DEVELOPMENT OF A WEB-BASED DOCUMENT REPOSITORY WITH PLAGIARISM CHECKER

Adegbola Adesoji[1], Akande Oyebola[2], Tunde-Idowu Inioluwa[3], Adebanjo Adedoyin[4],

Adewuyi Oluwaseyi[5], Mgbeahuruike Emmanuel[6], and Adediran Oluwaseyi[7]

[1,4,6]Department of Software Engineering, Babcock University, Ilisan, Nigeria.

[2,5]Department of Computer Science, Babcock University, Ilisan, Nigeria.

[7]Department of Information Technology, Babcock University, Ilisan, Nigeria.

Emails:

[1]adegbolaa@babcock.edu.ng, [2]akandeo@babcock.edu.ng;

[3]oluwatunmininu29603@gmail.com; [4]adebanjoa@babcock.edu.ng;

[5]adewuyio@babcock.edu.ng; [6]mgbeahuruikee@babcock.edu.ng; [7]adedirana@babcock.edu.ng

ABSTRACT: *In the digital age, managing vast volumes of documents and ensuring the originality of content has become a significant challenge. This paper presents the development of a web-based document repository integrated with a plagiarism checker, aimed at providing a comprehensive solution for storing, retrieving, and verifying the uniqueness of documents. The system allows users to upload, organise, and search documents efficiently while employing a robust plagiarism detection mechanism to ensure the integrity of content. By leveraging web technologies and plagiarism detection algorithms, this platform serves as a valuable tool for educational institutions, businesses, and content creators. The system enhances document management practices by offering a centralised, secure repository and reducing the risk of intellectual property infringement. This paper discusses the architecture, features, and implementation challenges of the system, along with its potential applications in various domains.*

KEYWORDS: Document, Repository, Plagiarism, Web-based.

INTRODUCTION

In the digital age, the management and integrity of information are critical for organisations, educational institutions, and businesses alike. The increasing reliance on digital documents for communication, research, and record-keeping has led to a growing need for efficient and secure document storage solutions (Huda, 2022; Iguehi et al., 2016). At the same time, with the vast availability of online content, the challenge of ensuring originality in written work has become more pronounced. This has driven the need for tools that can both manage large volumes of documents and detect instances of plagiarism.

The development of a web-based document repository with a plagiarism checker addresses these dual concerns by offering a centralised platform for storing, organising, and accessing documents, while simultaneously ensuring their authenticity. A web-based system allows users to upload, categorise, and retrieve documents from any location, providing flexibility and convenience. Meanwhile, the integration of a plagiarism checker adds a layer of security, allowing users to verify the originality of their documents and preventing the unauthorised use of existing works.

This system is particularly relevant in academic and corporate settings, where maintaining document integrity is paramount. Educational institutions can use such platforms to store research papers, theses, and assignments while checking for academic dishonesty. Similarly, businesses can store legal contracts, reports, and other sensitive materials, ensuring that these documents remain both secure and original.

The proposed system aims to streamline document management processes and safeguard the originality of content by leveraging cloud-based storage and advanced plagiarism detection algorithms. In this context, the web-based document repository will serve as a comprehensive solution for users seeking both efficient document management and content authenticity validation.

This paper outlines the design, development, and implementation of a web-based document repository with an integrated plagiarism checker. It explores the system architecture, key features, and technologies used, while also addressing the potential challenges and solutions associated with its development. The goal is to provide a scalable, secure, and user-friendly platform that can meet the evolving needs of document management and content verification in various fields.

Statement of the Problem

In academic institutions, businesses, and other professional settings, the effective management, storage, and retrieval of documents is essential for ensuring productivity and collaboration. However, as the volume of digital content continues to grow, so does the challenge of organising these documents in a systematic, secure, and easily accessible manner. In addition, the need to maintain originality and intellectual integrity in document submissions is critical, particularly in educational environments, where plagiarism has become a widespread issue.

Current systems often fail to combine the dual functionality of a comprehensive document repository with a robust plagiarism detection mechanism. This results in disjointed processes where users must navigate multiple platforms to store documents and separately check for

content originality. Such inefficiencies lead to time wastage, inconsistent document management practices, and the risk of intellectual property infringement.

There is a need for an integrated solution that not only provides secure and efficient document storage but also includes a built-in plagiarism detection tool to ensure the originality of the content. This system would streamline workflows, enhance document management, and improve the verification of document authenticity.

Thus, the development of a web-based document repository with a plagiarism checker aims to address these challenges by providing a centralised platform that simplifies document management and upholds content integrity. This system will cater to educational institutions, businesses, and any organisation that requires a secure repository for document storage along with the ability to detect and prevent plagiarism efficiently.

Aim and Objectives

This project is aimed at developing a web-based document repository that provides users with a secure platform for storing, organising, and sharing documents while incorporating an integrated plagiarism checker to ensure content originality and integrity.

The following are the objectives to achieve the aim –

i.   Create a robust and scalable architecture for the web-based repository that supports efficient document storage, retrieval, and user management.

ii.   Research, select, and integrate an effective plagiarism detection tool that allows users to check their documents for originality before submission.

iii. Establish security protocols to protect user data and documents, ensuring compliance with data protection regulations and safeguarding against unauthorised access.

iv.   Design and implement an intuitive user interface that facilitates easy navigation, document upload, and plagiarism checking.

v.   Perform rigorous testing of the repository and plagiarism checker to identify and resolve any issues, ensuring a reliable and user-friendly experience.

LITERATURE REVIEW

Document repositories are essential in the digital landscape for efficiently storing, managing, and retrieving information. Initially, repositories were offline systems primarily used for personal or institutional purposes, but the advent of the Internet has shifted these platforms to web-based solutions. Web-based repositories, like institutional repositories or cloud-based systems (e.g., Google Drive, Dropbox), provide access to information anytime and anywhere. These repositories serve not only as storage solutions but also as collaborative tools that support versioning, sharing, and metadata tagging for easier retrieval (Bailey, 2008). Educational institutions use such repositories to preserve academic documents, including research papers, theses, and dissertations, ensuring long-term access and organisation (Bankier and Smith, 2010).

The growth of academic document repositories is also driven by the need for open access and transparency. Many universities have moved towards institutional repositories to share research outputs publicly, allowing for wider dissemination and access. However, the key challenge remains the need for efficient search mechanisms and management of digital assets across diverse types of documents, formats, and sources (Lynch, 2003).

Plagiarism detection software has become a crucial tool in academic and professional writing, addressing ethical issues related to originality and intellectual property. The rise of digital content and the ease of copying and pasting from the web have led to an increase in plagiarism cases, making automated plagiarism checkers essential for maintaining academic integrity. These tools, such as Turnitin, Grammarly, and Copyscape, compare documents against vast databases of web content, academic papers, and other documents to detect similarities (Suseela, 2016).

Plagiarism detection algorithms typically employ two primary techniques: fingerprinting and string matching (Nkue and Taylor, 2021). Fingerprinting divides a document into small chunks and creates unique identifiers (hashes), which are then compared against other documents. String matching, on the other hand, identifies and compares sequences of words to detect common patterns. Both methods have their strengths and weaknesses, but recent advancements in natural language processing (NLP) are making these tools more sophisticated, capable of identifying not just direct copying but also paraphrasing and partial plagiarism (Roland and Suharjito, 2014).

A web-based document repository integrated with a plagiarism checker combines the benefits of both technologies. Users can upload, store, and share documents while simultaneously ensuring their originality through automatic plagiarism checks. This integration improves workflow efficiency, as users do not need to rely on separate platforms for plagiarism detection and document storage. Research on such integrated systems emphasizes their usefulness for academic institutions, where educators can assess student submissions for originality while maintaining a structured repository of academic work (Dharmapalan, 2017).

The seamless integration of these technologies poses certain challenges, particularly related to data privacy and processing speed. Since plagiarism checkers often require access to large databases, ensuring that documents within the repository are protected from unauthorised access becomes crucial. Moreover, processing large volumes of data for plagiarism checking can lead to performance issues, requiring efficient algorithmic solutions and cloud computing to handle high loads (Barreiro et al., 2019).

Cloud computing has transformed the way document repositories and plagiarism checkers operate. With the scalability and flexibility provided by cloud services, web-based systems can now handle large amounts of data while offering real-time access and processing. This shift allows institutions and individuals to manage document storage efficiently, eliminating the need for on-premise hardware and maintenance (Armbrust et al., 2010).

Recent literature highlights the benefits of cloud-based plagiarism checkers, particularly in terms of resource efficiency and scalability. Tools such as Urkund and Copyleaks offer cloud-based services that provide real-time plagiarism detection by comparing documents across various sources, including web content, publications, and previously submitted academic papers (Belli et al, 2020). Such systems ensure that repositories are both secure and scalable, catering to the growing need for large-scale document management and originality verification.

Another significant challenge is ensuring accuracy in plagiarism detection, particularly in detecting paraphrased content or translated works. Current algorithms primarily focus on exact matches, and the detection of sophisticated plagiarism techniques remains limited. Moreover, balancing the speed and accuracy of plagiarism checks in large-scale document repositories requires highly optimized algorithms and considerable computing power (Abdel-Hamid & Abdulghani, 2021).

Lastly, user experience and interface design are essential for the successful adoption of such systems. The literature points to the importance of intuitive design, ease of use, and seamless integration with existing tools, such as Learning Management Systems (LMS) or Content Management Systems (CMS), to ensure that users can efficiently manage and check documents for plagiarism within a single platform (Alqassimi, Saleh, & Ali, 2019).

The development of a web-based document repository integrated with a plagiarism checker is a timely response to the needs of modern academia and professional environments. While document repositories have long played a vital role in managing digital content, the integration of plagiarism detection systems addresses the growing demand for tools that ensure content originality. Cloud-based solutions provide a scalable infrastructure for these integrated systems, though challenges related to privacy, algorithm accuracy, and usability must be overcome to ensure widespread adoption and success. The combination of these technologies holds the potential to streamline document management while upholding ethical standards in content creation.

METHODOLOGY

In achieving the goal of the project, each of the objectives was carefully considered and executed.
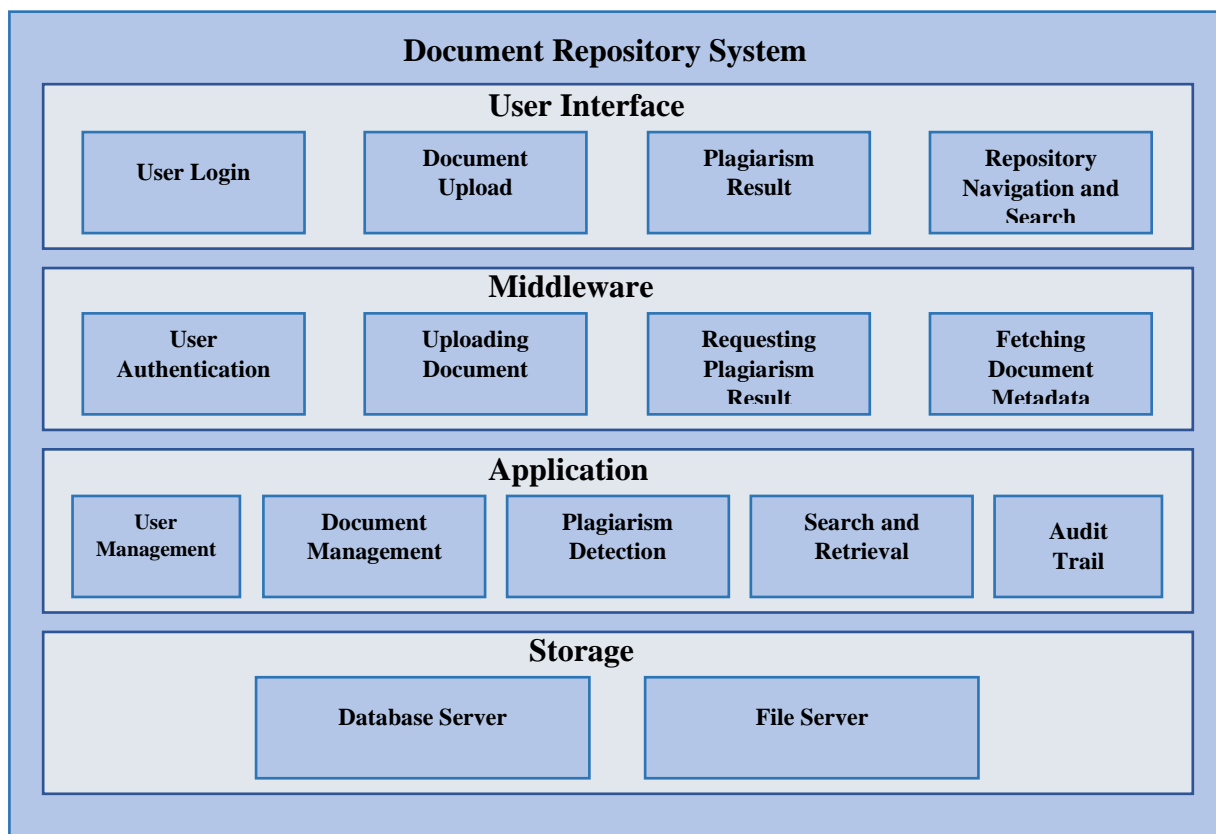
Figure SEQ Figure * ARABIC 1: Software Architectural Design for the Document Repository System

From Figure 1 above, the users interact directly with the system through the interface. It contains components for Document Upload, Repository Navigation/Search, Results of Plagiarism Checking, and user login.

The middleware layer provides communication between the frontend and backend for Uploading documents, fetching document metadata, Requesting plagiarism check results, User Authentication and Session Management. Web Application layer handles all business logic and processing. It consists of user management, document management, plagiarism detection, search and retrieval and audit trail. The storage layer stores the uploaded documents, indexed for efficient retrieval and search.

This architecture ensures scalability and efficiency in handling document storage and plagiarism detection.

Integrating Project Repository with Plagiarism Checker

To enhance the plagiarism detection functionality of the web-based digital project repository, a third-party plagiarism checker, Unicheck, was integrated into the system. Unicheck is a widely recognised and reliable plagiarism detection tool known for its accuracy and efficiency in identifying instances of plagiarism in academic works.

The integration process involved establishing a secure connection between the web-based repository and Unicheck's API (Application Programming Interface). This API allowed seamless communication and data exchange between the two systems. By leveraging Unicheck's robust plagiarism detection algorithms and databases, the system was able to compare uploaded documents against a vast repository of scholarly materials.

When a user uploads a project document or journal article to the repository, the integrated plagiarism checker automatically initiates a plagiarism check. The document is then processed and compared against Unicheck's extensive database, which includes academic papers, publications, and other relevant sources. The system generates a detailed report indicating any instances of potential plagiarism, highlighting the matched sources and providing a similarity score which is then sent to the user's email.

The integration with Unicheck offers several advantages. Firstly, it significantly reduces the burden on individuals who would otherwise have to manually compare documents for plagiarism. The automated process saves time and effort, enabling faster and more thorough plagiarism checks.

Furthermore, Unicheck's advanced algorithms ensure a high level of accuracy in detecting plagiarism. The system employs sophisticated text-matching techniques, including lexical analysis, semantic analysis, and citation analysis, to identify similarities and potential instances of plagiarism. This comprehensive approach helps maintain academic integrity and discourages academic misconduct.

Webhook-based Integration Approach

To seamlessly communicate with Unicheck and execute plagiarism checks within the web-based digital project repository, the system implemented a webhook-based integration approach. Webhooks serve as a mechanism for real-time communication between systems, allowing for immediate data transmission and event-triggered actions.

The integration process involved establishing a webhook connection between the web-based repository and Unicheck's plagiarism detection system. A webhook endpoint was set up within the repository's backend infrastructure to receive and process plagiarism check requests. When a user uploads a project document or journal article for plagiarism analysis, the web-based repository initiates a webhook request to Unicheck. This request contains relevant information about the document, such as its content and metadata. The webhook endpoint at Unicheck receives the request and triggers the plagiarism check process.
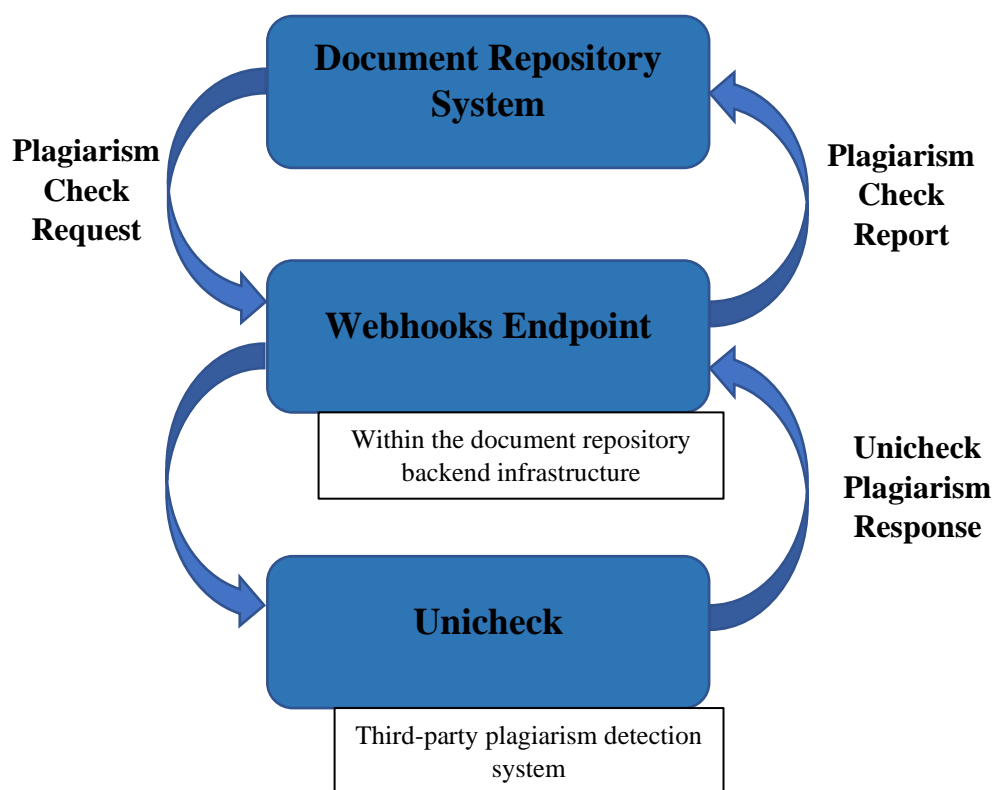
Figure SEQ Figure * ARABIC 2:System Integration with Unicheck

Figure 2: System Integration with Unicheck

Upon receiving the request, Unicheck's system applies its advanced plagiarism detection algorithms to analyse the document's content. The webhook integration enables the web-based repository to receive the plagiarism check report from Unicheck in real time. The report is then processed and made available to the user through the repository's user interface and the user's email. The detailed report provides information on matched sources, highlighting the specific sections or passages that may require further examination.

Implementation Tools

During the implementation of this project, several tools and technologies were utilised. The web components were developed using HTML5, Tailwind CSS, and Django framework. For the database, PostgreSQL was utilised with the assistance of Django's Object-relational mapping capabilities. In order to store static files, Amazon's S3 bucket was employed. To achieve the plagiarism detection functionality, Unicheck, a reliable plagiarism checker, was integrated into the project.

RESULTS

Landing Page

The landing page of the web-based digital document repository serves as the entry point for users. It provides a visually appealing and intuitive interface that welcomes users and highlights the key features and benefits of the system. Clear instructions and user-friendly forms are implemented to collect necessary information from users during registration, allowing them to create their accounts and gain access to the repository's features.
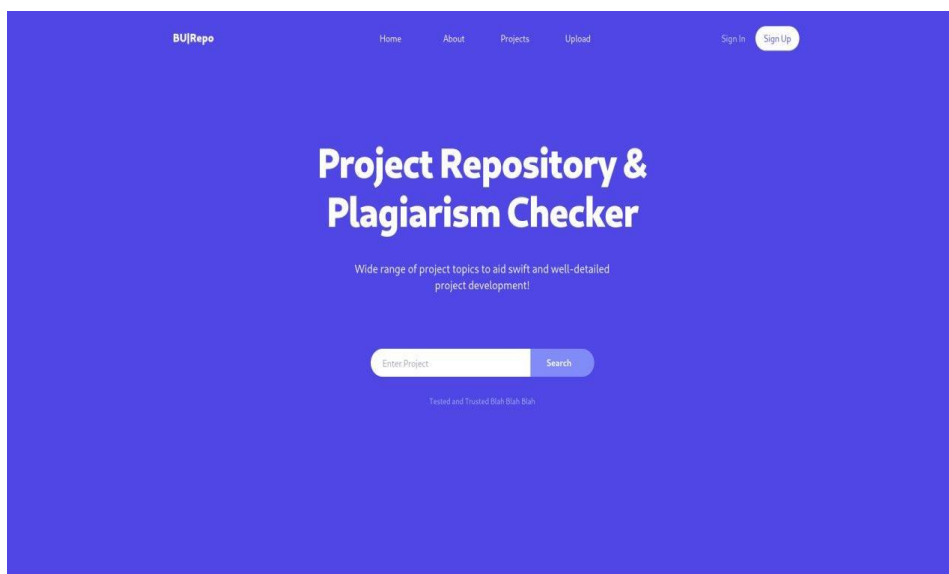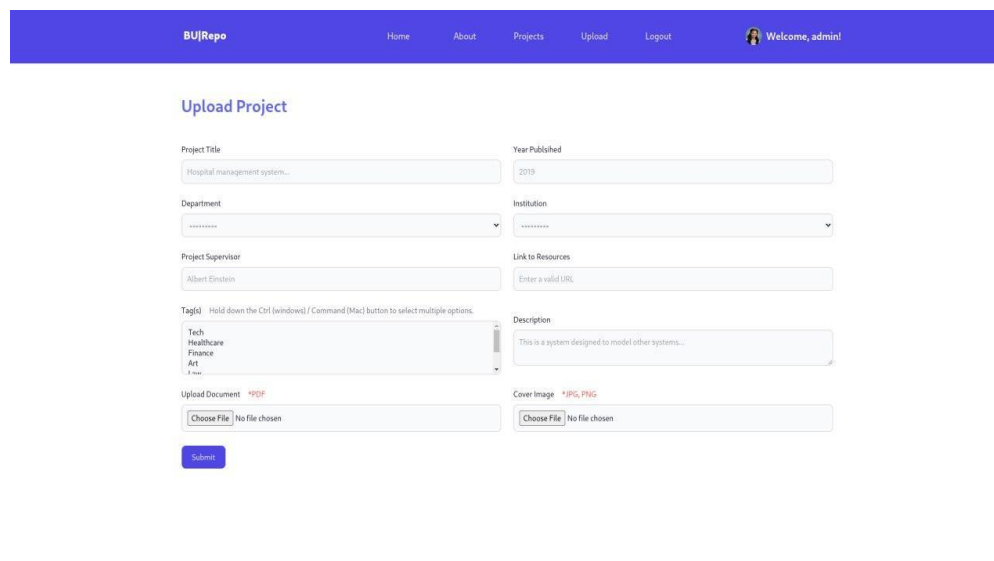
Figure 3: Document Repository System Landing Page
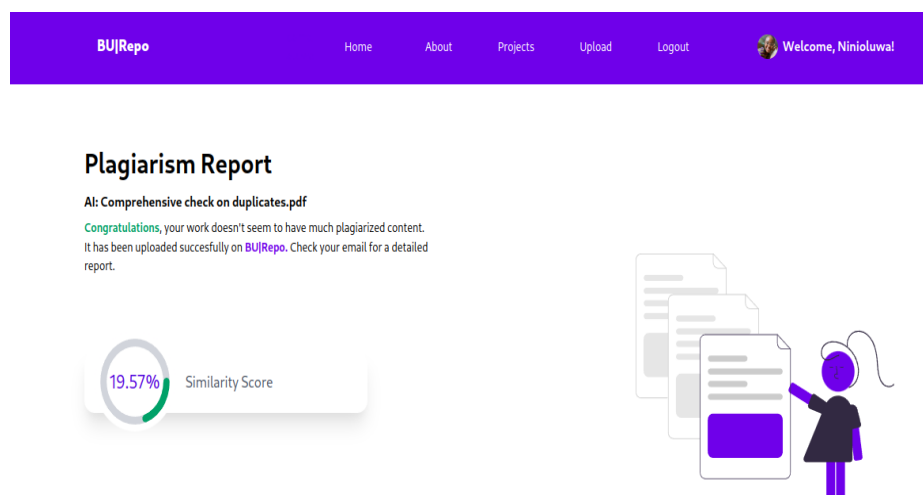
Document Registration and Plagiarism Check

Once registered and logged in, users can easily navigate to the document submission section. Here, they can provide detailed information about their document, such as the title, author(s), and description. Upon submitting the document, the system automatically triggers a plagiarism check process by sending a request to Unicheck through the webhook integration. Unicheck's plagiarism detection algorithms then analyse the uploaded document, comparing it against an extensive database of scholarly materials.

After the plagiarism check is complete, the web-based repository receives the plagiarism check report via the webhook. The system sends a comprehensive summary that includes information on matched sources, similarity scores, and highlighted sections for further examination. To ensure effective communication, the system sends an email to the user, notifying them that the plagiarism check is complete and providing detailed information about the results. This email notification keeps users informed and enables them to promptly review the results of the plagiarism check.

Figure 4: Document Registration Page

Figure 5: Plagiarism Report Page

By leveraging the integration with Unicheck, the system goes beyond traditional plagiarism checks limited to online databases and extends its functionality to documents stored within the repository itself. This expanded capability is demonstrated in Figure 7, where a 100% match is indicated. This match occurs when the exact document has previously been uploaded and stored in the repository, resulting in a perfect similarity score. The system's ability to perform plagiarism checks on repository-stored documents enhances its effectiveness and provides users with a comprehensive and thorough analysis of potential plagiarism instances, ensuring the highest level of accuracy in detecting similarities within the repository's collection of materials.
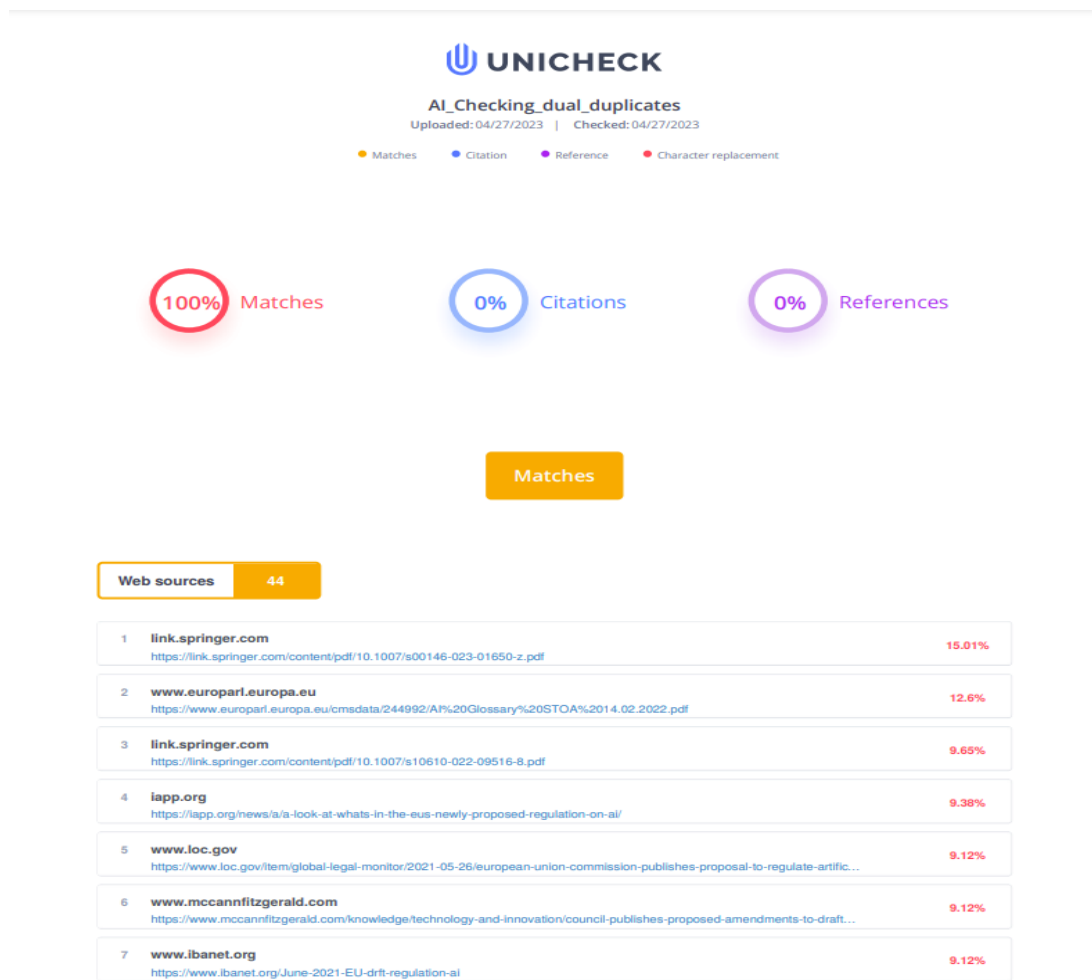
Figure SEQ Figure * ARABIC 6 Detailed Plagiarism Report

View and Search Project

The web-based repository provides users with convenient options to view and search for projects. Users can browse through a list of projects or utilise search filters to narrow down their search based on criteria such as title, author, or keywords. The system presents project details in a meticulously organised format, encompassing essential information such as the project title, author(s), description, and any associated files or resources. Notably, the system prominently displays the plagiarism score, a key metric in assessing the originality of the project. This score is visually highlighted in green if it falls below 25%, indicating a low similarity level, while it is marked in red if it exceeds 25%, indicating a higher degree of similarity. By effectively visualising the plagiarism score, users can quickly gauge the originality of a project and make informed decisions based on the extent of similarity detected. Users can access the full project document and explore additional materials related to the project. This comprehensive view and search functionality enables users to easily discover and access a wide range of research projects within the repository.

Figure 7: Search Result Page

CONCLUSION

The development of a web-based document repository integrated with a plagiarism checker provides a significant solution to document management and academic integrity challenges. This system offers a centralised platform for storing, organising, and retrieving documents while ensuring the originality of content through plagiarism detection. By utilising modern web technologies and plagiarism detection algorithms, the system enhances document accessibility, improves collaboration, and promotes ethical academic practices. It reduces the manual effort required for plagiarism detection and streamlines the workflow of managing academic or professional documents in an efficient, user-friendly manner.

FUTURE RESEARCH

Future versions of the system could include more advanced collaboration tools, document version control, and integration with learning management systems (LMS) to broaden its application in academic environments.

REFERENCES

● Huda, M. (2022), Empowering professional and ethical balance in digital record management, Organisational Cybersecurity Journal: Practice, Process and People, Vol. 2 No. 1, pp. 60-73. https://doi.org/10.1108/OCJ-06-2021-0016

● Iguehi Joy Ikenwe, Osahon Magnus Igbinovia, Ademakhe Ann Elogie (2016) - Information Security in the Digital Age: The Case of Developing Countries. Chinese Librarianship: an International Electronic Journal, 42. URL: http://www.iclc.us/cliej/cl42IIE.pdf

● Charles W. Bailey, Jr., (2005), The Role of Reference Librarians in Institutional Repositories, Reference Services Review 33, no. 3 (2005): 259-267. http://dx.doi.org/10.1108/00907320510611294.

● Jean-Gabriel Bankier & Courtney Smith (2010), Repository Collection Policies: Is a Liberal and Inclusive Policy Helpful or Harmful? Australian Academic & Research Libraries, 41:4, 245-259, DOI: 10.1080/00048623.2010.10721479

● Lynch, C. (2003) Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL Bimonthly Report, 226. https://doi.org/10.1353/pla.2003.0039

● V.J. Suseela (2016), Plagiarism: The Academic Dishonesty the Significance of Anti-Plagiarism Software (Tools) in Plagiarism Detection. Pearl A Journal of Library and Information Science 10(1):11 DOI:10.5958/0975-6922.2016.00002.4

● Nkue Dumka, Taylor Onate (2021), Plagiarism Detection System using an Enhanced String Matching Algorithm, Journal of Scientific and Engineering Research, 2021, 8(4):99-105

● Ronald Adam and Suharjito Suharjito (2014), Plagiarism detection algorithm using natural language processing based on grammar analysis, Journal of Theoretical and Applied Information Technology 10(63):168 – 180

● Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph (2010), A view of Cloud Computing, Communications of the ACM 53(4):50-58 DOI:10.1145/1721654.1721672

● Simone Belli, Cristian López Raventós, and Teresa Guarda (2020), Plagiarism Detection in The classroom: Honesty and Trust Through the Urkund and Turnitin Software, Springer Nature Switzerland AG 2020Á. Rocha et al. (Eds.): ICITS 2020, AISC 1137, pp. 660–668, 020.https://doi.org/10.1007/978-3-030-40690-5_63