ISSN: 3116-3475
International Journal
of Education, Research, and Innovation Perspectives
Volume 2 Issue 7 (July 2026)

HCCI eScholar: An Online Research Database Archival System with TF-IDF and Cosine Similarity Algorithms

Ronnie John Mark P. Sangil[1*], Marvin O. Mallari[1], John Elmo G. Mangune[1] and Jiane Princess S. Sigua[1]
[1]*Holy Cross Colleges, Inc. Sta. Ana, Pampanga*
[*]*ronniejohnmarksangil@gmail.com*

Date Submitted:
March 20, 2026

Date Accepted:
May 15, 2026

Date Published:
July 28, 2026

DOI:
10.5281/zenodo.21639964

ABSTRACT

The sudden increase in academic research outputs in higher education institutions presents challenges in file management, including redundancy of research topics, limited cross-department visibility, and difficulty in detecting similarities. The study developed HCCI eScholar: An Online Research Database Archival System with TFIDF and Cosine Similarity Algorithms, a centralized, searchable, and secure digital repository designed to archive undergraduate theses, capstone projects, faculty research, and institutional publications at a higher education institution in the Philippines. The system uses cross-department validation to prevent duplicate research submissions, utilizing an internal similarity-checking system that uses TF-IDF vectorization and the Cosine Similarity algorithms to check the originality of submitted abstracts without needing any third-party plagiarism detection services. To make sure that data is safe and can be stored for a long time, the system uses database normalization, encrypted PDF storage, structured directory management, and backup methods in a cPanel hosting environment. ISO/IEC 25010 Software Quality Model was used to test the system, and produced positive results. The system received excellent remarks across key software quality criteria, indicating that it satisfies essential functional and quality requirements. The results indicate that HCCI eScholar effectively enhances research management, reduces redundancy, and supports academic integrity by means of secure centralized archiving and similarity detection.

Keywords: *Cosine similarity, Institutional repository, Research archival system, TF-IDF vectorization*

INTRODUCTION

The rapid advancement of digital technology has fundamentally reshaped how academic institutions produce and manage scholarly work (Alayon, 2022). According to Aljounid (2025), there has been a surge in the volume of academic output, including undergraduate theses, capstone projects, master's and doctoral dissertations, faculty research, and institutional studies (Lynch, 2003). This increase reflects the expanding number of students and researchers and the growing demand for evidence-based instruction and academic productivity and presents a critical challenge in knowledge management. Without digital systems, a substantial portion of institutional research risks becoming fragmented, inaccessible, or lost over time (Alayon, 2022). This challenge is especially pronounced in institutions where record-keeping remains decentralized across academic departments. In these instances, departments may unknowingly accept projects with overlapping or closely related subjects due to an unavailability of recognition of previous research (Akhtar, 2023). This unintentional repetition undermines the originality of scholarly work and leads to the inefficient usage of institutional resources. The absence of a centralized system to perform similarity detection or verification hinders the detection of duplication, therefore threatening academic integrity (Suber, 2012). These flaws undermine the integrity of institutions and the precision

ISSN: 3116-3475
International Journal
of Education, Research, and Innovation Perspectives
Volume 2 Issue 7 (July 2026)

of academic records. At a higher education institution in the Philippines, the existing departmental-based system reflects the limitations of decentralized research management.

This setup prevents effective monitoring of completed studies across programs and increases the likelihood of duplicate or repetitive research topics. As a result, students and faculty may unknowingly reproduce previous work, and important academic resources may remain underutilized or unavailable for future reference. To address these concerns, this study proposes the development of HCCI eScholar: An Online Research Database Archival System with TF-IDF and Cosine Similarity Algorithms, a centralized, searchable, and secure digital repository that will archive undergraduate theses, capstone projects, faculty research, and institutional publications. In addition to the redundancy reduction and minimizing similarity, HCCI eScholar will also serve as a digital repository that protects institutional data. Centralized access to past research enhances the quality of literature reviews, facilitates academic planning, and provides evidence for accreditation.

Additionally, making a digital archive will encourage transparency, intellectual accountability, and cross-departmental research (Alayon, 2022). The successful use of institutional repositories at different Philippine universities shows that HCCI eScholar will greatly improve the quality of research and the amount of work done at a higher education institution in the Philippines. In summary, HCCI eScholar is envisioned not merely as a storage system, but as a comprehensive academic infrastructure that uses TF-IDF and Cosine Similarity Algorithms that preserves scholarly work, prevents duplication, strengthens ethical research practices, and ensures long-term accessibility of the institution's intellectual output.

METHODS

Research Design

The study used a quantitative developmental descriptive research design to collect and analyse numerical data using surveys, questionnaires, and measuring instruments. The study focused on the design, development, and evaluation of the HCCI eScholar Online Research Database Archival System. The researchers utilized the Agile approach, which enabled incremental improvements based on expert feedback and ISO 25010 quality guidelines. The researchers collected data on the accuracy, efficiency, and usability of the system, as well as the competence, satisfaction, and overall assessment of the experts. Quantitative methods, such as surveys, data collection, and analysis of text or image content, were adopted to provide more accurate and detailed data. The approach enabled the evaluation of the experts' assessment, challenges, and suggestions, which helped to guide future improvements and development of the system.

Research Locale

The system was developed within the college and based on real-world application and the way the Research and Development Unit will use it. Using the system, users will be able to upload, search, and review research papers. Also, the HCCI eScholar system was built with functionality that allows users to find similar documents based on their content using TF-IDF and cosine similarity algorithms. As part of the evaluation of the system, a sample group of selected IT professionals with knowledge of web-based development and/or software evaluation were used. Evaluation criteria included adopted ISO/IEC25010, focusing on the functional requirements of the eScholar system, usability, reliability and performance.

Sampling Technique

The researchers used a purposive sampling approach under nonprobability sampling to select respondents who were capable of assessing the HCCI eScholar system. They selected respondents based on their credentials and experience in web development. The selection of respondents was made to ensure that they could provide informed feedback. IT experts must have formal education or credentials in computing, software engineering, or other related fields. This criterion ensured that they possessed a basic understanding of software design and quality assessment. As stated by International Organization for Standardization and International Electrotechnical

ISSN: 3116-3475
International Journal
of Education, Research, and Innovation Perspectives
Volume 2 Issue 7 (July 2026)

Commission (2021), professionals must have hands-on experience in software development, system design, quality assurance, or other areas of IT under ISO/IEC 24773-3:2021. The experience was necessary for them to assess real-world software systems effectively. The experts must also be familiar with software quality standards, particularly ISO/IEC 25010. The familiarity was necessary for them to evaluate the system according to established quality attributes such as functionality, usability, reliability, and maintainability.

The data collection involves ten (10) IT experts with experience in web development or software engineering to evaluate the HCCI eScholar system. The sample size, consisting of experts from industry and academic research, follows Canlas (2021), who demonstrated that ten experts were sufficient to achieve a scale content validity index of 96.67% to 100% when validating an instrument grounded in ISO/IEC 25010. Adopting the sample size allows for precise, credible feedback while remaining manageable. The respondents' expertise in web development ensures the evaluation captures both technical functionality and practical usability, supporting the reliability of the data and aligning with best practices for expertbased software quality evaluations.

Statistical Treatment

Table 1. Descriptive Evaluations Chart per Likert Scale Point

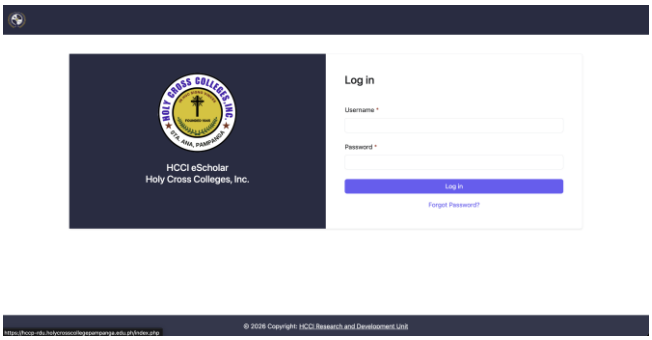
Point	Point Scale	Descriptive Interpretation
4	3.26 – 4.00	Strongly Agree
3	2.51 – 3.25	Agree
2	1.76 – 2.50	Disagree
1	1.00 – 1.75	Strongly Disagree

Table 1. shows different characteristics of the system, the researchers used descriptive statistical analysis on the data gathered. The survey used a four-point Likert scale, in which the respondents answered based on the following scale.

To interpret the gathered responses, the researchers employed several statistical techniques. The Weighted Mean ($\bar{x}$) was utilized to determine the average rating provided by respondents, serving as a measure of the overall user impression of the system. Additionally, the Standard Deviation (SD) was calculated to assess the consistency of the respondents' views and determine whether answers were clustered or varied widely.

RESULTS AND DISCUSSION

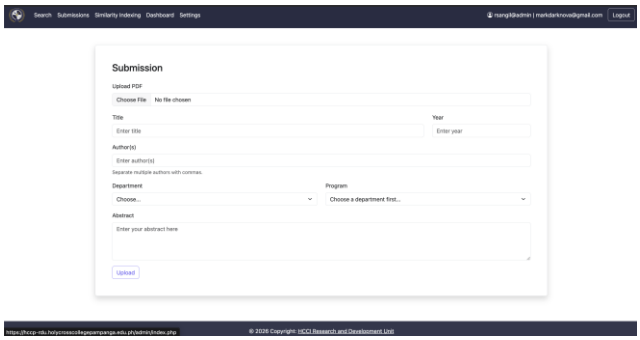
System User Interface



ISSN: 3116-3475
International Journal
of Education, Research, and Innovation Perspectives
Volume 2 Issue 7 (July 2026)

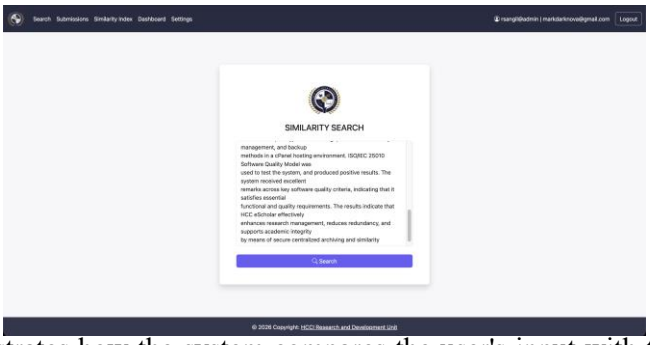
The figure displays the user authentication interface for HCCI eScholar. It requires a username and password to login.

Manuscript Submission



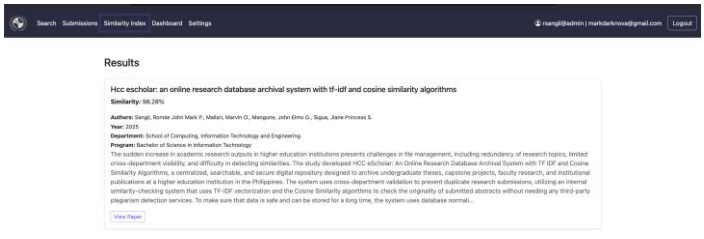
The figure illustrates the interface used for uploading academic papers. The page allows users to submit their completed research manuscripts to the HCCI eScholar repository for review and storage.

Similarity Search Page



The figure illustrates how the system compares the user's input with the existing research manuscripts. The process identifies documents that have similar content or topics to provide the most relevant results.

Similarity Results



The figure displays the list of academic papers found after a search. The page provides links to the uploaded manuscripts, allowing users to view or download specific research documents.

ISSN: 3116-3475
International Journal
of Education, Research, and Innovation Perspectives
Volume 2 Issue 7 (July 2026)

Preprocessing Algorithm

```php
function removeStopWords($text) {
    $stopWords = [
     'a','an','the','and','or','but','if','while','at','by','for','with','about','against','between','into',
        'through','during','before','after','above','below','to','from','up','down','in','out','on','off','over',
        'under','again','further','then','once','here','there','when','where','why','how','all','any','both',
        'each','few','more','most','other','some','such','no','nor','not','only','own','same','so','than','too',
        'very','can','will','just','don','should','now','what','which','who','whom','this','that','these','those',
        'is','am','are','was','were','be','been','being','have','has','had','having','do','does','did','doing',
        'using'
    ];

    $words = preg_split('/\s+/', strtolower(strip_tags($text)));
    $filtered = array_diff($words, $stopWords);
    return implode(' ', $filtered);
}
```

The figure illustrates the preprocessing step where common words, such as "the," "is," and "and," are removed from the text. The process helps the algorithm focus on the most important keywords in the research manuscript.

Cosine Algorithm

```php
function cosineSimilarity($vec1, $vec2) {
    $dot = 0;
    $normA = 0;
    $normB = 0;

    $keys = array_unique(array_merge(array_keys($vec1), array_keys($vec2)));

    foreach ($keys as $k) {
        $a = $vec1[$k] ?? 0;
        $b = $vec2[$k] ?? 0;
        $dot += $a * $b;
        $normA += $a * $a;
        $normB += $b * $b;
    }

    if ($normA == 0 || $normB == 0) return 0;
    return $dot / (sqrt($normA) * sqrt($normB));
}
```

The figure shows the internal similarity index checker was developed using cosine similarity on the existing research repository. The process involves preprocessing documents by eliminating stop words, text normalization, and optionally applying stemming. Each document, including new submissions, is converted into a vector using term frequency or TF-IDF. Cosine similarity is then computed between the submitted paper and stored records to measure textual overlap. By utilizing vectorized data stored in a MySQL database, the system efficiently identifies related papers without external similarity tools. Hence, the HCCI eScholar similarity checker is fast, secure, and fully integrated.

Instrument Evaluation Results

Table 2. ISO/IEC 25010 Evaluation Results

Criteria	Mean	Descriptive Interpretation	Standard Deviation
Functional Suitability	3.60	Strongly Agree	0.52
Performance Efficiency	3.30	Strongly Agree	0.48
Compatibility	3.60	Strongly Agree	0.52
Interaction Capability	3.60	Strongly Agree	0.52
Reliability	3.70	Strongly Agree	0.48
Security	3.40	Strongly Agree	0.51
Maintainability	3.70	Strongly Agree	0.48

ISSN: 3116-3475
International Journal
of Education, Research, and Innovation Perspectives
Volume 2 Issue 7 (July 2026)

Flexibility	3.50	Strongly Agree	0.53
Safety	3.80	Strongly Agree	0.42
OVERALL SCORE	3.58	Strongly Agree	0.50

The table 2 indicate that the system obtained an Overall Weighted Mean of 3.58, interpreted as Strongly Agree, which signifies that the respondents strongly agreed that the system met the basic quality requirements. The results also showed that the system achieved strong ratings across all criteria, with Safety obtaining the highest weighted mean score (3.80), while Performance Efficiency received the lowest weighted mean score (3.30) was due to the time taken for TF-IDF vectorization or the constraints of the cPanel hosting environment mentioned in the methodology, both still interpreted as Strongly Agree. The standard deviation values show that the evaluators' responses were generally consistent across criteria, with SD values ranging from 0.42 to 0.53.

CONCLUSION

The HCCI eScholar effectively addresses the research problems identified in the study. The system identifies redundancy across departments and ensures that all submissions undergo cross-department validation. The integration of the Cosine Similarity algorithm allows the system to assess the content of abstracts internally, eliminating the need for external plagiarism detection services. Data integrity is secured through a database normalization architecture, and encrypted PDF storage, which maintains the consistency of research files. The results show that the system effectively manages large volumes of research manuscripts using optimized database indexing and structured directory storage in a cPanel environment. Assessment results based on the ISO/IEC 25010 Software Quality Model confirm that the system meets high standards across all criteria. While Performance Efficiency received a lower score, it indicates a need for future improvements, particularly in upgrading the system infrastructure and implementing load balancing mechanisms. Overall, the findings indicate that the HCCI eScholar fulfills its objectives by improving research management, reducing research duplication, and enhancing academic integrity.

References

Akhtar, N., et al. (2023). On the Innovation of Refined Scientific Research Management in Application-Oriented Universities in the Context of Big Data. Atlantis Press: Advances in Health Information Science.

Alayon, S.B., 2022. Status of Open Access, Digitization Initiatives, and Institutional Repositories in the Philippines. Nazarbayev University Library

Aljounid, H. (2025). Impact of modern technology on social research methods in educational contexts. International Journal of Education and Teaching Zone, 4(1), 21–31. https://doi.org/10.57092/ijetz.v4i1.325

Canlas, R., Piad, R., Lagman, J., 2021. An ISO/IEC 25010 based software quality assessment of a faculty research productivity monitoring and prediction system.

Chen, Y., Zeng, H., 2021. Cloud-based research storage systems: Scalability, security, and data integrity. International Journal of Digital Archives, Volume 17(3), pp. 210–220. Available online at: https://www.digitalarchivesjournal.com/cloud-research

International Organization for Standardization, 2025. ISO/IEC 25010:2023 – Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models. Available online at: https://www.iso.org/standard/35733.html

International Organization for Standardization & International Electrotechnical Commission. (2021). Software and systems engineering — Certification of software and systems engineering professionals — Part 3: Systems engineering (ISO/IEC 24773-3:2021). https://www.iso.org/standard/75051.html

Li, B., Han, L., 2013. Distance weighted cosine similarity measure for text classification. In: Intelligent Data Engineering and Automated Learning (IDEAL 2013), Springer, pp. 611–618. doi: 10.1007/978-3-642-41278-3_74.

Lynch, C.A., 2003. Institutional repositories: Essential infrastructure for scholarship in the digital age. Portal: Libraries and the Academy, Volume 3(2), pp. 327–336. doi: 10.1353/pla.2003.0039.

Suber, P., 2012. Open Access. MIT Press.
University of Manitoba Academic Learning Centre, 2020. Avoiding redundancy. Available online at: https://umanitoba.ca/student/academiclearning