Metadata Quality of the National Digital Repository of Science, Technology, and Innovation of Peru: A Quantitative Evaluation

Miguel Valles(✉), Richard Injante, Victor Vallejos, Juan Velasco, and Lloy Pinedo

Universidad Nacional de San Martín, Tarapoto, Peru
mavalles@unsm.edu.pe

Abstract. The National Digital Repository of Science, Technology, and Innovation (ALICIA) of Peru is in charge of harvesting the scientific production of the universities from their institutional repositories. The aim is to determine and evaluate the availability of resources in the repositories, and if the quality of the metadata harvested by ALICIA has any relation to the metadata of the institutional repositories. To this end, a non-experimental, descriptive, comparative study was carried out after recovering the data from ALICIA, using its API rest and the data from the institutional repositories, and using organic techniques for validating broken links and web scraping, which are then stored in a database on which queries were made using non-SQL statements. There is 97.7% and 95% availability of resources in public and private institutional repositories respectively, and on average, the quality of the metadata describing the resource is between 54% and 57% . It is notable that despite all the documented interventions and the results found, there are still quality problems in the metadata registering process considering that universities are responsible for its publication.

Keywords: Data quality · Curation · Dublin core · Institutional repositories · Metadata

1 Introduction

The universities have an important mission since they are the main sources of knowledge generation and contribution to science, which among other things allows them to solve society's problems, from where the complaints about their insufficient and inefficient productivity are born and who gives them the benefit of their elitist academic and scientific position, as shown in [1].

However, even though most of the research funding is the result of government contributions [2], some of the results are available through restricted access with payments and/or editorial subscriptions to licenses [3]. This require investing time with uncertainty

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 96–105, 2021.
https://doi.org/10.1007/978-3-030-77417-2_7

in the achievement of the required content, being a nefarious factor in situations such as the current ones in which every minute counts.

In opposition, the open access movement is the immediate access, without payment, registration, or subscription to academic and scientific material, it has a wide community in the world that involves teachers, undergraduate, and graduate students [4], who, with their ingenuity, solve problems and share the results in a way that increases the scientific heritage and enriches knowledge [5]. Its aim is interoperability, compatibility among archives, long-term preservation, and universal access to information. In addition to promoting free access, it encourages its availability, distribution, and reproduction [6]. It emerged in response to the restricted access to knowledge in academic and scientific journals imposed by commercial publishers (Gideon, 2008), as cited by [7].

This same open access movement has caused universities to face today an enormous challenge related to the need and obligation to generate knowledge that acquires visibility, achieves impact factors that demonstrate its importance, influence, and contribution to the scientific community that uses the published results [2]. According to [8]:

> "A growing number of universities around the world have begun to establish open access policies regarding the academic results of their researchers, requiring them to publish mainly in open access journals and/or to archive their pre-prints or post-prints in institutional archives".

In this sense, in Peru, with the approval of the University Law in 2014 [9], the National Superintendence of University Education (SUNEDU) has the function of supervising the fulfillment of the basic conditions of quality. The state as a strategy to standardize and improve the visibility of the scientific production of the institutional repositories approved according to [10] creates the National Science Repository, Open Access Technology and Innovation (ALICIA). This strategy has allowed the implementation of technological infrastructure in universities to make available the results of scientific production that meet the needs of the community [11]. Without a doubt, we can say that institutional repositories are the cornerstone for open access [2]. By facilitating the systematization and access to publications, they contribute to the concept of open science [12, 13].

In South America there are similar strategies, adapted to each context, this promotes the exchange of scientific production with the implementation of infrastructures for harvesting the resources available in the repositories of the universities in each country [14]. Even within [15] the national nodes of the countries have been progressively integrating the region into a platform that has international standards of interoperability. This provides the possibility of generating a centralized scheme of a federated search for scientific results that facilitates the citation process and reference of resources in an open-access scheme.

According to [16, 17] and [18], ALICIA collects, integrate, and storage through a constant process based on the OAI-PMH interoperability protocol [19], known as harvesting, then presents the data and metadata from all the institutional repositories that are within the scope of Law No. 30035 [10]. By law, universities that receive state funding have to guarantee the availability of their repositories and their resources must remain efficiently usable, updated, and collected by "ALICIA" [20]. According to [18] all the repositories use DSPACE as an institutional repository solution.

The current ALICIA guidelines are based on the "OpenAIRE Guidelines for Literature Repository Managers v. 3.0", the Dublin Core metadata schema, and the OAI-PMH exchange protocol. 39 metadata must be considered for information registration in the institutional repositories: 23 are mandatory, 9 recommended, and 7 optional [21].

Although indeed, the resources of the institutional repositories, known as gray literature, have not undergone a peer-review process to ensure acceptable quality levels as in the case with scientific articles published in indexed journals according to [22]. This should not be an excuse for that the registration of metadata describing the resources, which are subsequently harvested by ALICIA, not to meet the quality criteria required by ALICIA guidelines [23].

We must understand that ALICIA needs to review the harvested content and determine whether the repositories that are part of the system [24] comply with the requirements of the directives [21]. Strategies are needed to improve the quality of the process of registering metadata [25] avoiding the same deficiencies that affect the process of bibliographic review of research conducted in universities, as well as its visibility and impact factor.

Not only that, ALICIA must look for mechanisms of identification, healing at the source, and correction at the destination, in case the results of their searches identify resources that at the time of access are fallen links, due to the dynamics of institutional repositories and the deficient skills of its staff to ensure a proper healing and updating process.

There are many research works worldwide that verify the quality of repository metadata, such as the [23], which evaluates the problem from those responsible for the registry, or the [26] which correlates the quality of metadata with its academic visibility in Google Scholar. However, there are few studies to determine if the harvesting process is correct. If there is a need for updates on the harvested metadata sources. Or any one that compares the quality of the data harvested by the harvesting infrastructures existing per country [14] against the existing metadata in the repositories.

Ensuring metadata quality that describe the resources available in the repositories, has much to do with the acceptance reflected in citations and impact factor of the research available in them and which are documented by [27], In addition to the qualitative criteria of the researchers who review these resources, as [28] claim, poor quality of metadata recording can have a negative impact not only on the way scientists retrieve, share and use research datasets, but also on the way that manage and audit repositories.

Therefore, we wonder whether the quality of the metadata harvested by ALICIA has any relation with the metadata of the institutional repositories. In this research, we seek to identify if the repository's administrator record the information according to established directives based on an indicator that we call MQ similar to that of [26]; How good ALICIA manages to harvest these data according to the correct configuration of protocols in the repositories through comparison. In what extent the repositories ensure the availability of the resources that ALICIA claims are available. Finally, we evaluate the need for a process of curing the ALICIA database by applying a comparative analysis for both.

2 Materials and Methods

2.1 Analysis Unit

This is a non-experimental, comparative descriptive cross-sectional research. The study universe is the metadata available in ALICIA and the resources available in the repositories of the 51 public universities and 92 private universities in Peru.

Initially, we have to specify that we carried out the whole process described in materials and methods during May 18–22, 2020, on a virtual private server on the amazon web services.

2.2 Data Recovery

We extracted the data harvested by ALICIA, consuming its API rest located and documented in https://alicia.concytec.gob.pe/vufind/api/v1, through a program built in python based on the libraries "requests", "JSON" and "BeautifulSoup" for text analysis.

We compared this metadata with data extracted through a web scraping process from the institutional repositories of all the universities that are part of ALICIA. This process is necessary because ALICIA does not harvest all the metadata available through the OAI-PMH protocol.

In detail, we obtain the metadata of each of the records available in the ALICIA database through a loop recovery process. For each record the metadata dc.identifier. Uri (URL of the resource) is identified, which is used as a parameter to perform an HTTP query in which organic broken-link validation techniques are applied [29, 30], based on HTTP responses to validate the availability of the resource in the institutional repository with a valid response code and not "not found".

If the resource is available in the repository, we apply web-scrapping techniques on grey literature documented by [31], to structure the data of the obtained response that we insert in a non-SQL database on which we perform queries on 20 Dublin Core elements, specified according to [21] to generate the presented results.

For all the algorithms built, we have worked with python and the request, JSON, and BeautifulSoup libraries. For the processing and generation of the descriptive statistics, we used No SQL statements (Fig. 1).

The algorithm used for data recovery is:

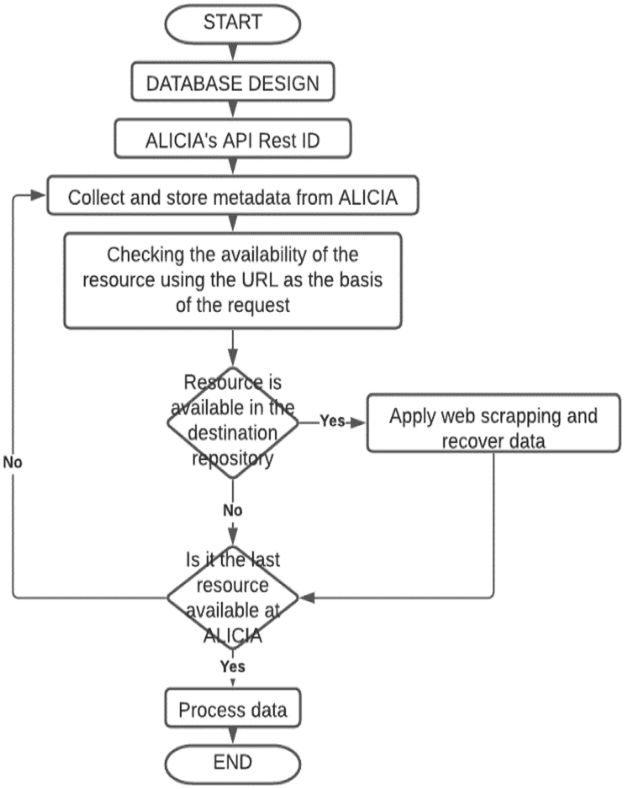

Fig. 1. Data recovery algorithm

2.3 Data Analysis and Processing

We based the evaluation criteria on the following indicators: Resource availability, Metadata quality.

Where

$$Availability = \frac{\sum_1^n resources\ that\ respond\ to\ the\ HTTP\ request}{\sum_1^m resources\ harvested\ by\ Alicia} * 100 \quad (1)$$

$$Metadata\ quality = \frac{\sum_1^n metadata\ harvested\ by\ Alicia}{\sum_1^m metadata\ specified\ in\ the\ standard} * 100 \quad (2)$$

3 Results

Thus, according to formula (1), In summary, we present the availability of resources in the repositories in Tables 1 and 2:

Table 1. Amount of available resources, as harvested by ALICIA and according to the web scraping we did from the repositories

Type of management	Harvested by alicia	Web scraping from the repositories	Availability
Public	111964	109393	97.7%
Private	156859	148949	95.0%

Source: own elaboration

At the date of data extraction, only 125 of 143 universities have repositories harvested in ALICIA; however, 15 of these have values in the URL metadata that do not respond to the HTTP checks performed by the algorithm. In addition, of these 87 have approved licensing, 38 have denied licensing.

According to information from SUNEDU, there are 51 state universities; however, only 39 of them have and maintain their institutional repository accessible.

Table 2. Available resources by type of university and type of resources, as harvested by ALICIA and according to the web scraping we did from the repositories

Management	Thesis type	Harvested by Alicia		Web scraping from the Repositories		Availability	
		Lic	No Lic	Lic	No Lic	Lic	No Lic
Public	Undergraduate	83.804	2.771	83.281	2.734	99.4%	98.7%
	Master's degree	16.491	4.077	16.464	2.126	99.8%	52.1%
	Ph.D.	4.616	205	4.606	182	99.8%	88.8%
Private	Undergraduate	92.563	20.300	88.732	19.043	95.9%	93.8%
	Master's degree	38.356	2.160	35.834	2.050	93.4%	94.9%
	Ph.D.	3.263	217	3.105	185	95.2%	85.3%

Source: own elaboration

Table 2 shows us an interesting characteristic, which according to our analysis, shows that universities with a denied license have problems guaranteeing the availability of the resources they have in their institutional repositories.

An important point to note is that the availability of resources harvested by ALICIA in the repositories of origin in public universities is 97.7%, in contrast to 95% in private universities. However [26], they find that 43% of public universities' resources and 60% of private universities' resources are correctly indexed in Google Scholar. The explanation can be found when we make a more detailed analysis of the content of the metadata since, apparently, according to [26] itself, the staff of private universities has more training in registering their resources in their repositories.

Continuing with the formula (2), following tables presents the quality of the metadata that describe the resources available in the repositories.

Table 3. Quantity of metadata harvested by type of university and type of resource.

Management	Thesis type	Amount of metadata per resource					
		Licensed			Licensed Denied		
		Max	Average	Min	Max	Average	Min
Public	Undergraduate	30	21.5	8	22	21	16
	Master's degree	29	22.5	14	23	21	20
	Ph.D.	29	21.8	14	22	21	20
Private	Undergraduate	34	22.1	12	26	22	13
	Master's degree	32	22.5	13	26	22	16
	Ph.D.	29	22.6	13	25	22	19

Source: own elaboration

An important fact from Table 3 is that, on average, public universities register 21.5 and private universities 22.5 metadata for undergraduate theses. In general, for both masters and doctoral theses, public universities register an average of 1 metadata less than private universities (Table 4).

Table 4. Quality of harvested metadata by type of university and type of resource.

Management	Thesis type	Amount of metadata per resource					
		Licensed			License Denied		
		Max	Average	Min	Max	Average	Min
Public	Undergraduate	77%	55%	21%	56%	54%	41%
	Master's degree	74%	58%	36%	59%	54%	51%
	Ph.D.	74%	56%	36%	56%	54%	51%
Private	Undergraduate	87%	57%	31%	67%	56%	33%
	Master's degree	82%	58%	33%	67%	58%	41%
	Ph.D.	74%	58%	33%	64%	56%	49%

Source: own elaboration

As for Table 5 (which is the summary of Table 3), on average, the public universities register 21.65 metadata and the private ones 22.29 metadata, that is, 0.64 metadata less. However, this analysis is only quantitative, and an analysis of the content of the metadata is necessary to determine the qualitative aspects of the data recording process to ensure that we are following the indications of [32] and the study of [25].

Table 5. Amount of metadata harvested by type of university.

Management	Amount of metadata per resource					
	Licensed			Not Licensed		
	Maximum	Average	Minimum	Maximum	Average	Minimum
Public	30	21,65	8	22	21	18
Private	34	22,29	12	26	22	14

Source: own elaboration

Table 6. Quality of harvested metadata by type of university.

Management	Metadata quality by resource					
	Licensed			Not Licensed		
	Maximum	Average	Minimum	Maximum	Average	Minimum
Public	77%	56%	21%	56%	54%	46%
Private	87%	57%	31%	67%	56%	36%

Source: own elaboration

Table 6 shows a very important indicator regarding the work and importance of those responsible for the repositories because according to [25], in general, they need training processes and strengthening of skills to ensure that the registration of metadata is done correctly, since as universities are responsible for the quality of the metadata recorded as mentioned [32].

4 Conclusions

After evaluating the resources harvested by ALICIA and their availability in university repositories, we concluded that public universities have 3% of unavailable resources compared to 5% of resources unavailable by private universities. The licensing process initiated 5 years ago with the enactment of the new university law could be the reason, in the process of which 25% of private universities have not achieved their licensing and have ceased to function.

Although many metadata's quality problems have been documented in the literature over the last few years (e.g., [33, 34] and even Concytec did its analysis with [25]), many of these problems are still present in the case of Peru's National Digital Repository of Science, Technology, and Innovation (ALICIA).

Finally, ALICIA must establish mechanisms for maintain available repositories of universities with denied licenses so that a significant number of researches are not left offline.

References

1. Concepción, D., Gonzáles, E., Miño, J.: Una visión actual de la ciencia como fuerza productiva directa. Rev. Univ. y Soc. **10**, 54–59 (2018)
2. Zacca-Gonzales, G.: Los repositorios en función de la ciencia abierta. Rev. Cuba. Inf. en Ciencias la Salud **30**(4), 1–3 (2019)
3. Banzato, G., Rozemblum, C.: Modelo sustentable de gestión editorial en Acceso Abierto en instituciones académicas. Principios y procedimientos. Palabra Clave (La Plata) **8**, e069 (2019). doi:https://doi.org/10.24215/18539912e069.
4. Silva-Rodríguez, A.: La sostenibilidad de la divulgación de la ciencia mediante modelos de negocios de acceso abierto. Rev. Digit. Int. Psicol. y Cienc. Soc. **2**, 21–39 (2016). https://doi.org/10.22402/j.rdipycs.unam.2.1.2016.73.21-39
5. Jacobs, N., Nixon, W.J.: Universities, Jisc and the journey to open. In: Technology, Change and the Academic Library, pp. 171–181. Elsevier (2021). https://doi.org/10.1016/b978-0-12-822807-4.00017-8
6. Cano, A., De Dios, R., García, O., Cuesta, F.: Los repositorios institucionales: situación actual a nivel internacional, latinoamericano y en Cuba. Rev. Cuba. Inf. en Ciencias la Salud **26**, 0–0 (2015)
7. Abrizah, A., Noorhidawati, A., Kiran, K.: Global visibility of Asian universities' Open Access institutional repositories. Malaysian J. Libr. Inf. Sci. **15**, 53–73 (2010)
8. Granholm, K.: Open access och spridning En kvantitativ analys av hur open access-publicerade artiklar citeras och sprids på webben. Uppsala University (2013)
9. Ley 30220: Ley Universitaria (2014). https://www.sunedu.gob.pe/wp-content/uploads/2017/04/Ley-universitaria-30220.pdf
10. Ley N° 30035: Ley que regula el repositorio nacioanl digital de ciencia, tecnología e innovación de acceso abierto 2 (2013)
11. Sandí, J., Cruz, M.: Repositorios institucionales digitales: Análisis comparativo entre Sedici (Argentina) y Kérwá (Costa Rica). e-Ciencias la Inf. **7**, 1 (2016). https://doi.org/10.15517/eci.v7i1.25264
12. Concytec: Autoridades académicas de Macrorregión Norte debaten sobre ciencia abierta y repositorios digitales (2019). https://portal.concytec.gob.pe/index.php/noticias/1791-ciencia-abierta-y-repositorios-digitales-debatidos-en-macrorregion-norte
13. Goben, A., Sandusky, R.J.: Open data repositories: current risks and opportunities. Coll. Res. Libr. News **81**, 62 (2020). https://doi.org/10.5860/crln.81.1.62
14. La Referencia. Nodos nacionales: el engranaje | LA Referencia. La Referencia (2020). https://www.lareferencia.info/legacy/nodos-nacionales-el-engranaje.html
15. La Referencia. ¿Qué es el buscador regional? | LA Referencia. La Referencia (2020). https://www.lareferencia.info/legacy/buscador-regional.html
16. Decreto Supremo N° 006-2015-PCM. Reglamento de la Ley N° 30035, Ley que regula el Repositorio Nacional de Ciencia, Tecnología e Innovación de Acceso Abierto 4 (2015)
17. Atamari Anahui, N., Díaz Vélez, C.: Repositorio Nacional Digital de Acceso Libre (ALICIA): Oportunidad para el acceso a la información científica en el Perú. An. la Fac. Med. **76**, 81 (2015). https://doi.org/10.15381/anales.v76i1.11081
18. Rivero, A.: Avances del repositorio nacional digital ALICIA: recolección del repositorio nacional digital ALICIA a las instituciones. Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica - Concytec (2018). https://repositorio.concytec.gob.pe/handle/20.500.12390/85
19. Haslhofer, B., Schandl, B.: Interweaving OAI-PMH data sources with the Linked Data cloud. Int. J. Metadata, Semant. Ontol. **5**, 17–31 (2010). https://doi.org/10.1504/IJMSO.2010.032648

20. Concytec. Reglamento RENATI (2016). https://busquedas.elperuano.pe/normaslegales/aprueban-reglamento-del-registro-nacional-de-trabajos-de-inv-resolucion-no-033-2016-sun educd-1425605-1/
21. Concytec: Directrices para el procesamiento de información en los repositorios institucionales (2020). https://portal.concytec.gob.pe/images/documentos/alicia/directrices_repositorio.pdf
22. Valderrama, J.: Literatura Gris. . Form. Univ. **4**, 1–2 (2011). https://doi.org/10.4067/S0718-50062011000600001
23. Stovold, E.: Metadata quality in institutional repositories may be improved by addressing staffing issues. Evid. Based Libr. Inf. Pract. **11**, 93–95 (2016). https://doi.org/10.18438/b81s7n
24. Concytec. Instituciones Integrantes - ALICIA (2019). https://alicia.concytec.gob.pe/vufind/
25. Francisco Talavera Chocano -Consultor, M.: Análisis de la calidad de metadatos en el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto – ALICIA. Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica - Concytec (2019). https://repositorio.concytec.gob.pe/handle/20.500.12390/399
26. Alhuay-Quispe, J., Quispe-Riveros, D., Bautista-Ynofuente, L., Pacheco-Mendoza, J.: Metadata quality and academic visibility associated with document type coverage in institutional repositories of peruvian universities. J. Web Librariansh. **11**, 241–254 (2017). https://doi.org/10.1080/19322909.2017.1382427
27. Repiso, R., Moreno-Delgado, A., Aguaded, I.: Factors affecting the frequency of citation of an article. Iberoam. J. Sci. Meas. Commun. **1**, (2020). https://doi.org/10.47909/ijsmc.08
28. Balatsoukas, P., Rousidis, D., Garoufallou, E.: A method for examining metadata quality in open research datasets using the OAI-PMH and SQL queries: the case of the Dublin Core 'Subject' element and suggestions for user-centred metadata annotation design. Int. J. Metadata, Semant. Ontol. **13**, 1–8 (2018). https://doi.org/10.1504/IJMSO.2018.096444
29. Hayat, S., Li, Y., Riaz, M.: Automatic recovery of broken links using information retrieval techniques. In: ACM International Conference Proceeding Series, pp. 32–36. Association for Computing Machinery (2018). https://doi.org/10.1145/3278293.3278296
30. Bashir, S.: Broken link repairing system for constructing contextual information portals. J. King Saud Univ. Comput. Inf. Sci. **31**, 147–160 (2019). https://doi.org/10.1016/j.jksuci.2017.12.013
31. Haddaway, N.: The use of web-scraping software in searching for grey literature. Grey J. **11**, 186–190 (2015)
32. Hrynaszkiewicz, I.: Publishers' responsibilities in promoting data quality and reproducibility. In: Bespalov, Anton, Michel, Martin C., Steckler, Thomas (eds.) Good Research Practice in Non-Clinical Pharmacology and Biomedicine. HEP, vol. 257, pp. 319–348. Springer, Cham (2019). https://doi.org/10.1007/164_2019_290
33. Swanepoel, M.: Digital repositories: all hype and no substance? New Rev. Inf. Netw. **11**, 13–25 (2005). https://doi.org/10.1080/13614570500268290
34. Palavitsinis, N., Manouselis, N., Sanchez-Alonso, S.: Metadata quality in learning object repositories: a case study. Electron. Libr. **32**, 62–82 (2014). https://doi.org/10.1108/EL-12-2011-0175