia, the number of thesis documents (#items theses in Alicia) was identified with "format" filter of the National Repository portal. Three thesis-type documents were considered for this count: bachelor thesis, master thesis and doctoral thesis.

The number of documents indexing in Google Scholar (#items in GS) for each analyzed repositories was extracted by direct queries in search engines using the command "site" with web address of the repository (e.g.: site: repository.domain.edu.pe).

To determine normality assumption in indicators (1), (2) and (3), we use Shapiro-Wilk test for samples < 50 cases. Normality test showed that values of analysis indicators were not distributed in a normal way, so we use Mann-Whitney's U Test, non-parametric equivalent of T-Test for the difference of two means in independent samples with assumptions of non-normality in populations.

Finally, to analyze the association between study variables, we use Spearman rho correlation coefficient after identifying non-normal distribution of ratios.

Results

Descriptive Analysis

The measure of central tendency of metadata quality is higher in national (media = 0.686) than in private (media = 0.608) universities, However, the standard deviation shows us that they are not statistically different.

Metadata Quality
Metadata quality ratio (MQ ratio) shows a median of 0.67 [RIC: 0.552 - 0.891] for national universities and 0.65 [RIC: 0.407 - 0.838] for private universities with a statistically non-significant difference (Mann-Whitney's U Test: p-value=0.542).

Web Visibility

Indexing in Google Scholar (IGS ratio) shows a median of 0.32 [RIC: 0.241 - 0.596] for national universities and 0.62 [RIC: 0.464 - 0.749] for private universities with a statistically non-significant difference Mann (Whitney U Test: p-value=0.054).

Coverage of Document Type, Thesis

Thesis coverage ratio shows a median of 0.67 [RIC: 0.494-0.891] for national universities and 0.42 [RIC: 0.115-0.679] for private universities; However, for this indicator, there is a statistically significant difference (Mann Whitney U Test: p-value=0.021).

Correlation Analysis

The correspondence analysis based on Spearman's rho between MQ and IGS ratio, and three indicators of coverage per document type (thesis, article, book, conference) for all sample repositories shows a moderate correlation (ρ = 0.594; p <0.01) between metadata quality and a document type indicator: thesis. In addition, Table 3 shows that there is a low correlation (ρ = 0.157) between the indexing of documents in Google Scholar and the proportion of documents harvested in Alicia, with respect to the number of documents in the source repository.

Discussion and Conclusions

Private universities show the highest number of items indexing in Google Scholar (60%) compared to public universities (43%), which evidences that there is more work at content visibility level in the private sector. However, the number of items deposited in public university repositories oversteps the number of items housed in a private university repository. In general, it is known that Google Scholar does not index all resources, for example, Orduña-Malea et al (2014) found that of 127 repositories, only 34% is indexed in google scholar and in the case of PDF files, the number is smaller, reaching only 2.5%. Also, Orduña-Malea and Delgado López-Cózar (2015) mention that 137 repositories (132 of them were institutional) have low indexing ratios and explain that these low ratios are due to the use of schemas of description that are not suitable for Google Scholar and to the use of
unsuitable web navigability.

In both cases, around half the repositories analyzed showed low indexing ratios in GS, which could indicate two points: a) lack of awareness or underestimation of web visibility that can be obtained by the institutional repository from Google Scholar (GS), or b) restriction of access (restricted or embargoed) to thesis, especially in private universities because these document types belong mostly to business sciences.

On the other hand, Alicia does not allow indexing its harvested records in Google Scholar, which could correspond to a limited capacity of its computer components since constant queries of Google bots and crawlers would generate an increase of logs that supposes a greater capacity of requirements in the server.

The proportion of items integrated into Alicia with respect to all repository items in public (69%) and private universities (61%) indicates that about half the items deposited in repositories do not meet CONCYTEC metadata quality requirements or they are document types non-contemplated for harvesting in Alicia.

Most coverage of documents from public and private universities integrated in Alicia are theses, followed by other types of documents such as articles, books and conferences. That could indicate that many Peruvian universities implement their institutional repositories to meet the requirements of national organisms like SUNEDU or CONCYTEC, but not to manage and disseminate research, and because it is an opportunity for institutional collaboration, search funds and institutional prestige improvement.

Declaration of possible conflicts of interests: DQR at the time of the research was working at CONCYTEC.

Supplementary material:

Additional information of institutional repositories analyzed is available in https://doi.org/10.6084/m9.figshare.5413231.v1

References


Bueno-De-La-Fuente, Gema, Hernández-Pérez, Tony, Rodríguez-Mateos, David, Méndez-


Roy, Bijan Kumar, Subal Chandra Biswas, and Parthasarathi Mukhopadhyay. 2016. “Global
Repository Movement in the Domain of Library and Information Science Discipline.”

Sandy, Heather Moulaison, and Felicity Dykas. 2016. “High-Quality Metadata and
Repository Staffing: Perceptions of United States-Based OpenDOAR Participants”.
Cataloging and Classification Quarterly 54 (2): 101–16.

Serrano Vicente, Rocío, Remedios Melero Melero, and Ernest Abadal. 2014. “Indicators for
the Evaluation of Open Access Institutional Repositories”. Anales de Documentacion
17 (2). doi:10.6018/analesdoc.17.2.190821.

Repositories” http://repositories.webometrics.info.

http://www.opendoar.org/.

Vílchez-Román, Carlos, and Diana Nakamura Shimabukuro. 2008. “Usability of a Full Text
Information Retrieval System: The Case of Cybertesis Peru.” ACIMED 17 (3).

Walsh, Maureen P. 2011. “Repurposing MARC Metadata for an Institutional Repository:
Working with Special Collections and University Press Monographs”. Library
Resources & Technical Services, 55(1), 33-44. doi: 10.5860/lrts.55n1.33.

Yang, Le. 2016. “Making Search Engines Notice: An Exploratory Study on Discoverability
doi: 10.1080/19322909.2016.1172539