Pamantasan ng Lungsod ng Maynila

REVAULT: A CLOUD REPOSITORY WITH SEO METADATA TAGGING FOR ARCHIVING RESEARCH-BASED WORKS OF PLM CISTM STUDENTS

A Thesis Presented to the
Faculty of Information Technology Department
College of Information System and Technology Management
Pamantasan ng Lungsod ng Maynila

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Information Technology
_____

By
Durante, KC D.
Insigne, Matthew Jacob B.
Lastra, Kristhia Cayle F.
Valeña, John Allen Troy E.
Yacub, Jetthro Karl C.

Thesis Adviser
Prof. Diony S. Abando

April 2025

Pamantasan ng Lungsod ng Maynila

TABLE OF CONTENTS

Pamantasan ng Lungsod ng Maynila

LIST OF FIGURES

Chapter One

INTRODUCTION

1.1 Background of the Study

Research archiving is essential in academic institutions where students produce technical projects and research papers. Traditional methods, such as printed copies or scattered digital storage, are inefficient, unstructured, and prone to loss. As reported by PLM faculty, stakeholders often struggle to locate research works. These issues hinder accessibility, making it difficult for students and faculty to locate relevant research for reference and academic development.

Digital repositories help address storage limitations, but many lack key features for effective research management. While digital repositories improve accessibility, privacy concerns around full-text storage remain a barrier according to the chief university librarian. Most systems do not support SEO metadata tagging which labels research based on themes and keywords, enhances searchability and organization. However, most repositories either lack SEO features or are designed for general document storage, making them not optimal for research archiving. This makes research papers more difficult to categorize and retrieve, leading to inefficient searches and duplication.

This study proposes ReVault, a cloud-based repository integrating version control and SEO metadata tagging for PLM CISTM students. By ensuring accessibility, organization, and continuous improvement, ReVault enhances research preservation and retrieval, fostering a more cohesive academic environment.

1.2 Statement of the Problem

This study aims to address the following questions:

1. How can research works be stored and retrieved more efficiently to reduce physical storage limitations and eliminate the need for labor-intensive, manual searching?
2. How can research records be systematically organized to prevent loss, ensure accessibility, and help students and faculty identify existing studies for reference and topic selection?
3. How can research retrieval be improved to allow users to find relevant studies based on content without relying solely on manual metadata tagging?

1.3 Objective of the Study

1.3.1 General Objective

To develop a cloud-based repository with SEO metadata tagging and ranks research papers based on query relevance, designed to streamline the archiving, retrieval, and management of research-based works for PLM CISTM students and faculty.

1.3.2 Specific Objectives

1.3.2.1. To design a Google cloud-based storage system that enhances research archiving and retrieval efficiency.

1.3.2.2. To implement a Schema.org/Dublin Core (DCMI) SEO metadata schema that systematically labels research works for structured indexing.

1.3.2.3. To integrate a full-text search and keyword extraction algorithm using BM25 that enables users to find relevant research papers based on content analysis.

1.4 Significance of the Study

This study aims to address the following questions: This study aims to develop ReVault, a cloud-based repository integrating version control and SEO metadata tagging to modernize the archiving and retrieval of research-based works within PLM CISTM. The system benefits the following stakeholders:

Pamantasan ng Lungsod ng Maynila (PLM): By establishing a structured digital repository, the university enhances research preservation, accessibility, and academic knowledge sharing, strengthening its research culture.

CISTM Faculty and Administrators: The system simplifies research management by providing an organized and searchable archive, reducing the burden of manual record-keeping and retrieval. It also ensures that research outputs remain accessible for academic evaluation and institutional reporting.

IT and Computer Science Students: ReVault assists students in finding relevant research works for topic selection, literature reviews, and reference purposes. Version control also supports iterative improvements, enabling students to refine and enhance their research over time.

Future Researchers: The repository ensures that research contributions are preserved for future use, preventing the loss of valuable academic works and minimizing topic duplication. The use of metadata tagging improves discoverability, allowing researchers to efficiently locate relevant studies.

Industry and Academic Partners: ReVault fosters collaboration by providing a centralized and accessible research archive, allowing external stakeholders to explore innovations and academic developments within CISTM.

1.5 Scope and Limitations

1.5.1 Scope

ReVault is a cloud-based repository designed for archiving and managing research-based works for 'Capstone Project' and 'System Integration and Architecture' created by undergraduate Information Technology students; Compiler Design and Thesis Writing created by Computer Science students at Pamantasan ng Lungsod ng Maynila (PLM) only. The system accommodates research papers completed within the past five years, ensuring accessibility to recent academic contributions. Accepted file formats are limited to .doc/.docx, .tiff, and .pdf to standardize document storage. The total storage capacity of the cloud storage for initial testing would be 5000 gibibyte (GiB) or 5.4 terabyte (TB).

1.5.1 Limitations

ReVault's archive is not exhaustive, as some research works may be unavailable due to loss, restricted access based on institutional permissions, or refusal to submit due to privacy concerns. The system exclusively stores documents and does not support supplementary materials such as source code, multimedia files, or datasets. The repository does not integrate with external university platforms, and all metadata tagging relies on the accuracy of provided information.

1.6 Definition of Terms

BM25   A full-text search ranking function integrated into ReVault that evaluates and scores research documents based on the relevance of their content to a user's search query.

Capstone Project   A comprehensive research project undertaken by PLM IT students that is stored, managed, and made accessible within ReVault.

Cloud Storage   The remote storage environment where users' documents are securely hosted after upload, ensuring scalable and reliable access.

Compiler Design   A research area for PLM CS students focused on the principles and construction of compilers, included in ReVault's collection.

Data Indexing   The process of organizing research documents and their associated metadata to enable fast, efficient searches.

Digital Document   An electronic version of any academic or research paper stored within the ReVault repository.

Digital Repository A centralized system for storing and managing digital research documents,

facilitating efficient access and long-term preservation.

Document Upload	The action by which users submit their research papers into ReVault, initiating the archiving process.
Dublin Core Metadata Initiative (DCMI)	A set of standardized metadata terms integrated into ReVault to systematically describe and index digital research documents.
File Format Standardization	The requirement that research documents be submitted in specific formats (e.g., .tiff, .doc, .docx, .pdf) to ensure consistency and compatibility.
Full-Text Search	A search function that indexes the complete text of each research document, allowing users to locate information beyond basic metadata.
Google Cloud Storage	The specific cloud infrastructure used by ReVault to host and manage archived research documents.
Keyword Extraction	The automated process of identifying significant words and phrases from research documents for generating effective metadata tags.

Metadata	Descriptive information attached to research documents—including keywords, author names, and publication details—used for categorization and retrieval.
Metadata Schema	The structured framework defining the types and formats of metadata used to tag research documents in ReVault.
Query Relevance	The metric used by BM25 in ReVault to determine how closely a research document's content matches a user's search query.
Research Archiving	The systematic process of storing research documents in a digital format to ensure their long-term accessibility and preservation.
Research Discoverability	The ease with which research documents can be located and accessed through effective metadata tagging and advanced search algorithms such as BM25 and TF-IDF.
Research-based Work	An academic document produced by PLM undergraduate students that presents original research findings and is archived in ReVault.

Research Preservation	Strategies and technologies employed to protect research documents from loss or deterioration, ensuring future accessibility.
Research Retrieval	The process of locating and accessing stored research documents within ReVault using search functions and metadata.
ReVault	The cloud-based repository system developed in this study for archiving and managing research-based works of PLM CISTM students and faculty.
SEO Metadata Tagging	The process of attaching descriptive keywords and data to research documents to enhance their discoverability within the repository.
System Integration and Architecture	A category of research focusing on the design and integration of complex systems, archived as part of PLM IT outputs.
TF-IDF (Term Frequency-Inverse Document Frequency)	An algorithm used in ReVault to extract important keywords from research

papers, aiding in the SEO metadata tagging process.

Thesis Writing

The process of composing a detailed academic thesis, which is one of the key document types of PLM CS outputs. archived in ReVault.

REVIEW OF RELATED LITERATURE

2.1   Related Literatures

2.1.1 Local Literature

Comparison and Evaluation of Different Methods for the Feature Extraction from Educational Contents

Aguilar, et al. (2020) conducted a comparative study of feature extraction techniques for analyzing educational resources modeled with the Learning Object Metadata (LOM) standard. Among the evaluated methods BM25, LSA, Doc2Vec, and LDA. BM25 emerged as a robust approach for semantic representation and retrieval tasks. The study assessed these techniques across diverse datasets, including scientific publications, learning objects, and patents, using unsupervised metrics like entropy, correlation, and coherence. The results showed that no single method universally outperformed the others, as performance depended on the dataset and evaluation criteria. Notably, the study highlights that while no technique achieves recommendation matches with 80% similarity, BM25 provides reliable results at a 70% threshold, making it a practical choice for educational recommendation systems.

Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method

Lan (2022) proposed a hybrid text similarity algorithm combining TF-IDF with semantic term analysis to overcome two key limitations of traditional methods: TF-IDF's inability to capture word meanings and semantic methods' tendency to create high-dimensional vectors. The study compared three approaches: standard TF-IDF, pure semantic analysis, and the proposed hybrid method. Evaluation using K-means clustering showed TF-IDF alone achieves 70-

75% accuracy, with optimal performance when the similarity threshold is set at 0.75 - beyond this point, clustering quality declines. The hybrid method demonstrated superior results by strategically combining TF-IDF's statistical power with semantic relationships from HowNet via a Term Similarity Weighting Tree structure, improving accuracy by approximately 2 percentage points while maintaining computational efficiency. The findings confirm that integrating statistical and semantic approaches yields better performance than either method alone for text similarity tasks, particularly in balancing accuracy with manageable dimensionality.

Dimensionality Reduction for Classification of Filipino Text Documents based on Improved Bayesian Vectorization Technique

Sueno et al. (2020) addressed the challenge of dimensionality reduction in classifying Filipino text documents by proposing an improved Bayesian vectorization technique. The study aimed to reduce the size of feature vectors used in text mining tasks while improving classification accuracy. The improved Bayesian vectorization technique was compared to the Term Frequency-Inverse Document Frequency (TF-IDF) method, with results evaluated using precision, recall, f-score, and accuracy metrics. The findings revealed that the improved Bayesian vectorization achieved 98% classification accuracy, significantly outperforming the TF-IDF method, which achieved only 76% accuracy. The study concluded that the improved Bayesian vectorization is highly effective for Filipino text classification but recommended using larger and more evenly distributed datasets for future research. Additionally, the study encouraged further exploration of Filipino text documents in machine learning applications.

Incorporating Rule-based Pattern Recognition Approach for Document Structure Classification on Cloud-based Document Management System

Buctuanon et al. (2021) developed an intelligent document organizing system named Docudile, which uses a rule-based pattern recognition approach to classify and organize documents into a hierarchical structure. The system addresses challenges in manual document organization, such as inconsistency, human error, and time consumption. Docudile integrates a cloud-based document management system to ensure accessibility from remote locations and uses Term Frequency-Inverse Document Frequency (TF-IDF) for document retrieval. The system achieved 98% accuracy in document classification and 89% accuracy in document retrieval, outperforming Naïve Bayes Classifier and Support Vector Machine (SVM) algorithms. The study highlighted the system's ability to classify documents into categories like letters, memos, and unclassified documents. However, performance testing revealed that the system's processes, such as uploading, hierarchical structure generation, and cloud syncing, were resource-intensive and internet-dependent. To address this, the system offers local storage as a backup. The study recommends integrating additional document types, such as certificates and research papers, into the classification process for future improvements.

Document Management System Using Optical Character Recognition, Clustering, Watermarking and QR Coding Algorithms

Document Management System (DMS) was designed for the Department of Budget and Management Regional Office VI, incorporating Optical Character Recognition (OCR) to scan and convert documents into text. The system employs the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm for clustering and applies watermarking and QR coding algorithms for document authentication and security. Developed using the Evolutionary Prototyping model, the system was

evaluated by employees and was found to generate authenticated and secure documents with accurate watermark and QR code imprints Santos et al. (2023).

Students' use of Cloud Storage in their Studies: A Case of a Private University in the Philippines.

The study conducted by Santos, et al. 2023) examines the adoption and use of cloud storage systems, such as Google Drive, Google Cloud, iCloud, and Microsoft OneDrive, among students at a private university in the Philippines. It analyzes the factors influencing their use and finds that performance expectancy and social influence significantly impact students' intention to utilize cloud storage. The study highlights that students are more inclined to adopt cloud storage when they perceive it as beneficial for efficiently managing academic tasks, such as storing, sharing, and accessing documents from any device or location. For instance, the seamless integration of Google Cloud with productivity tools like Google Workspace enables students to collaborate on group projects in real time and securely store large volumes of data. Additionally, the study points out that social influence, such as encouragement from peers and instructors, further promotes the adoption of cloud storage systems.

Determining factors affecting Filipino consumers' behavioral intention to use cloud storage services: An extended technology acceptance model integrating valence framework

Cloud storage services offer users the ability to digitally store, manage, back up, and share data remotely, providing significant advantages over traditional storage methods, such as accessibility from various devices and locations. However, despite these benefits, adoption rates remain low in developing countries like the Philippines. A study investigated factors influencing Filipino consumers' intentions to use cloud storage, employing the Extended Technology Acceptance

Model (ETAM) integrated with the Valence Theoretical Framework. The study analyzed twelve variables, including perceived benefit, perceived usefulness, job relevance, perceived risk, and subjective norms, through data collected from 431 cloud users, mainly students and professionals. Job relevance also played a key role, indicating that users are more inclined to adopt cloud storage when it supports their professional or academic needs. In contrast, factors such as cost, experience, and voluntariness showed minimal impact on adoption decisions. The study recommends that cloud storage providers emphasize practical benefits and address security concerns to build user trust and encourage widespread use according to Cruz et al. (2024).

Integrated web-based research document tracking system with barcode for Medical Colleges of Northern Philippines and International School of Asia and the Pacific

The study developed an integrated web-based research document tracking system for medical colleges in Northern Philippines and the International School of Asia and the Pacific, aiming to streamline document management through automation and cloud storage integration. By utilizing cloud storage, the system ensures secure, centralized data access and real-time updates, enabling efficient retrieval and archival of research documents. The use of barcodes further automates data entry and retrieval, minimizing manual errors and enhancing productivity. The study employed a descriptive developmental research design and the Waterfall Model Methodology to develop and evaluate the system's usability and effectiveness. Data was gathered through questionnaires and interviews with students, teachers, and staff, highlighting the importance of combining cloud storage with workflow tools to address challenges related to volume, security, and real-time document tracking. The findings underscore the potential of integrating

cloud storage in research document tracking systems to optimize workflows and promote efficient document management (Delos Santos et al., 2024).

Central Repository System: A Quantitative Analysis of Its Feasibility and Applicability in MMSU

The study explored the feasibility and applicability of a Central Repository System (CRS) at Mariano Marcos State University (MMSU), focusing on the integration of cloud storage through Microsoft 365 to optimize knowledge management and streamline university processes. Cloud storage, particularly through Microsoft OneDrive and SharePoint, played a crucial role in centralizing data, facilitating seamless collaboration, and ensuring secure and efficient document management. By leveraging cloud technology, the CRS allowed for real-time access to important files and enhanced workflow efficiency across various departments. Despite encountering challenges related to login issues, data inconsistency, and resistance to change, the study found that employees were generally familiar with Microsoft 365 and recognized the value of cloud storage in improving productivity. To maximize the system's effectiveness, the study recommended comprehensive training, support initiatives, and addressing technical obstacles to fully realize the benefits of cloud storage in advancing university operations (Ruguian and Pola, 2023).

Data Archiving for Philippines Higher Education in Compliance with Accrediting Bodies – A CloudBased Data Processing System for Universities

Yap & Velasco (2024) proposed a Cloud-Based Data Processing System designed for higher education institutions in the Philippines to address inefficiencies in traditional data management practices. The system leverages cloud storage to securely archive and manage large volumes of academic data while enhancing scalability, flexibility, and cost-effectiveness. By utilizing modern cloud

technologies, the system ensures that data is stored safely and can be accessed efficiently, even as the volume of stored information grows. Evaluation of the system revealed efficient processing of various file types, with CSV files demonstrating faster execution times compared to more complex formats like JPEG and PDF. Despite some variability in processing speed, especially with image-based files, the system consistently delivered reliable data archiving and retrieval. The proposed solution addresses the challenges faced by universities in managing vast amounts of academic data, offering a practical and modern approach to data processing and secure cloud storage.

Evaluation of State University and College Institutional Website in the Philippines

Ventayen (2019) evaluated the institutional websites of 114 State Universities and Colleges (SUCs) in the Philippines, focusing on four categories: accessibility, experience, marketing, and technology. Using the Nibbler testing tool, the study scored each website from 0 to 10 and ranked them based on their overall performance. The results showed that the University of the Philippines and Visayas State University ranked first, while Pangasinan State University ranked fifth with an overall score of 8.3, a significant improvement from its 2015 score of 4.9. These findings are highly relevant to the development of efficient and user-friendly digital platforms, as they emphasize the importance of accessibility, user experience, and technology in enhancing website performance. By adopting similar evaluation metrics and benchmarking practices, digital platforms can ensure high-quality performance and user satisfaction. The study also highlights the importance of institutional recognition and quality measures, which are critical for creating a reputable and reliable repository for research-based works.

Technical Review and Analysis on Students Academic Website Projects Using GTmetrix Web Speed and Optimization Tool

Quinito & Catipay (2019) conducted a technical review and analysis of three local government website projects developed by Bachelor of Science in Information Technology (BSIT) students at Mindanao University of Science and Technology-Jasaan Campus. The study aimed to evaluate the websites' performance in terms of access speed and page efficiency using the GTmetrix web speed optimization tool. The results revealed that the websites were unoptimized based on GTmetrix's predefined speed categories and PageSpeed rules. Key areas for improvement included image optimization, use of expired headers, implementation of Content Delivery Networks (CDNs), and compression of resources using gzip. The study emphasized the importance of following global best practices in web development to enhance website performance and user experience. Recommendations included deferring JavaScript parsing, combining CSS files, and using cookie-free domains to minimize data traffic. The study concluded that academic institutions should prioritize teaching web optimization techniques to improve student projects and align them with industry standards.

Social Media Marketing Strategies: Implementation, Optimization, and Integration among Private Higher Education Institutions

Aquino (2024) explored the implementation, optimization, and integration of social media marketing (SMM) strategies among private higher education institutions (PHEIs) in Calamba City. The study aimed to understand how PHEIs develop and execute SMM plans, focusing on content creation, audience targeting, platform selection, and integration with other marketing channels like email marketing and search engine optimization (SEO). Using a survey, the study assessed the level of SMM implementation, optimization, and integration among respondents. Results showed a high level of SMM implementation and

optimization, with significant differences observed based on respondents' roles, length of involvement, and student enrollment. However, no significant differences were found in integration levels across profile variables. The study concluded that PHEIs need a proposed action plan to sustain their SMM efforts and recommended further integration of SMM with other marketing strategies to enhance student recruitment and brand awareness.

Influence of Search Engine Optimization (SEO) Towards Purchase Intention of Online Shoppers

Caro, et al. (2024) investigated the influence of Search Engine Optimization (SEO) on the purchase intention of online shoppers, focusing on Zalora, an e-commerce platform. The study assessed how SEO factors like quality content, website interface, and ranking algorithms impact customer purchase intention. Using stratified random sampling, 375 online shoppers in Region III (Central Luzon) were surveyed. Results showed that quality content, user-friendly design, and personalized search results significantly influenced purchase intention, with p-values less than 0.05. The study concluded that optimizing these SEO factors enhances customer engagement and satisfaction, driving purchase decisions. These findings are relevant to digital platforms aiming to improve content discoverability, user experience, and engagement through effective SEO strategies.

Organic Search Engine Optimization for Museum Websites in 2023: Strategies for Improved Online Visibility and Access

Rivera (2023) examined the use of organic search engine optimization (SEO) strategies in museum websites across the United States, aiming to enhance online visibility and access to collection metadata. The study involved technical audits and manual reviews of 214 art museum websites, identifying common challenges such as the lack of technical optimizations, insufficient use of

Schema.org markup for rich search results, and inconsistent application of SEO practices across digital platforms. Incorporating SEO from the outset can minimize future rework and help maintain a consistent online presence. Furthermore, the research highlighted the potential benefits of participating in data aggregation platforms like Google Arts & Culture, which automatically enrich metadata and improve discoverability. The study concluded that to maximize public engagement, museums must adopt adaptive and ongoing SEO practices, using automated tools where possible and regularly auditing their digital presence to ensure accuracy and effectiveness.

Towards the Feasibility of Instituting a Philippine Digital Audio Library: A Case Study

The study explored the feasibility of establishing a Philippine Digital Audio Library by examining the integration of existing audio collections into a unified digital archive. One of the significant challenges identified was the organization and preservation of vast amounts of audio data, where metadata tagging played a crucial role in ensuring data accessibility and sustainability. Through meticulous metadata tagging, the project efficiently cataloged audio recordings, capturing essential details such as audio quality, recording context, and cultural significance. This practice greatly enhanced data retrieval and preservation, allowing the library to maintain a high level of accessibility for academic and cultural research. Moreover, metadata tagging streamlined the integration of new collections, making the digital library a reliable repository of Philippine audio heritage. The study ultimately emphasized that metadata tagging is fundamental to the success of digital audio libraries, particularly when managing large and diverse collections (Li, 2021).

Evidence Summary Theme: Information Access & Retrieval

It addressed challenges related to information access and retrieval within library systems by analyzing how metadata tagging can significantly improve cataloging and data organization. By leveraging metadata tagging practices, the library systems enhanced discoverability and streamlined information retrieval, reducing manual search time and improving user experience. The research utilized publicly available catalogue metadata to assess tagging efficiency and demonstrated that comprehensive metadata schemas were crucial for managing both digital and physical collections. Furthermore, the study highlighted the importance of continuously updating metadata structures to keep pace with technological advancements and evolving user needs, making metadata tagging a vital aspect of modern library management (Inglis, 2023).

Legacy Language Materials in the Ernesto Constantino Collection: Challenges and Lessons for Building a Philippine Language Archive

This study examined the use of legacy materials from the Ernesto Constantino Collection, highlighting the importance of metadata tagging in organizing linguistic data and preserving cultural heritage. Without consistent metadata practices, the archive faced significant challenges related to data fragmentation and disorganization. To address this, the study employed metadata tagging to systematically capture linguistic features, dialect variations, and contextual information, which not only improved data coherence but also facilitated cross-referencing and comparative analysis. By implementing metadata tagging, the project ensured the sustainable management and long-term preservation of valuable linguistic materials, making the archive more accessible and useful for researchers and cultural scholars (Or & Estrellado, 2023).

Enhancing university operations: A study of the electronic document management systems (EDMS) of one higher education institution

The study conducted by Aliazas, et. al (2024) examines the implementation of an Electronic Document Management System (EDMS) in a higher education institution in the Philippines to enhance operational efficiency. One of the critical components that contributed to the system's success was the use of metadata tagging, which played a pivotal role in organizing and categorizing documents for easy retrieval and management. Metadata tagging allowed the EDMS to facilitate quick access to research materials and institutional records, enabling users to perform full-text searches and efficiently locate relevant documents. This enhanced data accessibility not only streamlined administrative workflows but also promoted better knowledge sharing and collaboration among departments. The study also identified challenges such as user adaptation and technical difficulties but highlighted how metadata tagging significantly addressed issues related to document organization and retrieval.

Assessment of enhanced online documents repository service

The study conducted by Amado & Gamba (2023) assessed an enhanced online document repository service for school records management in Matnog I District. The research aimed to evaluate existing online records management systems, identify areas for improvement, and implement advanced features such as standalone execution, network adaptability, cross-site scripting prevention, and one-time password functionality. One of the most significant improvements introduced was the incorporation of metadata tagging, which played a crucial role in organizing and indexing documents for efficient retrieval and management. Through metadata tagging, users could swiftly search and access relevant school records, significantly enhancing the usability and accessibility of the system. The findings indicated that most of the 23 participating schools lacked a proper online

13

records management system before the enhancement, but respondents strongly agreed that the improved repository addressed critical issues related to GUI, accessibility, usability, and security. Additionally, the study proposed a training program to ensure proper implementation and utilization of the enhanced repository, emphasizing the importance of metadata in maintaining a secure and well-organized digital record system.

2.1.2 Foreign Literature

Design and Implementation of Institutional Repository with Dublin Core Metadata

Institutional repositories are essential for managing academic collections, including undergraduate theses, journals, and research papers. Reza et al. (2022) designed an institutional repository utilizing Dublin Core metadata at Ibnu Sina University to enhance information management and accessibility. The repository was developed using the open-source DSpace platform and employed PostgreSQL as its storage database. The implementation followed the waterfall model, comprising five phases: needs analysis, system design, coding, testing, and implementation. By incorporating Unified Modeling Language (UML) diagrams such as use case and sequence diagrams, the repository ensured a structured system architecture. The study found that utilizing Dublin Core metadata significantly improved the organization and retrieval of academic resources. Furthermore, the system facilitated efficient indexing and search capabilities, making it easier for users to locate relevant documents. This research highlights the benefits of metadata-driven institutional repositories in higher education by improving document organization and long-term accessibility.

Integrating Public Reported Evidence Collection, Public Court Records Archive, and Realizing Secure and Decentralized Case Document Management Using IPFS and Hyperledger Fabric Blockchain: An Implementation Study

Banjan et al. (2023) explored a cloud and blockchain-enabled document management system that integrates InterPlanetary File System (IPFS) and Hyperledger Fabric blockchain to store legal documents. The study highlights the significance of secure, decentralized, and transparent storage mechanisms in legal document archiving. The proposed system enhances document integrity by allowing public evidence collection while ensuring privacy and confidentiality through blockchain security measures. By leveraging Google Firebase, the system provides a centralized digital repository with enhanced accessibility. The findings emphasize that a digital repository integrating blockchain technology can improve document authenticity and traceability. This approach not only modernizes traditional case management systems but also enhances public access to court records, fostering legal transparency.

Publishing E-resources of Digital Institutional Repository as Linked Open Data: An Experimental Study

Kar and Das (2020) examined the transformation of institutional repositories into Linked Open Data (LOD) frameworks using the DSpace digital library system. Their study focused on how bibliographic metadata stored in PostgreSQL databases could be converted into structured LOD datasets. The research emphasized the importance of semantic web technologies in enhancing interoperability between digital repositories across institutions. By integrating Apache Jena Fuseki, metadata could be converted into Resource Description Framework (RDF) models and queried using SPARQL. The study demonstrated that implementing LOD in institutional repositories enables the seamless linking of

related research resources, improving accessibility and knowledge-sharing among institutions. This research underscores the necessity of standardized metadata schemas like Dublin Core for ensuring structured indexing in digital repositories.

The Cloud-enabled Architecture of the Clinical Data Repository in Poland

Augustyn et al. (2021) analyzed the design of a cloud-enabled clinical data repository in Poland, focusing on high availability and scalability. The system incorporated a PostgreSQL HA cluster—Crunchy PostgreSQL—within the Google Cloud Platform to optimize data storage and accessibility. Traditional clinical repositories were often constrained by limited local storage, security risks, and interoperability issues, but cloud computing resolved these limitations by offering flexible deployment options. The study proposed an architecture adaptable to various environments, including cloud-agnostic, cloud-dependent, and on-premise solutions. The findings demonstrated that cloud-enabled repositories ensure better data exchange between institutions while enhancing security and accessibility. The research highlights the transformative potential of cloud-based digital repositories in various fields, particularly for managing large-scale datasets.

When Rain Clouds Gather: Digital Curation of South African Public Records in the Cloud

Shibambu and Ngoepe (2020) investigated the adoption of cloud storage for public record management in South Africa. The study revealed that government departments were reluctant to fully transition to cloud-based repositories due to concerns about data security, jurisdiction, and legal implications. Although public officials occasionally saved records in cloud systems like Microsoft Azure and Google Cloud, the lack of a cohesive legal framework hindered full-scale implementation. The findings underscored the necessity of a standardized policy for cloud-based digital repositories to ensure legal compliance and data integrity.

The study concluded that while cloud storage offers significant advantages for digital record preservation, government institutions must rigorously evaluate security and regulatory concerns before widespread adoption.

Analysis of Frequently Appearing Words in the Titles of 2023 Research Grant Winners in Indonesia Using the TF-IDF Method

Setiawan et al. (2023) utilized the TF-IDF method to analyze the most frequently occurring words in the titles of research grants awarded in Indonesia. The study aimed to identify trends in research topics by quantifying term importance within a dataset of funded research titles. TF-IDF was used to weigh words based on their frequency in individual titles compared to the entire corpus. The findings revealed dominant themes and keywords that characterized successful research proposals, offering insights into topic selection strategies for future applicants. The study emphasized the role of TF-IDF in identifying research trends and optimizing proposal alignment with prevailing academic interests. By applying statistical analysis to title patterns, the research demonstrated the effectiveness of TF-IDF in keyword extraction and thematic analysis, enhancing the understanding of grant-winning tendencies.

Extending Neural Keyword Extraction with TF-IDF Tagset Matching

Koloski et al. (2021) explored the enhancement of neural keyword extraction methods through TF-IDF tagset matching. Their research compared transformer-based keyword identification models with a traditional TF-IDF approach, demonstrating how TF-IDF can complement deep learning techniques to improve recall. The study utilized datasets from European news media in multiple languages, applying TF-IDF to refine keyword selection by cross-referencing extracted terms with predefined tagsets. Results indicated that integrating TF-IDF with neural models significantly enhanced keyword detection, making it suitable

for large-scale content categorization. The findings underscored TF-IDF's ongoing relevance in hybrid keyword extraction frameworks, particularly for improving the contextual relevance of identified terms.

Final Year Project Repository with Automated Classification Using TF-IDF

Bakri et al. (2023) developed a repository system for final-year projects (FYP) at Universiti Teknologi MARA (UiTM), incorporating TF-IDF for automated classification. The study addressed the challenge of students lacking access to prior FYPs for reference and topic selection. By applying TF-IDF, research projects were categorized based on term importance, enabling efficient retrieval and structured organization. The repository improved accessibility by systematically indexing projects according to extracted keywords, facilitating student learning and academic planning. The study highlighted the value of TF-IDF in academic repositories, demonstrating its utility in enhancing document classification and retrieval efficiency.

TF-IDF Implementation for Similarity Checker on the Final Project Title

Meidelfi et al. (2021) applied TF-IDF in conjunction with cosine similarity to develop a title similarity checker for student final projects at Politeknik Negeri Padang. The study aimed to streamline the manual verification process used by academic committees to identify duplicate project titles. The proposed system utilized TF-IDF to assign weight to words, followed by cosine similarity calculations to determine title resemblance. Experimental results indicated that TF-IDF significantly improved the accuracy and efficiency of title matching, reducing redundancy in project submissions. The study validated the use of TF-IDF for academic document comparison, providing an automated solution for ensuring title uniqueness in educational institutions.

TF-IDF Keyword Extraction Method Combining Context and Semantic Classification

Wang and Ning (2020) proposed an advanced TF-IDF keyword extraction technique that incorporated context vectors and semantic classification. Their study addressed limitations in traditional TF-IDF models, which often overlook contextual influence and word polysemy. The researchers introduced an expandable linked list structure to manage multiple meanings of words and applied semantic similarity measures before inserting terms into keyword sets. By integrating these enhancements, their method improved keyword relevance and accuracy in text classification tasks. The experimental results, conducted using the Sogou news dataset, demonstrated superior performance compared to conventional TF-IDF models. This study showcased TF-IDF's adaptability when combined with contextual and semantic processing techniques, refining its applicability in keyword extraction.

A Page-topic Relevance Algorithm Based on BM25 and Paragraph-Semantic Correlation

The relevance of web pages or documents plays a crucial role in the efficiency of search engines, directly impacting a user's ability to find the information they need. Shi et al. (2021) explored the role of BM25 as a fundamental ranking algorithm in information retrieval, particularly in webpage ranking. They acknowledged BM25's reliance on term frequency to establish a retrieval probability model between search keywords and page content. However, traditional BM25 implementations struggle with ambiguous search terms, often retrieving irrelevant information. To address this limitation, the study proposed an enhanced topic relevance algorithm that integrates BM25 with paragraph-semantic correlation, leveraging deep neural networks for improved accuracy. Their experimental findings demonstrated that the combined approach outperforms the

traditional BM25 algorithm by refining search relevance through semantic classification. This hybrid methodology highlights the adaptability of BM25 when augmented with contextual models, reinforcing its role as a reliable ranking function in search engine development.

Improving Ranking in Document-based Search Systems

Menon et al. (2020) investigated the performance of BM25 compared to alternative ranking models in document-based search systems. They emphasized that BM25 is a feature-based baseline model that determines relevance by analyzing the frequency of query terms within documents rather than evaluating their proximity. Their study examined neural network-based ranking techniques, comparing semantic similarity approaches with traditional term-matching models. Although modern ranking methods enhance retrieval accuracy, their computational demands pose significant challenges, making BM25 a favorable alternative due to its efficiency. Their findings confirm that BM25 remains highly effective in ranking document-based search results, particularly in systems requiring scalable and interpretable ranking models. The study underscores the continued relevance of BM25 in balancing retrieval performance with computational feasibility.

Integrating Okapi BM25 and Jaccard Algorithms in Thesis Search Engine

Bronda-Ecraela and Doctora (2024) explored the implementation of BM25 in an academic thesis search engine, integrating it with the Jaccard algorithm to optimize document retrieval. The research focused on ranking thesis documents efficiently while addressing computational performance. The study evaluated the effectiveness of BM25 in document ranking, demonstrating its superior ability to retrieve relevant search results compared to the Jaccard similarity measure. While Jaccard was useful in identifying similar documents, its tendency to return all related items resulted in information overload. Computational efficiency tests

further validated BM25's advantage in delivering prompt and accurate search results. The research concluded that BM25 provides a robust ranking mechanism for academic repositories, offering insights into optimizing document search systems to enhance user experience and retrieval efficiency.

On the Interpolation of Contextualized Term-based Ranking with BM25 for Query-by-Example Retrieval

Abolghasemi et al. (2022) examined the generalizability of BM25 in query-by-example (QBE) retrieval, where a full document is used as a query rather than short keyword-based searches. The study compared BM25's ranking performance against deep contextualized term-based models such as TILDE and TILDEv2. While transformer-based ranking models offer semantic advantages, their reliance on BERT inference at query time introduces computational complexity. The study revealed that BM25 remains competitive in QBE retrieval, even surpassing the performance of TILDE and TILDEv2 in some instances. This finding challenges previous assumptions about BM25's limitations in long-query scenarios. Furthermore, the researchers explored score interpolation techniques, demonstrating that combining BM25 with transformer-based models results in improved retrieval effectiveness. Their study highlights BM25's adaptability and relevance in information retrieval systems beyond traditional keyword-based search paradigms.

Text Mining in Digital Libraries using OKAPI BM25 Model

Tinega et al. (2022) investigated the application of BM25 in digital libraries, emphasizing its role in enhancing text mining and retrieval efficiency. The researchers noted that many digital libraries suffer from ineffective information retrieval models that lack relevance ranking, leading to suboptimal search results. Their study implemented BM25 within a library search framework and compared

its performance against Boolean and Vector Space models. Experimental results indicated that BM25 significantly improved document ranking by prioritizing terms based on their relative frequency and distribution. The findings support the adoption of BM25 as a foundational ranking algorithm in digital libraries, ensuring that retrieved documents are ordered according to their relevance. This study reinforces BM25's importance in structuring and optimizing search capabilities within academic and research-focused repositories.

A heuristic approach on metadata recommendation for search engine optimization

Traditional search engine optimization (SEO) techniques lack a logical foundation for selecting metadata, yet metadata plays a crucial role in search engine indexing and ranking. An and Jung (2020) propose a heuristic approach to metadata recommendation for improving rankings in Search Engine Result Pages (SERP). Their method consists of two steps: first, extracting influential keywords and metadata from high-ranked websites based on frequency and weight, and second, prioritizing terms according to semantic relevance using a competitive learning model. The study evaluates the effectiveness of this method through experimental validation using Google search queries. The results demonstrate that their approach increases website traffic by strategically using high-ranked keywords and semantic relevance in metadata selection. This research highlights the significance of structured metadata in improving online visibility and suggests that similar methods could be beneficial in academic repositories to enhance research discoverability.

Data Deposit in a CKAN Repository: A Dublin Core-Based Simplified Workflow

Many data repositories employ the Dublin Core metadata schema to enhance searchability and access. Karimova et al. (2020) examine a CKAN-

powered repository that integrates a simplified Dublin Core-based dataset deposit process. They outline a workflow where researchers describe their data and complete publication in a structured manner. The study highlights the challenges researchers face in organizing, describing, and depositing data due to limited knowledge of metadata and data publication processes. Through user feedback, the research identifies difficulties in data description but also demonstrates researchers' willingness to participate in metadata-enhanced data publication. The study underscores the importance of structured metadata schemas in ensuring standardized and efficient data deposits, which can improve research accessibility and long-term preservation.

Digitization of Archival Data and Metadata in Archaeology: The Case of Ancona

In the process of digitizing a vast amount of archaeological data, the implementation of metadata standards plays a critical role in ensuring accessibility and preventing information loss. Iacopini (2024) discusses the digitization of approximately 12,000 files related to archaeological contexts in Ancona, Italy. To avoid these digital records becoming "silent" documents, the study employs a metadata system based on the Dublin Core Metadata Initiative (DCMI) standard. This metadata framework enhances the findability and usability of archived data by providing structured descriptions and improving interoperability across various systems. The study demonstrates how metadata can transform raw digital archives into searchable and systematically categorized resources, ensuring long-term preservation and accessibility of historical and scientific documentation.

Improving the Documentation and Findability of Data Services and Repositories: A Review of (Meta)data Management Approaches

Metadata is often perceived as an obligatory aspect of data management rather than a tool for enhancing data findability. Řezník et al. (2022) challenge this notion by presenting metadata as an essential component for improving the accessibility and organization of digital repositories. Their study identifies Dublin Core and ISO 19115 as the predominant international metadata standards for geosciences, with widespread adoption in regions like Australia, China, Europe, and the United States. They categorize metadata management into four levels: unstructured data, schema-based metadata with literal values, metadata with URIs, and linked open (meta)data. The research emphasizes the necessity of integrating structured metadata into search engine algorithms, particularly through Schema.org, to optimize data discoverability. The study provides valuable insights into metadata management strategies that enhance the usability of digital repositories and ensure seamless access to academic and scientific research.

Tools for Interoperability Between Repositories of Digital Resources

Interoperability among digital repositories is essential for effective information exchange and accessibility. Guerrero et al. (2020) explore various technological tools that facilitate interoperability by utilizing metadata standards such as Dublin Core and interoperability protocols like the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The study employs qualitative research methods to analyze different tools, along with a case study that demonstrates the application of Python-based technologies for interoperability. Using the Rapid Application Development (RAD) methodology, the researchers assess the effectiveness of metadata-driven interoperability solutions, with a focus on the Sickle tool. Their findings highlight both the benefits and challenges of implementing metadata standards for repository integration, demonstrating that

structured metadata plays a crucial role in enhancing research accessibility across multiple digital platforms. Despite limitations such as insufficient documentation for some tools, the study underscores the importance of well-structured technological solutions in improving interoperability and metadata-based search functionalities.

2.2 Related Studies

2.2.1 Local Studies

The Development of A Cloud-Based University Research Repository Software Using A Configurable Subscription Model

Alvez (2022) developed a cloud-based university research repository that utilizes cloud storage to address the limitations of traditional hardbound archives, which often remain unused on library shelves. By leveraging cloud technology, the system makes academic research more accessible and impactful through electronic publication, subscription, and access to research papers. The repository's configurable subscription model allows users to customize how they publish and access materials while offering remuneration to researchers when their work is cited or downloaded. This approach not only incentivizes continuous scholarly output but also enhances research visibility and collaboration among researchers, students, and institutions by creating a unified platform for interaction. The use of cloud storage ensures that research papers are easily accessible from any location, promoting long-term relationships between stakeholders while minimizing the workload related to maintaining physical archives. Furthermore, the system addresses challenges posed by the global pandemic by ensuring heterogeneous access to resources, making it highly beneficial and sustainable for educational institutions.

The Presentation and Long-Term Preservation of the Oral Histories in the Houston Asian American Archive

The Houston Asian American Archive (HAAA) utilizes cloud storage as a fundamental component in preserving and managing complex oral histories, making it highly relevant to the use of cloud storage in research repositories. This system, developed with the support of Rice University, leverages cloud-based tools and storage to efficiently track, store, and preserve digital oral histories, including transcripts, consent forms, photographs, and time-syncing data. Initially, physical storage methods like CD-ROMs and thumb drives were employed, but as technology advanced, the program transitioned to a fully cloud-based storage system. This shift not only enhanced accessibility and organization but also streamlined the process for archivists and team members. Through cloud storage, the archive ensures long-term preservation while enabling remote access and collaboration, which is vital for maintaining comprehensive and reliable records. Moreover, this model demonstrates how cloud storage can be effectively utilized to manage large volumes of research data, offering valuable insights for developing modern research repositories, especially in contexts where accessibility and data preservation are critical. This approach can also be particularly beneficial for documenting Filipino-American oral histories within the archive, promoting cultural preservation and representation through digital means (Focke, 2019)

Actual usage assessment among cloud storage consumers in the Philippines using a machine learning ensemble approach

The study assessed the actual use of cloud storage among Filipinos using a machine learning ensemble approach. Key factors influencing usage were identified as voluntariness, subjective norm, perceived benefit, perceived usefulness, and perceived ubiquity. The research demonstrated that both personal and professional use of cloud storage, including platforms like Google Drive, has

become increasingly popular due to its convenience and efficiency in managing and retrieving data remotely. It achieved an impressive accuracy rate of 93%, highlighting that users' willingness to adopt cloud storage is influenced by its practicality and social encouragement from peers. Practical recommendations were provided to enhance cloud storage uptake, including marketing strategies and efforts to boost digital literacy among users (Ong et al., 2024).

File and storage management system towards cloud computing technology

The study conducted at the Asian College of Technology, Cebu focused on enhancing file management efficiency through the implementation of a cloud-based File Storage and Management System (FSMS). The previous system, assessed during the academic year 2014–2015, was found to be inefficient and cumbersome for the 43 employee respondents. To address these challenges, the study developed a new FSMS that leveraged cloud storage to centralize and streamline file management processes. Unlike traditional methods where files were stored locally or on isolated servers, the cloud storage approach enabled real-time access, easy retrieval, and secure file management from any location. The integration of cloud storage not only ensured better organization and accessibility of documents but also minimized data loss and improved collaborative efficiency among employees. This innovation proved essential for optimizing the school's file management tasks, demonstrating the transformative potential of cloud storage in modernizing institutional workflows (Sagun, et. al 2021).

eDALAYON: A Document Management and Monitoring System for the Department of the Interior and Local Government, Philippines

Improving document handling processes, eDALAYON was designed as a document management and monitoring system for the Department of the Interior and Local Government (DILG) in Negros Occidental, Philippines. Featuring four

core modules—user access roles, document management, document tracking, and reporting—the system was developed using the Agile Scrum methodology. It integrates cloud-based storage, version control, and access control mechanisms to enhance security and collaboration. Evaluations by DILG personnel revealed significant reductions in workload and document retrieval time, with high usability and user satisfaction. Despite its success, further quality assurance testing and considerations for adaptation to broader contexts were identified as areas for improvement (Dela Cruz et al., 2023).

Challenges to Digital Services in Philippine Academic Libraries

Lagas & Isip (2023) investigated the challenges to digital services in Philippine academic libraries, emphasizing the difficulty in sustaining digitization programs. Despite the growing need for digital repositories, barriers such as inadequate staffing, lack of physical space, and insufficient funding hinder implementation. The study reveals that most digitization efforts are reactive rather than strategic, often driven by external factors rather than being part of a comprehensive plan. This disconnect between the demand for digital resources and the libraries' capacity to provide them became even more apparent during the COVID-19 pandemic. To address these gaps, the study recommends integrating digital archiving strategies into the curriculum and forming partnerships for sharing resources and expertise. This highlights the importance of sustainable and scalable digital repositories, like cloud-based systems, that can accommodate academic content without being limited by physical infrastructure or staffing challenges.

Web-based Document Management System

Nagrama, et al. (2024) developed a web-based document management system to digitize paper-based processes in an educational office in Pagadian City, Philippines. The system replaced traditional methods of document dissemination

and collection with an online platform that allows authorized personnel to post, access, download, and submit documents efficiently. It features secure access, automated processing, and real-time updates, streamlining the entire document management process. The success of this system demonstrates how a web-based approach to document management can significantly enhance administrative efficiency and transparency. This study supports the idea that a cloud-based research repository can offer advantages by centralizing academic documents, making them accessible to authorized users from anywhere, and improving overall efficiency in managing academic records.

Centralized documents repository

A Centralized E-Document Repository System (CERS) was developed to enhance document and record management in organizations. The system improves office productivity by streamlining file management, sharing, and distribution while ensuring security. By eliminating physical file storage, CERS saves space and time, integrating a secure database with backup and disaster recovery features. Evaluations by end-users and IT experts resulted in a high effectiveness, efficiency, and usability rating, with a grand mean score of 4.826. The study concludes that CERS significantly enhances document management processes, offering users convenience and easy access (Garcia et al., 2023).

An Agile Methodology Approach to a Record Management and Monitoring System for Company B

Andaya et al. (2020) implemented a web-based Record Management and Monitoring System for Company B to address persistent challenges related to manual record-keeping, financial transparency, and queue management. The system was developed using Agile Methodology, which allowed for iterative improvements and active user feedback throughout the development process. By

implementing SEO and metadata tagging, the system enhances data retrieval speed and accuracy, making information more accessible to authorized users. This approach not only optimizes transparency and consistency in business processes but also ensures that stored data can be efficiently managed and retrieved when needed. Such practices could serve as a model for academic repositories, where properly tagged and searchable research files are essential for maintaining organized and accessible digital libraries.

Computer laboratory network centralization files management system

Computer laboratory network centralization files management system (Ledonio, 2021) developed a centralized file management system designed to improve file organization and automate backups in computer laboratories. The system was built through the Prototyping methodology, allowing for iterative testing and refinement based on real-world feedback. It features a file backup daemon that regularly schedules and transfers newly created files from client computers to a central server, thereby minimizing data loss and ensuring timely backups. This centralized approach significantly enhances data organization and retrieval by consolidating files into a single, secure location. Furthermore, by integrating SEO and metadata tagging, the system ensures that files can be efficiently tracked and searched, which is vital in maintaining a well-organized data management structure. This system not only meets performance and security requirements but also demonstrates the practical benefits of employing metadata strategies to streamline file management, particularly in academic settings where data accuracy and availability are crucial.

2.2.2 Foreign Studies

Effect of Institutional Digital Repository Information Services on Utilisation of Library Resources by Postgraduate Students at Selected University Libraries in Kenya

Kavila, Gichohi, and Mwiti (2024) investigated how institutional digital repositories impact the utilization of library resources among postgraduate students in select Kenyan universities. Their study revealed that digital repositories significantly enhance accessibility to research materials, with 87.7% of respondents expressing satisfaction with repository services. This demonstrates that institutional repositories play a crucial role in improving research quality and resource management. The study utilized the Diffusion of Innovation Theory (DIT) as a framework, examining the usage patterns and challenges faced by postgraduate students. A sample of 493 respondents, including both students and librarians, was surveyed to gather qualitative and quantitative data on repository engagement. The study's findings indicated a statistically significant positive effect ($\beta = 0.109$, $p < 0.05$) of digital repository services on the utilization of library resources. Additionally, the research highlighted the necessity of consistent internet access and training workshops to optimize repository use. The authors recommend that university administrations strengthen digital repository infrastructures to further enhance accessibility and research engagement.

Quantitative Assessment of Metadata Collections of Research Data Repositories

Metadata plays a crucial role in structuring and retrieving research data within digital repositories, as analyzed by Strecker (2021). The study assessed metadata quality in research data repositories, identifying the extent to which metadata completeness and consistency affect repository effectiveness. Findings

revealed that metadata completeness varies across repositories due to differences in schema obligation levels, influencing the ease of data retrieval and reuse. Additionally, repositories with formal certifications demonstrated more comprehensive metadata records, suggesting that certification status has a more significant impact on metadata quality than repository type. The research also found that metadata descriptions become less homogeneous as their level of detail increases, indicating evolving data stewardship practices. Another key observation was that metadata records undergo frequent modifications, especially in institutional repositories, reflecting a dynamic and adaptable approach to data management. The study concluded that maintaining high metadata quality is essential for optimizing research discoverability and usability, reinforcing the importance of structured indexing in digital repositories.

Extractive Text Summarization of Online Scientific Articles for Digital Library Repository

Gunathilaka and Jahan (2023) conducted a study on extractive text summarization techniques for online scientific articles, aiming to enhance digital library repositories. The research evaluated multiple algorithms, including TF-IDF and K-Means, to determine their efficiency in summarizing large-scale academic content. The findings indicated that TF-IDF outperformed K-Means in terms of execution time, recording a processing speed of 4.338225 seconds compared to 0.891249 seconds for K-Means. Additionally, the system exhibited improved performance when the content size of the article increased. This suggests that TF-IDF remains a robust and scalable approach for extractive summarization, particularly for lengthy academic texts. The study highlights the potential of TF-IDF in streamlining information retrieval and optimizing content management in digital repositories, ensuring that users can efficiently access relevant scientific knowledge.

Research on the TF–IDF Algorithm Combined with Semantics for Automatic Extraction of Keywords from Network News Texts

Wang (2024) explored enhancements to the traditional TF-IDF algorithm by incorporating semantic analysis techniques for keyword extraction in network news texts. The study introduced a novel approach that integrated title weighting and the Word2Vec model to refine the accuracy of extracted keywords. Experimental results demonstrated that this semantic-enhanced TF-IDF method significantly outperformed existing keyword extraction techniques, achieving superior precision, recall rates, and F-measure scores. Additionally, the extracted keywords closely aligned with manually labeled keywords, reinforcing the algorithm's reliability. These findings underscore the effectiveness of augmenting TF-IDF with semantic models, particularly in improving keyword extraction accuracy and consistency. By leveraging semantic relationships between words, the enhanced TF-IDF approach provides a more sophisticated mechanism for information retrieval, making it particularly useful for applications requiring high-quality text indexing and summarization.

Best Matching Algorithm to Identify and Rank the Relevant Statutes

Automated identification of relevant documents based on queries plays a crucial role in information retrieval. Kayalvizhi et al. (2020) explored the effectiveness of the BM25 ranking algorithm in identifying and ranking relevant statutes. Their study was conducted under the Artificial Intelligence in Legal Assistance (AILA) task, where artificial intelligence techniques were applied to retrieve prior legal documents relevant to a given query. By utilizing BM25, the researchers calculated the ranking scores for various statutes and sorted them accordingly to improve retrieval efficiency. Their results demonstrated that BM25 outperformed most competing approaches, achieving a Mean Average Precision (MAP) score of 0.2975, a BPREF score of 0.2531, a reciprocal rank of 0.4769, and

a P@10 score of 0.15. These findings underscore the algorithm's capability to rank legal texts effectively, making it a suitable choice for legal document retrieval. This highlights BM25's efficiency in structured text ranking and its potential applications in other fields requiring high-precision document retrieval.

Document Ranking for COVID-19 Researcher Articles Using BM25 and BERT as a Re-ranker

BM25 remains a competitive ranking algorithm in information retrieval despite the growing popularity of deep learning-based models. Olaisen (2021) conducted a study comparing BM25 with BERT for ranking COVID-19 research documents in the CORD-19 dataset. The study explored two implementations of BERT—the bi-encoder and the cross-encoder—to determine whether deep learning could outperform traditional term-matching models like BM25. The research revealed that BM25 achieved an nDCG(10) score of 0.658, outperforming the bi-encoder, which scored 0.561. However, the cross-encoder model, used as a re-ranker, improved results by raising the nDCG(10) score to 0.713. While this demonstrates the potential of deep learning models in refining ranking results, the study also noted that replicating this approach for other datasets might be challenging. The findings reaffirm BM25's strength as an efficient baseline ranking model, particularly in cases where computational resources for deep learning models are limited or where full-ranking tasks are necessary.

Solving SEO Issues in DSpace-based Digital Repositories

Formanek (2021) explores the significance of search engine optimization (SEO) in enhancing the discoverability of academic digital repositories. Institutional repositories are widely used for the dissemination of scientific knowledge, yet many fail to optimize their visibility in academic search engines such as Google Scholar and Microsoft Academic. Through a case study of a

DSpace-based digital repository, Formanek identifies prevalent SEO deficiencies by utilizing three independent auditing tools. The study details various SEO challenges, including poor metadata structuring and inadequate indexing, which hinder accessibility. By implementing targeted optimizations, the study demonstrates a 59% improvement in SEO-related variables, significantly enhancing document visibility. Moreover, an external assessment of global DSpace repositories reveals similar SEO shortcomings, emphasizing the need for systematic improvements in metadata standardization and indexing techniques. The findings underscore the critical role of SEO in ensuring that academic repositories effectively serve their purpose as accessible, research-driven knowledge bases.

Digital Library Design for Visibility: A Case Study on University Digital Libraries

Weideman (2020) examines the accessibility of academic research documents stored in university digital libraries, focusing on their visibility in search engine results. Many digital library designers assume users navigate repositories from the homepage, yet most researchers access documents through direct search engine queries. The study evaluates doctoral theses from globally ranked universities, assessing their search engine visibility using a structured query generation method. Results indicate that 61% of the investigated theses were not listed in search results, and only 25% appeared as the top result. Weideman suggests that poor visibility stems from digital libraries failing to provide indexable rich text, such as abstracts, on their web pages. The study advocates for improving metadata structuring and ensuring research documents are easily crawlable by search engines to enhance global accessibility. These findings highlight the importance of integrating SEO principles and metadata optimization in digital libraries to maximize academic research discoverability.

Implementation of Google Cloud Platform as Data Storage in Industry

Irawan (2024) examined the adoption of Google Cloud Platform (GCP) for data storage in industries through a case study of Roche, a global pharmaceutical company. The study revealed that migrating to GCP improved data security, reduced costs, and enhanced agility. GCP's security and compliance features were key factors influencing Roche's decision, along with its scalability and flexibility. Compared to alternatives like AWS and Microsoft Azure, GCP demonstrated strong AI and machine learning capabilities, enabling advanced data processing and analysis. However, challenges such as vendor lock-in and a steeper learning curve were noted. The study concluded that while GCP offers significant benefits in scalability and security, organizations must carefully assess their needs before implementation. It also highlighted the importance of expertise in managing cloud storage systems effectively to optimize efficiency and data management outcomes.

Performance Analysis of Cloud Storage Services for Mobile Multimedia Health Record Management

Akter et al. (2020) investigated cloud storage performance for mobile multimedia health record management, comparing different cloud platforms, including Google Cloud. Their findings indicated that Google Cloud's data compression techniques reduced network traffic and improved storage efficiency. Google Cloud had the fastest upload speed (2026 seconds for 1 GB) and lower control packet usage compared to Microsoft's cloud storage services. The study analyzed cloud storage performance in terms of data synchronization, transmission efficiency, and protocol overhead. It emphasized that Google Cloud's data handling capabilities significantly enhanced performance in healthcare applications by optimizing storage and retrieval speeds. However, challenges such as data security and control data overhead were also discussed. The research concluded that while Google Cloud demonstrates superior efficiency, particularly in data compression

and transmission, further improvements in security and synchronization mechanisms are necessary for broader adoption in medical data management systems.
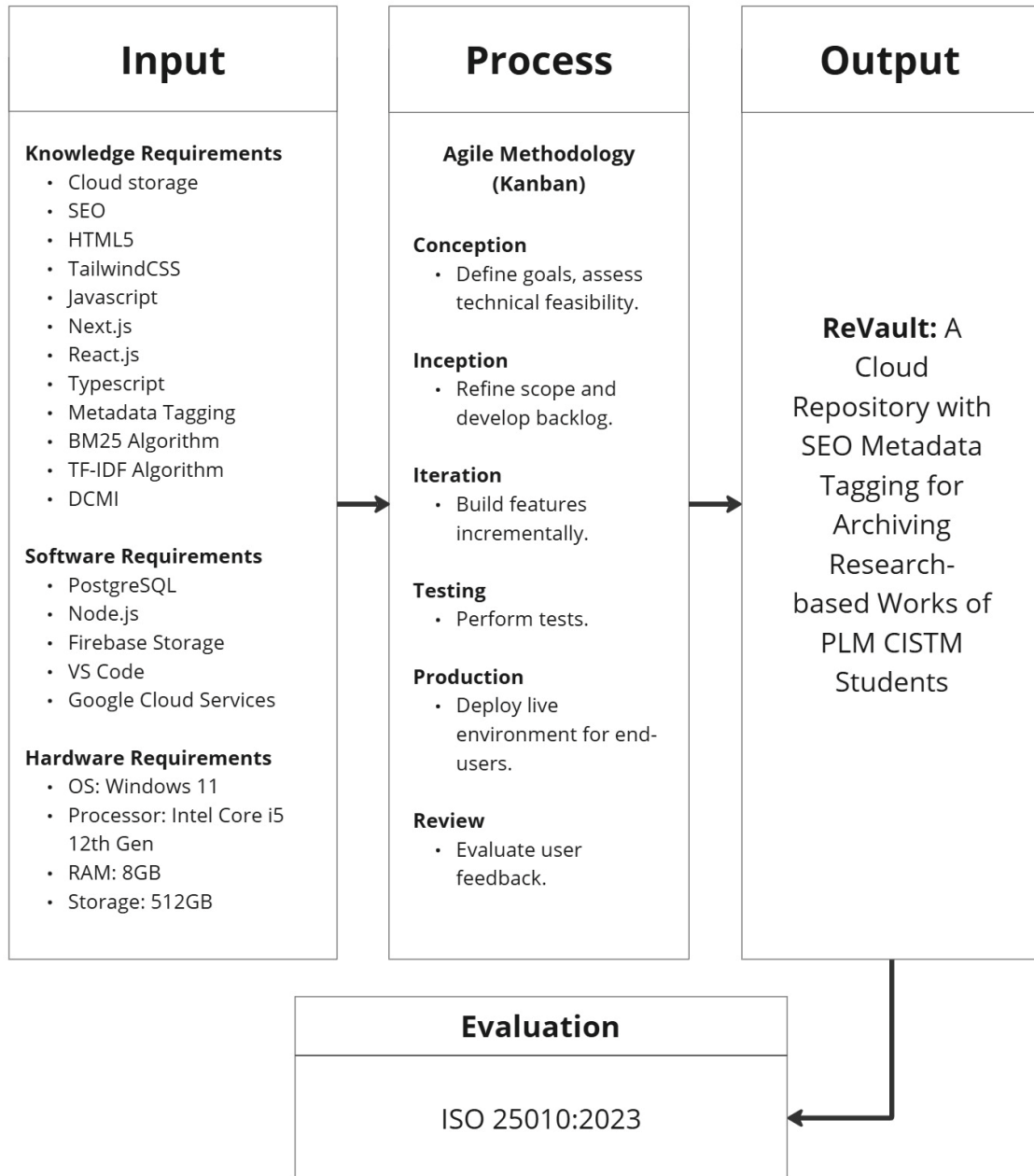
2.3 Conceptual Framework

Input	Process	Output
Knowledge Requirements • Cloud storage • SEO • HTML5 • TailwindCSS • Javascript • Next.js • React.js • Typescript • Metadata Tagging • BM25 Algorithm • TF-IDF Algorithm • DCMI **Software Requirements** • PostgreSQL • Node.js • Firebase Storage • VS Code • Google Cloud Services **Hardware Requirements** • OS: Windows 11 • Processor: Intel Core i5 12th Gen • RAM: 8GB • Storage: 512GB	**Agile Methodology (Kanban)** **Conception** • Define goals, assess technical feasibility. **Inception** • Refine scope and develop backlog. **Iteration** • Build features incrementally. **Testing** • Perform tests. **Production** • Deploy live environment for end-users. **Review** • Evaluate user feedback.	**ReVault:** A Cloud Repository with SEO Metadata Tagging for Archiving Research-based Works of PLM CISTM Students

Evaluation

ISO 25010:2023

Figure 2.1. The conceptual framework diagram for ReVault.

The development of a cloud repository for archiving research works with SEO metadata tagging requires expertise in cloud storage, SQL, Next.js, React.js, TypeScript, JavaScript, and SEO. Supported by PostgreSQL, Node.js, and Google Cloud Services for backend operations, the system demands hardware including a Windows 11 environment, Intel Core i5 processors, 8GB RAM, and 512GB storage. Following Agile-Kanban methodology, the process begins with Conception to define goals (e.g., centralized archiving) and stakeholder needs, followed by Inception to refine scope, prioritize metadata schema design, and track progress via Kanban metrics. Iteration incrementally builds features like user interfaces and version history using sprints and a Kanban board (To Do, Doing, Done), while Testing validates metadata accuracy, storage reliability, and security. Production deploys the live system with performance monitoring, and Review incorporates user feedback to optimize workflows and plan updates.

The output, ReVault, is a platform enabling structured archiving of academic works with SEO-enhanced metadata to improve accessibility. System quality aligns with ISO 25910:2023, ensuring compliance with international standards for software reliability and performance.

LIST OF REFERENCES

Abolghasemi, A., Askari, A., & Verberne, S. (2022). On the interpolation of contextualized term-based ranking with BM25 for query-by-example retrieval. ACM Digital Library. https://doi.org/10.1145/3539813.3545133

Aguilar, J., Salazar, C., Velasco, H., Monsalve-Pulido, J., & Montoya, E. (2020). Comparison and evaluation of different methods for the feature extraction from educational contents. Applied Sciences https://www.mdpi.com/2079-3197/8/2/30

Akter, M., Gani, A., Rahman, M. O., Hassan, M. M., Almogren, A., & Ahmad, S. (2020). Performance analysis of cloud storage services for mobile multimedia health record management. IEEE Access, 8, 84640-84652. https://doi.org/10.1109/ACCESS.2020.2992638

Alavez, R. G. (2022). The development of a cloud-based university research repository software using a configurable subscription model. Acta Informatica Malaysia, 6(1), 7–12. https://www.researchgate.net/profile/Reynaldo_Alvez/publication/359149731_The_Development_of_A_Cloud-Based_University_Research_Repository_Software_Using_A_Configurable_Subscription_Model/links/663df73135243041538656f5/The-Development-of-A-Cloud-Based-University-Research-Repository-Software-Using-A-Configurable-Subscription-Model.pdf

Aliazas, J. V., Cruz, R. D., & Ilagan, N. (2024). Enhancing university operations: A study of the electronic document management systems (EDMS) of one higher education institution. TWIST, 19(3), 229-237. https://twistjournal.net/twist/article/view/337

Amado, J. A., & Gamba, M. P. (2023). Assessment of enhanced online documents repository service. United International Journal for Research & Technology, 4(6), 37-41. https://uijrt.com/articles/v4/i6/UIJRTV4I60004.pdf

An, S., & Jung, J. J. (2020). A heuristic approach on metadata recommendation for search engine optimization. Concurrency and Computation: Practice and Experience, 32(10), e5407. https://doi.org/10.1002/cpe.5407

Andaya, U. D. M., Baek, J., Cruz, M. R. S., & Ng, G. (2020). An agile methodology approach to a record management and monitoring system for Company B. https://d1wqtxts1xzle7.cloudfront.net/112750420/Andaya-Baek-Cruz-libre.pdf?1711412217=&response-content-disposition=inline%3B+filename%3DAn_Agile_Methodology_Approach_to_a_Recor.pdf&Expires=1741102607&Signature=J~tWdPyymJaxRPYv9izTgHEQFHcuk~3ig0WT~jhJyyJ7alviTQnfAtr1ExeEoM4vNooDUgETGhqKp3pUqfbWjxinHfTsp8jk6oxdpjp1hacH657Ais1kouP~k-yvCRK9pebJS~Q5pnXEQu-NfkHNonqpbRytlcWMPCMPUxkmlOGBB12sTc-XaL9zm-da7~LvX6y77Q13H2vXwdzNIHHCvxOMNNPRamA9qMDjrFL3t~LosM4ptUmUWyyzRLL5eolja1Or906XJDlCggDHyY9Z1hryWzHpid1ULc9fZvlQmEhgbvCaza4tlbmb7Z3DaceEGvkLRziYFtb2466f-25iKg__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Aquino, J. F. (2024). Social media marketing strategies: Implementation, optimization, and integration among private higher education institutions. International Journal of Marketing, Communication, and Emerging Research. https://www.ijmcer.com/wp-content/uploads/2024/07/IJMCER_L0640150177.pdf

Altes, G. C., Ong, A. K. S., & German, J. D. (2024). Determining factors affecting Filipino consumers' behavioral intention to use cloud storage services: An extended technology acceptance model integrating valence framework. Heliyon, 10(4).https://www.cell.com/heliyon/fulltext/S2405-8440(24)02478-2

Augustyn, D. R., Wyciślik, Ł., & Sojka, M. (2021). PostgreSQL & Google Cloud: The cloud-enabled architecture of the clinical data repository in Poland. Sustainability, 13(24), 14050. https://doi.org/10.3390/su132414050

Bakri, M., Danish, M., & Hamiz, M. (2023). Final year project repository with automated classification using TF-IDF. IEEE Xplore. https://doi.org/10.1109/ICAC60006.2023.10284625

Banjan, K., Anilkumar, J., Singh, H., Sunny, K., & Kouser, R. (2023). Integrating public reported evidence collection, public court records archive and realizing secure and decentralized case document management using IPFS and Hyperledger Fabric blockchain: An implementation study. ResearchGate. https://www.researchgate.net/publication/371093970

Bronda-Ecraela, F. R., & Doctora, R. L. (2024). Integrating Okapi BM25 and Jaccard algorithms in thesis search engine. Journal of Information Technology and Computing, 2(1), 1-12. https://innocon.innotcs.org/index.php/jitc/article/view/59

Buctuanon, M. M., Gadiane, J. L. A., Margallo, F. A., & Lucero, P. R. (2021). Incorporating rule-based pattern recognition approach for document structure classification on cloud-based document management system. Mindanao Journal of Science and Technology. https://mjst.ustp.edu.ph/index.php/mjst/article/view/999

Caro, A. D. G., Mendoza, A. J. A., Noble, H. O., & Tanglao, K. A. B. (2024). Influence of search engine optimization (SEO) towards purchase intention of online shoppers. Journal of Digital Marketing and Consumer Behavior, 18(2). https://doi.org/10.32996/jbms.2024.6.3.24

Dohinog, D. A. (2022). Centralized documents repository. Globus-An International Journal of Management and IT, 13(2), 12-14. http://dx.doi.org/10.46360/globus.mgt.120221003

Focke, A. (2020). The Presentation and Long-Term Preservation of the Oral Histories in the Houston Asian American Archive. Transnational Asia, 3(1). https://transnationalasia.rice.edu/index.php/ta/article/view/55

Formanek, M. (2021). Solving SEO issues in DSpace-based digital repositories. Information Technology and Libraries, 40(1), 1-15. https://doi.org/10.6017/ital.v40i1.12529

Gunathilaka, R. D. R. M., & Jahan, M. A. C. A. (2023). Extractive text summarization of online scientific articles for digital library repository. Southeastern University Repository. http://ir.lib.seu.ac.lk/handle/123456789/6779

Guerrero, J., Sánchez, D., Menéndez, V., Castellanos, M. E., & Gómez, J. (2020). Tools for interoperability between repositories of digital resources. IATED Library. https://library.iated.org/view/GUERRERO2019TOO

Hovd, O. J. (2021). Document ranking for COVID-19 researcher articles using BM25 and BERT as a re-ranker [Master's thesis, NTNU]. https://ntnuopen.ntnu.no/handle/11250/2826393

Iacopini, E. (2024). Digitization of archival data and metadata in archaeology: The case of Ancona. MDPI Proceedings, 96(1), 9. https://doi.org/10.3390/proceedings2024096009

Inglis, F. (2023). Evidence Summary Theme: Information Access and Retrieval. Evidence Based Library and Information Practice, 18(1), 1-2. https://epe.bac-lac.gc.ca/100/201/300/evidence_based/pdf/2023/v18n1-2023.pdf

Irawan, A. L. H. (2024). Implementation Google Cloud Platform as data storage in industry. Journal of Research, 7(2), 469-471. https://doi.org/10.24815/jr.v7i2.37699

Kar, S., & Das, R. (2020). Publishing e-resources of digital institutional repository as linked open data: An experimental study. Library Philosophy and Practice (e-journal), 8815. https://digitalcommons.unl.edu/libphilprac/8815

Karimova, Y., Castro, J. A., & Ribeiro, C. (2020). Data deposit in a CKAN repository: A Dublin Core-based simplified workflow. Lecture Notes in Computer Science, 12051, 216-228. https://doi.org/10.1007/978-3-030-11226-4_18

Kavila, P. M., Gichohi, P. M., & Mwiti, F. M. (2024). Effect of institutional digital repository information services on utilisation of library resources by postgraduate students at selected university libraries in Kenya. University of Embu Repository. http://41.89.236.15/xmlui/bitstream/handle/1/4357

Kayalvizhi, S., Thenmozhi, D., & Aravindan, C. (2020). Best matching algorithm to identify and rank the relevant statutes. CEUR Workshop Proceedings, 2826, 1-10. https://ceur-ws.org/Vol-2826/T1-6.pdf

Koloski, B., Pollak, S., Skrlj, B., & Martinc, M. (2021). Extending neural keyword extraction with TF-IDF tagset matching. arXiv. https://arxiv.org/abs/2102.00472

Lagas, S., & Isip, J. (2023). Challenges to digital services in Philippine academic libraries. Philippine Journal of Librarianship and Information Studies, 43(1), 27–38. https://www.phjlis.org/index.php/phjlis/article/view/133

Lan, F. (2022). Research on Text Similarity Measurement Hybrid Algorithm with Term Semantic Information and TF-IDF Method. *Advances in Multimedia*, *2022*, 1–11. https://doi.org/10.1155/2022/7923262

Ledonio, A. L. (2021). Computer laboratory network centralization files management system. Development, 9(1), 64–72. https://research.lpubatangas.edu.ph/wp-content/uploads/2022/02/APJMR-2021.9.1.08.pdf

Li, J. (2021). Towards the Feasibility of Instituting a Philippine Digital Audio Library: A Case Study. Journal of ICT in Education, 8(2), 1-12.https://pdfs.semanticscholar.org/4b47/0f994b3b80d458db034452b97e317c3f5fe4.pdf '

Meidelfi, D., Yulherniwati, Rahmayuni, I., Hidayat, T., & Chandra, D. (2021). TF-IDF implementation for similarity checker on the final project title. International Journal of Advanced Studies in Computer Engineering, 5(1), 1-8. https://www.ijasce.org/index.php/IJASCE/article/view/3

Menon, R. R. K., Kaartik, J., Nambiar, E. T. K., Athul, K. T. K., & Kumar, S. A. (2020). Improving ranking in document-based search systems. IEEE Xplore. https://doi.org/10.1109/ICCMC48092.2020.9143047

Nagrama, N. D. C., Lingating, M. L. D., Calleno, J. T., Rato, R. K. A., Catungal, M. L. P., & Encarnacion, P. C. (2024). Web-based Document Management System. International Journal, 13(3). https://d1wqtxts1xzle7.cloudfront.net/115808636/ijsait031332024-libre.pdf?1717924849=&response-content-disposition=inline%3B+filename%3DWeb_based_Document_Management_System.pdf&Expires=1741334735&Signature=RrP5A5EjjAyg7TV3u0e3E5mKsOrv-ABfWru~zKSGZOmrX1mE7UdCP5mCnWLLPLETtC8f3TuYgSkwGjY7e7SUQ9n4CmAgy7wP2hnaPrtqkj89s0SjMoYutm7M~D24YFDGr5JUGK7Qjrf81b3d7jeu2-LNixutHn0XroERifXxPo3R8bzau6-lDYbfbQFoToD~i1FhN8T5V-q82ndLoD5qhEN7uKrixB3UXLMpvHAJ2Jzj-I~4Fnu~EQNfrdd0WzlR6tD0nu4z3gwaxHBXGKJPdadq-X9FkYJB8tw4yWElSZl1l2lf7odgC-P-4WcAAf6~vKVy4HHLi~CsMcxBFSmTeQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Ong, A. K. S., Altes, G. C., & German, J. D. (2024). Actual usage assessment among cloud storage consumers in the Philippines using a machine learning ensemble approach. Dental Science Reports, 14(1). https://doi.org/10.1038/s41598-024-80676-9

Or, E. M. T., & Estrellado, D. M. O. (2023). Legacy language materials in the Ernesto Constantino Collection: Challenges and lessons for building a Philippine language archive. ResearchGate.https://www.researchgate.net/publication/379822090_Legacy_Language_Materials_in_the_Ernesto_Constantino_Collection_Challenges_and_Lessons_for_Building_a_Philippine_Language_Archive

Pajas, A. L. D. M. P. (2024). Integrated web-based research document tracking system with barcode for Medical Colleges of Northern Philippines and International School of Asia and the Pacific. College of Information Technology and Engineering, 1(1). https://dawa.isap.edu.ph/index.php/cite/article/view/63

Peng, S., Xie, X., Zhai, J., Jia, Y., & Gong, Y. (2021). A page-topic relevance algorithm based on BM25 and paragraph-semantic correlation. Journal of Physics: Conference Series, 1757(1), 012115. https://doi.org/10.1088/1742-6596/1757/1/012115

Quinito, F. B., & Catipay, J. P. (2020). Technical review and analysis on students academic website projects using GTmetrix web speed and optimization tool. Journal of Information Technology Education: Research, 18. https://www.researchgate.net/profile/Floreto-Quinito/publication/360118042_Technical_Review_and_Analysis_on_Students_Academic_Website_Projects_Using_GTmetrix_Web_Speed_and_Optimization_Tool/links/6262e5108e6d637bd1f510aa/Technical-Review-and-Analysis-on-Students-Academic-Website-Projects-Using-GTmetrix-Web-Speed-and-Optimization-Tool.pdf

Reza, F., Indah, I. K. D., & Ropianto, M. (2022). Design and implementation of institutional repository with Dublin Core metadata. Jurnal KomtekInfo, 9(1), 318. https://jkomtekinfo.org/ojs/index.php/komtekinfo/article/view/318

Řezník, T., Raes, L., Stott, A., De Lathouwer, B., P erego, A., Charvát, K., & Kafka, Š. (2022). Improving the documentation and findability of data services and repositories. Computers & Geosciences, 159, 105020. https://doi.org/10.1016/j.cageo.2022.105020

Rivera, A. (2023). Organic Search Engine Optimization for Museum Websites in 2023: Strategies for Improved Online Visibility and Access (Doctoral dissertation). https://jscholarship.library.jhu.edu/server/api/core/bitstreams/fe8c2420-2d77-43a1-97a6-398eeb8c428d/content

Rodriguez III, B., Verallo, C. J. A., Aquiatan, V. T., Agpad, S. B., De Loyola, R. C., & El Jireh, P. B. (2024). eDALAYON: A Document Management and Monitoring System for the Department of the Interior and Local Government, Philippines. Philippine Journal of Science, Engineering, and Technology, 1(1), 34-41. https://pjset.org/index.php/journal/article/view/57/7

Ruguian, S. N. M., & Pola, B. A. U. (2023). Central repository system: A quantitative analysis of its feasibility and applicability in MMSU. Journal of Computer and Communications, 11(3). https://www.scirp.org/journal/paperinformation?paperid=140964

Sagun, L. J. B. D., Asuncion, J. E., & Rafols, H. C. G. (2021). File and storage management system towards cloud computing technology. Theoretical & Applied Science, (6), 101-110. https://www.researchgate.net/profile/Joel-Asuncion/publication/352968118_FILE_AND_STORAGE_MANAGEMENT_SYSTEM_TOWARDS_CLOUD_COMPUTING_TECHNOLOGY/links/613866b39520966a6b03f5db/FILE-AND-STORAGE-MANAGEMENT-SYSTEM-TOWARDS-CLOUD-COMPUTING-TECHNOLOGY.pdf?__cf_chl_tk=Dd2kgFG3HnGKVv2AwA.ez_fBGufzianz46lVdIZbq9U-1741064448-1.0.1.1-nezVkQGg7MbP8rTmPsNFsO69o2WqBQAxoAXJNJKZQhM

Santos, M., Cruz, J., & Reyes, R. (2023). Document management system using optical character recognition, clustering, watermarking, and QR coding algorithms. Journal of Information Systems and Technology, 11(7), 78-92. https://www.proquest.com/openview/581031dca1f8325f5d007334793ed39c/1?pq-origsite=gscholar&cbl=18750&diss=y

Setiawan, R., Kisman, Z., & Imam, A. (2023). Analysis of frequently appearing words in the titles of 2023 research grant winners in Indonesia using the TF-IDF method. Indonesian Scientific Journal, 1(1), 1-10. https://www.journal-isi.org/index.php/isi/article/view/599

Shibambu, A., & Ngoepe, M. (2020). When rain clouds gather: Digital curation of South African public records in the cloud. South African Journal of Information Management, 22(1), a1205. https://doi.org/10.4102/sajim.v22i1.1205

Strecker, D. (2021). Quantitative assessment of metadata collections of research data repositories [Doctoral dissertation, Humboldt-Universität zu Berlin]. https://edoc.hu-berlin.de/handle/18452/24086

Sueno, H. T., Gerardo, B. D., & Medina, R. P. (2020). Dimensionality reduction for classification of Filipino text documents based on improved Bayesian vectorization technique. International Journal of Advanced Trends in Computer Science and Engineering, 9(5), 7526–7531 https://www.researchgate.net/publication/344827951_Dimensionality_Reduction_for_Classification_of_Filipino_Text_Documents_based_on_Improved_Bayesian_Vectorization_Technique

Tinega, G. A., Mwangi, W., & Rimiru, R. (2022). Text mining in digital libraries using OKAPI BM25 model. SlideShare. https://www.slideshare.net/text-mining-bm25

Tubay, J. B. (2021). Students' use of Cloud Storage in their Studies: A Case of a Private University in the Philippines. 8(1), 16–25. https://doi.org/10.20448/JOURNAL.509.2021.81.16.25

Ventayen, R. J. M. (2020). Evaluation of state university and college institutional websites in the Philippines. Journal of Web Engineering and Technology, 16(4). https://sciresol.s3.us-east-2.amazonaws.com/IJST/Articles/2020/Issue-35/Article2.pdf

Wang, X., & Ning, H. (2020). TF-IDF keyword extraction method combining context and semantic classification. ACM Digital Library. https://doi.org/10.1145/3414274.3414492

Wang, Y. (2024). Research on the TF-IDF algorithm combined with semantics for automatic extraction of keywords from network news texts. Journal of Intelligent Systems, 33(1), 1-15. https://doi.org/10.1515/jisys-2023-0300

Weideman, M. (2020). Digital library design for visibility: A case study on university digital libraries. Proceedings of WBIMLC 2020, 36-50. https://www.researchgate.net/publication/333996189

Yap, V. A., & Velasco, R. (2024) Data Archiving for Philippines Higher Education in Compliance with Accrediting Bodies–A Cloud-Based Data Processing System for Universities. https://www.researchgate.net/profile/Rose-Mary-Velasco/publication/380002922_Data_Archiving_for_Philippines_Higher_Education_in_Compliance_with_Accrediting_Bodies_-A_Cloud-_Based_Data_Processing_System_for_Universities/links/6627208139e7641c0be2f4d1/Data-Archiving-for-Philippines-Higher-Education-in-Compliance-with-Accrediting-Bodies-A-Cloud-Based-Data-Processing-System-for-Universities.pdf

APPENDICES

APPENDIX A: REQUEST LETTERS

PAMANTASAN NG LUNGSOD NG MAYNILA
(University of the City of Manila)
Intramuros, Manila

College of Information Systems and Technology Management (CISTM)
Information Technology Department

March 18, 2025

Mr. Jesus Raymond R. Mijares
PLM Chief University Librarian
Pamantasan ng Lungsod ng Maynila
Gen. Luna Cor. Muralla Streets, Intramuros Manila, Philippines 1002

Dear Mr. Mijares:

We, third-year Bachelor of Science in Information Technology (BSIT) students from Pamantasan ng Lungsod ng Maynila (PLM), are currently undertaking our SIA project entitled **"ReVault: A Cloud Repository with SEO Metadata Tagging for Archiving Research-based Works of PLM CISTM Students."** This system is designed to address the challenges of research preservation, retrieval, and accessibility within the College of Information Systems and Technology Management (CISTM). The project serves as a final requirement for our System Integration and Architecture (SIA) course.

Furthermore, we request authorization to access any available digital manuscripts and hardbound copies of undergraduate research works from the past five years to structure our system accordingly. We fully acknowledge the sensitivity of these materials and will strictly adhere to all institutional policies regarding confidentiality and access. Should full direct access not be possible, we would appreciate your approval to seek copies directly from the authors, following appropriate guidelines set by the college.

We understand the importance of institutional procedures and data security, and we are fully committed to ensuring compliance with all necessary protocols. If there are any additional requirements we must fulfill to facilitate this request, we are more than willing to comply.

We deeply appreciate your time and consideration of this matter. If there are matters that you want to discuss with our subject adviser, please contact him through this email, dswabando@plm.edu.ph.

Respectfully,

KC D Durante

Matthew Jacob B. Insigne

Kristhia Cayle F. Lastra

John Allen Troy E. Valeña

Jetthro Karl C. Yacub

Noted by:

Prof. Diony S. Abando
SIA Adviser

Signed by:

Mr. Jesus Raymond R. Mijares
PLM Chief University Librarian

PAMANTASAN NG LUNGSOD NG MAYNILA
(*University of the City of Manila*)
Intramuros, Manila

College of Information Systems and Technology Management (CISTM)
Information Technology Department

March 12, 2025

Prof. Ariel Antwaun Rolando C. Sison
Chairperson, Information Technology Department
Pamantasan ng Lungsod ng Maynila
Gen. Luna Cor. Muralla Streets, Intramuros Manila, Philippines 1002

Dear Mr. Sison:

We are third-year Bachelor of Science in Information Technology (BSIT) students from Pamantasan ng Lungsod ng Maynila (PLM), currently working on our project titled **"ReVault: A Cloud Repository with SEO Metadata Tagging for Archiving Research-based Works of PLM CISTM Students."** This system is designed to address the challenges of research preservation, retrieval, and accessibility within the College of Information Systems and Technology Management (CISTM). The project serves as a final requirement for our System Integration and Architecture (SIA) course.

As part of our research, we are requesting your permission to conduct an interview regarding the current archival process of Information Technology research-based works, particularly from courses such as System Integration and Architecture and Capstone Project. Understanding the department's archiving practices will allow us to design an effective system that preserves and improves access to academic research.
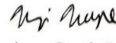
Additionally, we seek authorization to access digital and hardbound copies of undergraduate Information Technology research works within the past five years. We understand the sensitivity of these materials and will adhere to all necessary confidentiality and access policies. If full access to digital copies is not possible, we would also like your approval to request copies directly from the authors.

We would be grateful for any guidance you can provide regarding these matters. We are once again anticipating your cooperation with this request and giving you our utmost gratitude. If there are matters that you want to discuss with our subject adviser, please contact him through this email, dswabando@plm.edu.ph.

Thank you for very much.
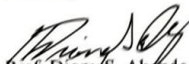
Respectfully,

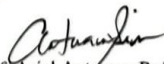KC D. Durante
Lastra

Matthew Jacob B. Insigne

Krishia Cayle F.

John Allen Troy E. Valeña

Jetthro Karl C. Yacub

Noted by:

Prof. Diony S. Abando
SIA Adviser

Signed by:

Prof. Ariel Antwaun Rolando C. Sison
Chairperson, Information Technology Department

APPENDIX B: TRANSCRIPT OF INTERVIEW WITH
UNIVERSITY LIBRARIAN

Transcript of Interview

Meeting for SIA: Interview with Library Office - Transcript of Dialogue

Date: March 25, 2025; Tuesday

Location: PLM University Library

Participants of the Meeting:

Mr. Jesus Raymond R. Mijares, PLM Chief University Librarian

Matthew Jacob B. Insigne

John Allen Troy E. Valeña

Jetthro Karl C. Yacub

Proceedings

[The meeting started at 12:30 PM]

Questions About the General Archiving Practices

Q: How are IT and Computer Science research papers currently stored and archived in the department?

Mr. Mijares: Currently, we don't have any [archived research] papers stored in the library, because we are not yet [into] digitization of thesis [works].

Q: What challenges do students and faculty face when trying to retrieve past research works?

Mr. Mijares: So usually… uhh… to retrieve the data when they go to search uh… sometimes it's [the] internet connection, that's the number one thing because if you already install the-- your data, it's very easy to access [the paper/data], however retrieving the data it depends on the-- how the-- how you are going to input the data. Usually if you're trying

to see the whole thesis-- perhaps but with our guideline for example only the summary, the abstract, okaya… the first 3 chapters only. 'Yun lang 'yung ma-ilalagay namin but we cannot put all the thesis. (Secretary of Mr. Mijares: the full text) The full conte-- the full text, oo. 'Yun lang yung ano- yung challenges diyan… so syempre ineexpect niyo 'no, na makikita niyo buo [yung paper], so hindi naman. Kasi tinitignan natin yung data privacy.

Questions About Challenges in Research Accessibility & Organization

Q: How does the department ensure that research papers are well-organized and easy to find?

Mr. Mijares: Naka classify. We have the-- in librarian point of view, naka metadata yan. Ano ibig sabihin ng metadata? Kapag sinearch mo halimbawa ay 'management', keyword 'management', tapos yung thesis mo may title na 'management' lalabas 'yun, so masasama siya—E-books, Printed books, E-Journals, Journals, plus Thesis. Nakikita yun [doon]. So madali siyang ma-ano… madali siyang… ma-search. So ayun yung kaganadahan niyan pag organized.

Q: Have there been instances where students or faculty struggled to locate relevant past studies? Can you describe what happened?

Mr. Mijares: Depende 'no kung hahanapin nyo [halimbawa paper na nagsisimula sa letter] E, napakadali [hanapin] pero yung location for the printed thesis, yung iba na mis-shelved ibig sabihin… supposedly dito lang siya [sa isang section] minsan nalalagay dito [sa kabilang section], so kung sa printed merong [struggle sa paghahanap] pero sa online hindi mangyayari 'yan kasi nga meron tayong- sabi niyo nga sa question number… 3 kung organized, definitely hindi ka maliligaw. Ganun din sa printed, kung very-well organized ang pagkaka-shelved ng thesis hindi mahihirapan ang kliyente.

Questions About the Issues with Traditional Storage & Loss of Research

Q: Have there been cases where research papers were lost, misplaced, or became inaccessible? What were the causes?

Mr. Mijares: Sa'king experience 'no uhm… mis-placed (mis-shelved) pwede pa pero yung lost… hindi ko pa na-experience 'no… kasi kapag sinabi mong lost, dalawa lang maiisip naming (laugh)… maaring (may nagnakaw) may kumuha—ninakaw yung thesis 'no, nang walang paalam, at tapos yung-- o kaya ay-- hindi mo masabing tinago kasi kapag tinago, misplaced lang 'yun so kaya siya inaccesible… o kaya mis-shelved ang tawag namin diyan, hindi siya naka shelved sa tamang ayos. Yun lang.

Q: How does the department address situations where research papers need to be retrieved but aren't properly stored or recorded?

Mr. Mijares: Kung kulang ka sa documents-- so that would be the challenges kasi kaya hindi siya na-properly stored maaring kulang yung documents na sinubmit [ng studyante]… kasi may mga ganiyang instances. Halimbawa, uhm… dapat 'pag sinubmit mo sa amin ang paper mo dapat talagang pasado kana. May iba kasi dumadaan sa proseso… na andyan yung paper nila 'no, minsan kasama yung librarian sa pumipirma… sa ibang school ganon eh 'no (Secretary: Opo)… kasama yung librarian sa pag-- siya yung final uh… magsasabi na pwede na'to… okay na 'to. Ngayon, [maaring] hindi ka pa niya io-okay kasi meron ka pang kulang na kailangan mong isubmit, dapat idaan mo muna sa isang department na 'to ganiyan…may ibang ganon. Pero sa PLM since wala pa rin naman tayong digitization pa, sa ngayon hindi pa 'to nangyayari.

(Secretary of Mr. Mijares: Yung iba sir kulang… hindi nag susubmit ng electronic format.)

Mr. Mijares: Ayun pa.. maaring isa lang sinubmit nyo dapat usually alam ko dapat dalawa.

Questions About Updating and Maintaining Research Versions

Q: When research work is updated or revised, how is that tracked in the current system?

Mr. Mijares: Etong uh… siguro revise lang ang sasabihin ko 'no... once we already submit-- or the documents 'no, yung thesis niyo nailagay na namin sa metadata…usually very minimal na lang yung revision, kasi the thesis that you-- the moment that you submitted the [paper, or the] department [submitted the] thesis to us, kompleto na 'yon ibig sabihin yung ginawa niyong thesis approved na. Ipapasa niyo na lang- ipapasa nalang nila

sa amin, kumpleto na. So…wala akong masyadong nakikitang ano dito uhm… uh challenges, oo—unless na lang yung cataloger namin mali ang nailagay niya na subject so ayun lang naman 'yon…but totally wala na.

Q: What challenges do you see in maintaining updated versions of research works while keeping past versions accessible?

Mr. Mijares: Unless may gumawa ng bagong thesis nyo 'no. Kasi diba sa last part of your thesis may further recommendation kayo 'no? So, ang kagandahan non—merong gumagamit- I mean, merong tumingin ng thesis niyo, tapos nakita niya yung recommendation ninyo—yung recommendation niyo, 'yun ang ginawa niyang thesis. So, makikita mo na nag u-update-- upgrade 'no. Ikalawa, kung meron tayong uh… tinatawag na Scopus 'no, ito yung tools na dun mo makikita kung maraming mga researchers ang tumitingin or cited sa iyong thesis. So ayun… sa ngayon wala pa tayong Scopus pero yung updated [na paper] nakikita naman namin kasi same subject lang naman yung ilalagay tsaka basically kung ano yung title mo- 'di naman nag-kakalayo yung title... nag-iba lang ang title pero the subject is [still] there. So… ganon lang 'yun.

[The meeting adjourned at 1:00 PM]

APPENDIX C: TRANSCRIPT OF CONVERSATION WITH DEPARTMENT CHAIRPERSON

Transcript

Date: March 28, 2025; Monday

Location: Facebook Messenger

Participants:

Prof. Ariel Antwaun Rolando C. Sison, Chairperson of IT Department

Jetthro Karl C. Yacub

Questions About the General Archiving Practices

[2:00 PM] **Q:** How are IT and Computer Science research papers currently stored and archived in the department?

[2:02 PM] **Mr. Sison:** Currently naka stored sa library yung mga thesis.

[2:05 PM] **Q:** What challenges do students and faculty face when trying to retrieve past research works?

[2:08 PM] **Mr. Sison:** Nahihirapan maghanap ng mga lumang thesis. Dahil minsan wala nang kopya sa department at library.

Questions About Challenges in Research Accessibility & Organization

[2:12 PM] **Q:** How does the department ensure that research papers are well-organized and easy to find?

[2:20 PM] **Mr. Sison:** Requirements sa graduating students na mag pasa ng 2 kopya ng thesis. Para sa library at department.

[2:22 PM] **Q:** Have there been instances where students or faculty struggled to locate relevant past studies? Can you describe what happened?

[2:25 PM] **Mr. Sison:** Maraming beses may naghahanap ng capstone kaso hindi mahanap kung saan nakalagay.

Questions About Issues with Traditional Storage & Loss of Research

[2:27 PM] **Q:** Have there been cases where research papers were lost, misplaced, or became inaccessible? What were the causes?

[2:30 PM] **Mr. Sison:** Same sa answer sa number 5 may naghahanap pero di makita.

[2:32 PM] **Q:** How does the department address situations where research papers need to be retrieved but aren't properly stored or recorded?

[2:35 PM] **Mr. Sison:** Sa ngayon inuutos ni dean humingi ng softcopy ng research at i-stored online.