Aggregation of institutional repositories for the analysis of the scientific performance of Peruvian universities

Miguel Valles-Coral¹, *, Richard Injante¹, Edwin Hernández-Torres¹, Lloy Pinedo¹, Jorge Raúl Navarro-Cabrera¹, Luis Salazar-Ramírez¹, Ángel Cárdenas-García¹, Eddy Huancaruna¹

¹ Faculty of Systems Engineering and Informatics, National University of San Martín, Tarapoto, Peru.
* Corresponding author.
Email: mavalles@unsm.edu.pe. ORCID: https://orcid.org/0000-0002-8806-2892.

ABSTRACT
Objective. The scientific performance of Peruvian universities was examined and characterized by aggregating information from their institutional repositories.

Design/Methodology/Approach. A protocol was followed that included the following stages: (1) analysis and optimization of the DSpace database model; (2) data recovery and validation with the OAI-PMH Protocol; (3) obtaining an index of resources available in the repositories; (4) metadata extraction and processing; and (5) extraction and inclusion of keywords in the database.

Results/Discussion. A diverse use of keywords was identified. 69.1% of them appeared only once. A significant gap was evident in producing theses between public and private universities.

Conclusions. The discrepancy between the research areas of Peruvian universities and regional needs highlights the importance of greater alignment between academic production and the demands of society in each region; this underlines the need to adapt research programs to address specific problems and contribute more effectively to regional development.

Keywords: research evaluation; scientific production; institutional repositories; Peruvian universities.

1. INTRODUCTION

Institutional repositories (IR) are platforms that facilitate access to academic products. They provide vital support for university students’ and teachers’ learning and research (Bamigbola and Adetimirin, 2020). Although a central pillar of IR is to promote open access, there are cases where access restrictions are implemented for specific resources, either making them exclusively accessible from institutional campuses or only to particular users. This occurs to safeguard sensitive or protected information (Gunderman, Scherer, and Behrman, 2020). IR play a crucial role in metadata collection and interoperability (Chisita and Chhiparausha, 2021) and in its monitoring. Accessing their statistics can be challenging, as they

Received: 21-03-2023. Accepted: 12-10-2023. Published: 23-10-2023

Editor: Adilson Luiz Pinto


Copyright: © 2023 The author(s). This is an open access article distributed under the terms of the CC BY-NC 4.0 license which permits copying and redistributing the material in any medium or format, adapting, transforming, and building upon the material as long as the license terms are followed.
Valles-Coral, et al.

IR management is a constant process, a large volume of data requires continuous supervision and correct decision-making about managing data, elements, metadata field, and treatment (Shelley, 2020). Some research has emphasized that IR administrators do not have extensive knowledge of the processes and maintenance of repository functions (Fujita et al., 2023; Masinde & Sanya, 2022).

The technological infrastructure of IR can be expanded and optimized to manage metadata and effectively promote the research contained through, for example, improvements in search engine optimization and advanced web analytics. In this way, IRs would house and enhance the global presence and recognition of academic research and products (Gunderman et al., 2020). The lack of tools to visualize large data sets in Peruvian universities negatively impacts the interaction of researchers with IR (Santos-Hermosa, 2023). Proposals such as those of Orrego Granados et al. (2022) seek to solve this problem. Although the web platforms of these repositories offer sections of statistical information, their value is limited and does not contribute to meaningful analysis. Furthermore, due to the insufficient adequacy of data and metadata, it is impossible to evaluate the performance of universities to determine whether their research responds to the demands of society in their respective areas of influence. This lack also limits understanding and decision-making.

The web platform of each university presents sections of statistical information of little value, which does not benefit the practical analysis of the representatives. The lack of adequate processing of data and metadata prevents the evaluation of the performance of institutions to determine whether their research aligns with the needs of society in their local context. This limitation restricts the ability to understand and make informed decisions due to the quality of the design of the repositories’ web portal (Subiyakto et al., 2021).

Universities have integrated data analysis into their processes, playing a fundamental role in strategic decision-making (García Estrella et al., 2021; Schoen et al., Kisa, 2019). This analysis has as its primary focus maximizing the benefits derived from the information generated, providing a deeper understanding of patterns and trends. However, managing large volumes of information also entails the complexity of occasionally lack the functionalities necessary for intuitive and effective navigation (Vardell, 2020).

Most universities have implemented an IR. Peruvian universities are no exception since, as of June 2020, more than 80% of the 92 universities licensed by the National Superintendency of Higher University Education (SUNEDU, in Spanish) have an institutional repository (Blanco-Olea, 2021). These licensed universities have implemented an IR to manage their academic and scientific production. These IRs store both the content and its metadata, playing a crucial role in organizing and preserving these resources (White and Radio, 2020). Optimal information management is achieved through IR, and more effective data recovery is completed. Researchers and students can quickly access relevant materials (Velasco-Mieses, 2022).

These repositories provide the opportunity to establish interoperability features. This encourages collaboration and the exchange of information between different academic and scientific entities. However, to date, IR in Peruvian universities have not incorporated functionalities that enable graphical visualization of metadata aggregation. This limitation prevents a holistic and visual understanding of the institution’s distribution of stored resources and focus areas. Additionally, these repositories lack performance indicators of the functional units and dependencies belonging to the institution. The absence of this function prevents an evaluation regarding the use of resources and their impact on academic and scientific production in different areas. The incorporation of these characteristics would significantly enhance the usefulness of IR. In turn, this would provide a more complete and strategic vision in decision-making (Tsunoda et al., 2016).

A repository administrator performs a vital task in identifying and managing data in academic products as researchers, editors, and other interested parties seek to publish research to receive recognition. For this reason, it is crucial that the repository administrator can identify and monitor the data of researchers and their academic activities (Mering, 2019). If dissatisfaction with IR is generated, researchers could publish their research on other platforms or portals (Price and Murtagh 2020). Although IR management is a constant process, a large
a sophisticated analysis process (Vázquez-Ingelmo, García-Peñalvo, and Therón, 2021). According to Susnjak, Ramaswami, and Mathrani (2022), when supported by layers of information provided by repositories, decision-making gives users a broad and enriching perspective when defining strategies. This integration of analytics and repositories makes it easier to identify meaningful findings and empowers universities to adopt more effective approaches.

Calle Paz (2021) has implemented a data analysis panel, which constitutes a central tool for decision-making. Our board has been designed to effectively visualize the data hosted in the IR of Peruvian universities, and dashboard-type visualization approaches are used (Subiyakto et al., 2021). This approach empowers repository administrators and other users with the ability to design strategies (Miguel, González, and Ortiz-Jaureguizar, 2018). Through the aggregation and meticulous processing of metadata extracted from IR, valuable information has been generated about the scientific performance of Peruvian universities (Bandeira Andriola and Castro Araújo, 2018).

This initiative is fundamental to more efficient management and data-based decision-making (Valles-Coral, 2023).

The general objective of this work is to analyze the scientific performance of Peruvian universities intending to (a) generate relevant information by extracting metadata using interoperability protocols, (b) aggregate and process said data in a management information system using a dashboard to visualize scientific production, and (c) describe the scientific performance of universities.

2. METHODOLOGY

We conducted research in the information systems laboratories of the Faculty of Systems Engineering of the National University of San Martín, Peru, from June 2022 to May 2023. To achieve this objective, we follow the methodologies proposed by Balatsoukas et al. (2018), Haddaway (2015), Meschenmoser et al. (2016), Nieto et al. (2020), and Valles et al. (2020). From these methodologies, we formulate the protocol presented in Figure 1.

2.1. Analysis and renormalization of the DSpace Database Model. Through a “pivot” transformation, we obtained an entity-relationship model in MySQL that has an optimized structure with three entities: "Universities", "Published Items" and "Keywords". Then, we register the licensed universities of Peru. These metadata were related to the variables taken into account to analyze the performance of the universities.
2.2. Data recovery and validation through the OAI-PMH Protocol. In an iterative process, we accessed the IR data using the OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) protocol. We previously verified its availability online.

2.3. Obtaining the index of resources available in the repositories. The index of resources available in each university repository was obtained using the OAI-PMH protocol. This allowed us to extract and analyze metadata.

2.4. Extraction and treatment of metadata in XML format. Again, iteratively, within the index of available resources, we proceeded to extract the metadata associated with the stored resources. We subjected them to a validation process to obtain the metadata: title (resource available in the repository), authors, abstracts, and dates. Finally, we proceeded to insert it into the database.

2.5. Extraction of keywords and their insertion into the database. Finally, we iterate on the metadata obtained to extract keywords associated with each resource we enter into the database, establishing links with their respective “items”.

Using the data collected and available in the database created in step 2.1, we evaluated universities’ performance. This process involved applying SQL scripts for aggregation. We use Group By commands to analyze the number of theses and keywords by university and region (geographic and political). Vue 3.0 with InertiaJS was used to develop the interfaces and implement the graphics. On the backend, PHP 8.0 was used, both for data query and subsequent processing. This classification allowed us to identify the variability and concentration of keywords in the repositories. This provided a clearer understanding of the research topics addressed by Peruvian universities.

3. RESULTS AND DISCUSSION

3.1. Implications of the recovery of IR metadata from Peruvian universities through interoperability protocols

Figure 2 provides information on the variability and concentration of keywords in these repositories. The majority (69.1%) appear only once, indicating a marked diversity in their choice. Although a minority are repeated more frequently, the majority of them remain unique. The skewness (134.277) suggests the distribution is skewed to the right. The extremely high kurtosis (29,284,193) means that the frequency distribution has heavy tails and a significant concentration of keywords in the low values. This could indicate the need to review and improve keyword selection practices.
These results indicate considerable variability and lack of uniformity in keyword use. This may have implications for information accessibility and retrieval for users, who could benefit from greater standardization and review of indexing practices. This distribution reflects a complex interaction of actors and factors in the indexing process. Luiz Pinto et al. (2022) highlight the importance of shared responsibility and the vital role of training, guidelines, feedback, and continuous improvement in optimizing metadata in repositories. Collaboration and commitment from all stakeholders are essential to address this challenge and improve the accessibility and visibility of academic research in the Peruvian context.

3.2. Visualization and analysis of information on the scientific production of universities

Scientific production in the university environment acquires precise and revealing contours through collecting and processing crucial indicators. This section focuses on presenting the analysis results. Thus, this summary reveals its evolution, providing a quantitative vision of the trends.

The data in Figure 3 highlight a gap in thesis production between public and private universities. Although public ones represent 45.45%, their contribution is only 30.79%. On the other hand, private companies (54.55%) lead the production of theses with 69.21%. The private César Vallejo University stands out with 116,058 theses, representing 24% of the total. Likewise, we appreciate its increase since 2014, coinciding with the approval of Peruvian University Law No. 30220. This suggests that the reform could have influenced the promotion of research in universities, as Bibi et al. (2023) considered when talking about improving the higher education system. This information makes it possible to evaluate and understand the academic and scientific impact and propose reforms to university public policies as proposed by Estrada Araoz et al. (2023). These indicators reflect figures and the constant effort and commitment of universities in solving society’s problems.
By reviewing the increase in thesis production at César Vallejo University since 2014, a space is opened to examine how government policies can influence the behavior of academic institutions, which are addressed by the studies of Deroncele-Acosta et al. (2023); Gonzales et al. (2019) and especially Navarro Vilchez (2021). The identified trend suggests a stimulus for research and raises the need to investigate further the dynamics between government regulations and research practices in the university field. In addition to the number of theses produced, we must also evaluate their actual impact on society, highlighting the need to assess the quality and relevance of the research generated by these institutions (Arbieto Mamani et al., 2023). These insights provide a more complete and contextualized vision of the results, inviting a deeper reflection on the underlying dynamics in the academic production of Peruvian universities.

Beyond the numbers, possible factors contributing to this disparity are explored, such as differences in resource allocation and pedagogical approaches between the two types of institutions. This divergence could be influenced by the specialization of private universities in specific areas of research that enjoy more significant funding or social demand. Furthermore, the question arises as to whether the quality of theses, beyond quantity, varies between these two educational sectors. This is crucial to evaluating the real impact of generating knowledge and problem-solving.

### 3.3. Description of the scientific performance of universities

In Figure 4, an analysis of the geographic distribution of the research reveals exciting patterns. In the Selva region, business and administration, agronomy, and topics such as management and taxation strongly emphasize economic and agricultural issues. In the Coast region, business and administration, education, nursing, law, industrial engineering, and productivity suggest significant diversification compared to the Jungle region. In the Sierra, education and administration highlight the importance of academic training in universities. Furthermore, the presence of topics such as mining and mineral processing possibly reflects the relevance of the mining industry in this region.

This shows how these universities adapt and contribute to each region’s specific needs and demands. This has been evidenced in the studies of Deroncele-Acosta et al. (2023) on research trends in Peru and Villacorta Chambi et al. (2022) on the current situation of geoscience in said country. Are these theses addressing local and regional issues, or are there opportunities for greater alignment with societal demands in different parts of Peru? According to Belter et al. (2019), this is a crucial question as future academic policies are planned and the impact of research on society and problem-solving is assessed.

This could reflect universities’ adaptation to their regions’ needs and the specific research approaches they are developing, as mentioned (Albertus, Espinoza, and Fort, 2020). We must consider that universities can adapt to address the specific needs of their environment and respond to the demands of the labor market and the local community (Luna, Chong, and Djurica, 2023; Novella, Rosas-Shady, and Alvarado, 2023).

### 4. CONCLUSIONS

Based on the information generated from the aggregation of keywords from the repositories, it has been identified that Peruvian universities develop theses in various areas of study, given the diversity of keywords used. The keywords that stand out are business, education, nursing, law, civil engineering, economics, agronomy, and mining. This highlights a diversity of research disciplines. Likewise, thanks to visual exploration, it has been identified that there is a gap in the production of theses between public and private universities that gives us a valuable perspective to propose improvements in the management of information in repositories. Finally, the discrepancy between the research areas of Peruvian universities and regional needs highlights the importance of greater alignment between academic production and the demands of society in each region. Added to this is the need to adapt research programs to address specific problems and contribute more effectively to regional development.

### Conflict of interests

The authors declare that there are no conflicts of interest in this work.
Contribution statement


Acquisition of funds; investigation; project administration: M. Valles-Coral, R. Injante, E. Hernández-Torres.

Supervision: M. Valles-Coral, L. Salazar-Ramírez.


Writing - original draft; writing - review and editing: M. Valles-Coral, E. Huancaruna, L. Pinedo, J. R. Navarro-Cabrera.

Statement of data consent

The data generated during the study has been included in the article.

Funding

The National University of San Martín has financed the project “Integration of the UNSM Institutional Repository in the National Information Network on Science, Technology and Innovation of Peru” through Resolution No. 483-2022-UNSM/CU-R of 06/06/2022.

REFERENCES


Gonzales, B. B., Iraola-Real, I., Lluluy, D., Menezes-Claudio, B., and Roman-Gonzalez, A. (2019). Relationship Between Research Production And Funding...
ORIGINAL ARTICLE

At The Universidad de Ciencias y Humanidades. In 2019 IEEE World Conference on Engineering Education (EDUNINE) (pp. 1-5). IEEE. https://doi.org/10.1109/EDUNINE.2019.8875754


