**Data Extraction Of FYP Title Page**



Session (Fall 2020 – Spring 2024)

Program

**Bachelor of Studies in Computer Science**

Submitted By

Usman Javed 301-20164

Muhammad Shoaib 301-20165

Rimsha Bibi 301-20181

Supervised By

Mr Imran Khan

Lecturer Department of CS & IT

**Hazara University Mansehra**

**APPROVAL SHEET**

A project report submitted to Department of Information Technology, Hazara University, in partial fulfillment of the requirement for the award of BSCS Degree.

**Submitted by:**

Usman Javed 301-20164 Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Muhammad Shoaib 301-20165 Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Rimsha Bibi 301-20181 Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

It is our judgment that this project report is of sufficient standard to warrant its acceptance by the Department of computer science and IT, Hazara University Mansehra.

**Approved by:**

**Internal Supervisor:**

Mr. Imran Khan Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Assistant Professor

Department of Computer Sciences & IT

Hazara University Mansehra

**External Examiner:** Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Head of Department**

Dr. Mushtaq Ahmad Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Department of Computer Sciences & IT

Hazara University, Mansehra

**DECLARATION**

It is here by declared that this software, neither as a whole nor as a part has been copied out from any source. It is further declared that We developed this software and this report entirely on the basis of our personal efforts made under the sincere guidance of our project Supervisor.

If any part of this software is proved to be copied or found to be a report of some other, We standby the consequences. No portion of the work presented in this report has been submitted in support of any application from any other degree of qualification of any other university or institute of learning. We further declare that this software and all associated documents, reports and records are submitted as partial requirement for the degree of BSCS (Bachelor of Studies in Computer Science). We therefore understand and transfer copyrights for these materials to Hazara University Mansehra.

**Usman Javed -----------------------**

**MuhammadShoaib -----------------------**

**Rimsha Bibi -----------------------**

**ACKNOWLEDGMENT**

With the great name of **ALLAH**, the most gracious and merciful, who gifted us with blessings, strength and serenity, without which we could not complete this project. Before we get into things, we would like to add a few heartfelt words for the people for their help and support in this project. People who gave unending support right from the stage when project idea were conceived. In particularly, we are extremely indebted to my teacher and supervisor **Mr Imran Khan** for his guidance to complete this academic for the award of BSCS degree. Class fellows and friends for their support and motivation. We are extremely thankful to our beloved parents and family whose prayers and continuous moral and financial support enabled us to make the successful completion of this project possible.

**DEDICATION**

To

Our beloved parents, respected teachers and

our supervisor

Mr. Imran Khan

Whose support have given us the strength Determination

& Fortitude to Accomplish our Goal.

They all are Assets of our Life.

**PROJECT IN BRIEF**

|  |  |
| --- | --- |
| Project Title: | Data Extraction Of FYP Title Page |
| Developed By: | Usman Javed  Muhammad Shoaib  Rimsha Bibi |
| Supervised By: | Mr Imran Khan |
| Operating System Used: | Windows 10 |
| System Used: | Dell Core i5, 16gb RAM |
| Tools Used: | VSCode Editor, Streamlit, Python,OCR, |
| Starting Date: | Dec-2023 |
| Completion Date: | May-2024 |

**TABLE OF CONTENT**

**Chapter 1.**

**INTRODUCTION**

[1.1. Introduction 1](#_Toc172151142)

[1.2. Problem Statement 1](#_Toc172151143)

[1.3. Scope of the Project 2](#_Toc172151144)

[1.4. Advantages of proposed system 3](#_Toc172151145)

[1.5. Future Work 4](#_Toc172151146)

[1.6. Feasibility of the Project 4](#_Toc172151147)

[1.6.1. Economic Feasibility 4](#_Toc172151148)

[1.6.2. Technical Feasibility 5](#_Toc172151149)

**Chapter 2.**

REQUIREMENT SPECIFICATION

[2.1 Introduction 6](#_Toc172151150)

[2.2 Present System 6](#_Toc172151151)

[2.3 Why We Need New System 7](#_Toc172151152)

[2.4 Drawbacks and Limitations of The Present System 7](#_Toc172151153)

[2.5 Tools Used In The Project: 9](#_Toc172151154)

[2.5.1. Software Tools Used: 9](#_Toc172151155)

[2.5.2. Technologies and Language: 10](#_Toc172151156)

[2.5.3. Hardware Requirements: 11](#_Toc172151157)

[Chapter 3.](#_Toc168568968)

[DESIGN OF THE PROPOSED SYSTEM](#_Toc168568969)

[3.1. Overview of Proposed System 12](#_Toc172151158)

[3.2. Key Features of the Proposed System: 12](#_Toc172151159)

[3.2.1. Image-to-Text Conversion: 12](#_Toc172151160)

[3.2.2. Centralized Database: 12](#_Toc172151161)

[3.2.3. User-Friendly Interface: 12](#_Toc172151162)

[3.2.4. Advanced Search Capabilities: 12](#_Toc172151163)

[3.2.5. Report generation and and Export : 12](#_Toc172151164)

[3.2.6. Scalability: 12](#_Toc172151165)

[3.2.7. Integration and Compatibility: 12](#_Toc172151166)

[3.3. Functional Requirements 13](#_Toc172151167)

[3.4. Use Case Diagram: 14](#_Toc172151168)

[3.4.1. Flow of Actions 16](#_Toc172151169)

[3.5. Sequence Diagram: 16](#_Toc172151170)

[3.6. Entity-Relationship Diagram: 18](#_Toc172151171)

[3.7. Non-Functional Requirements: 20](#_Toc172151172)

[Chapter 4.](#_Toc168568980)

[TESTING](#_Toc168568981)

[4.1. Introduction 21](#_Toc172151173)

[4.2. Unit Test 21](#_Toc172151174)

[4.3. System Test 23](#_Toc172151175)

[4.4. conclusion 24](#_Toc172151175)

[Chapter 5.](#_Toc168568985)

RESULTS AND [SCREENSHORTS](#_Toc168568986)

[5.1. First Interface 26](#_Toc172151176)

[5.2. Browse File 26](#_Toc172151177)

5.3. Extracted Data from Image……………………………………………………………………...27

[5.4. Saving Record](#_Toc172151178) 28

[5.5. Search Results by Title 28](#_Toc172151179)

[5.6. Search Results by Session](#_Toc172151180) 29

[5.7. Search Results by Supervisor 29](#_Toc172151181)

[5.8. Download PDF 30](#_Toc172151181)

[Chapter 6.](#_Toc168568991)

[REFEREN](#_Toc168568992)CE MANUAL

6.1. Reference Manual ………………………………………………………………………………………………………………………32

6.1.1. Figure 5.1Initial Interface …………………………………………………………………………………………………….32

6.1.2. Figure 5.2 Image Upload and Processing ……………………………………………………………………………..32

6.1.3. Figure 5.3 Extracted Text Display …………………………………………………………………………………………32

6.1.4. Figure 5.4 Saving Extracted Text …………………………………………………………………………………………. 32

6.1.5. Figure 5.5 Search by Title ……………………………………………………………………………………………………. 32

6.1.6. Figure 5.6 Search by Session …………………………………………………………………………………………………32

6.1.7. Figure 5.7 Search by Supervisor ……………………………………………………………………………………………33

6.2. Future Scope……………………………………………………………………………………………………………………………….33

[REFERENCES …….……. 34](#_Toc172151183)

**List of Figures**

[Figure 3.1: Use Case Diagram 14](#_Toc168951506)

[Figure 3.2: Sequence Diagram 17](#_Toc168951507)

[Figure 3.3: Entity-Relationship Diagram 19](#_Toc168951508)

[Figure 5.1: First Interface 26](#_Toc168951518)

[Figure 5.2: Browse File 27](#_Toc168951519)

[Figure 5.3: Extracted Data from Image 27](#_Toc168951520)

[Figure 5.4: Saving Record 28](#_Toc168951521)

[Figure 5.5: Search by Title 28](#_Toc168951522)

[Figure 5.6: Search by Session 29](#_Toc168951523)

[Figure 5.7: Search by Supervisor Name 30](#_Toc168951524)

**List of Tables**

[Table 4.1: Upload Image Module 21](#_Toc168951509)

[Table 4.2: Conversion of Image 21](#_Toc168951510)

[Table 4.3: Save Extracted Data in Database 22](#_Toc168951511)

[Table 4.4: Side Bar Navigation Module 22](#_Toc168951512)

[Table 4.5: Search Module 22](#_Toc168951513)

[Table 4.6: Fetching Record Module 22](#_Toc168951514)

[Table 4.7: Delete Record Module 23](#_Toc168951515)

[Table 4.8: Total Record Module 23](#_Toc168951516)

[Table 4.9: System Test 23](#_Toc168951517)

## Introduction

Managing academic records, especially students thesis, presents significant challenges for universities due to the sheer volume of documents and the necessity for precision, accessibility, and security. Traditionally, this process has depended on manual, paper-based systems that are labor-intensive, error-prone, resource-heavy, and inefficient for data retrieval and preservation. As educational institutions grow and the volume of academic output increases, there is a pressing need for a more efficient, reliable, and scalable solution to manage these records.

The "Data Extraction of FYP title page" project tackles these challenges by introducing a comprehensive software application designed to digitize and streamline thesis document management. The core idea of this project is to replace manual data entry and physical document storage with an automated, digital system that uses Optical Character Recognition (OCR) technology for accurate data conversion and a centralized database for secure storage. This solution aims to enhance accessibility, improve data security, and support advanced search capabilities, thereby facilitating better management and utilization of academic records. The project's objectives include reducing the time and resources needed for record management, ensuring data integrity, and unlocking new research potentials by providing easy access to a vast repository of digitized theses. The necessity for such a system is underscored by the inefficiencies and limitations of current methods, highlighting the importance of adopting innovative technologies to meet the evolving needs of academic institutions.

## Problem Statement

In the academic world, particularly within universities, the management of thesis records and related documents presents various challenges and issues that need to be addressed. Let's delve into these problems in detail:

1. **Manual Data Entry:** The current system heavily relies on manual data entry for the conversion of physical documents into digital formats. This process is not only time-consuming but also prone to errors, jeopardizing the accuracy and integrity of academic records.
2. **Resource Intensiveness:** The manual data entry and document archiving process require a substantial allocation of human resources. University staff spends valuable time on transcription tasks, diverting their attention from more critical academic responsibilities.
3. **Limited Accessibility:** Accessing specific thesis documents can be challenging, especially when they are archived or stored in physical formats. This limited accessibility hinders the dissemination of knowledge and research.
4. **Inefficient Search and Retrieval:** The lack of a standardized, digitized database makes it difficult for students, faculty, and researchers to efficiently search, retrieve, and cross-reference academic records and theses.
5. **Data Security Concerns:** Physical documents are vulnerable to loss, damage, or unauthorized access, which raises concerns about data security and the protection of sensitive academic information.
6. **Ineffective Record Preservation:** Traditional methods of preserving academic records are often insufficient. Physical documents may deteriorate over time, potentially leading to data loss.
7. **Missed Research Opportunities:** The current system does not harness the research potential within academic records. Data analysis and insights that could contribute to academic advancements remain largely untapped.
8. **Lack of Scalability:** As universities expand and the volume of academic records grows, the current system struggles to scale efficiently to accommodate this expansion.
9. **Inconsistent Data Format:** The lack of a standardized data format for academic records makes it challenging to perform systematic data analysis and reporting.
10. **Environmental Impact:** The excessive use of physical documents contributes to resource consumption, including paper and storage space, with environmental implications.

## Scope of the Project

The scope of the "Data Extraction of FYP title page" is to develop a comprehensive software solution that addresses the challenges associated with managing academic theses and related documents within university settings. This project aims to revolutionize the way academic records are handled by introducing an efficient and user-friendly approach to digitizing and archiving these documents. The system's key objectives are to eliminate the time-consuming manual data entry process, reduce resource intensiveness, enhance accessibility to academic records, improve data security, and unlock the research potential within the archives. The system's scope encompasses multiple facets, including the development of image-to-text conversion using Optical Character Recognition (OCR) technology, the establishment of a centralized and secure database, the creation of a user-friendly interface, advanced search capabilities, document security, scalability to accommodate the growing volume of records, integration with existing university databases, and a focus on environmental sustainability. The project also seeks to provide a seamless transition from traditional paper-based record-keeping methods to a digital solution and ultimately contribute to the efficient management of academic records in the long term.

## Advantages of proposed system

The following are the advantage of proposed system.

1. **Efficiency Improvement**: By automating the data entry process through OCR technology, the system significantly reduces the time and effort required to digitize thesis documents, freeing up university staff for more critical tasks.
2. **Enhanced Accessibility**: Digitized records stored in a centralized database ensure that students, faculty, and researchers can easily access and retrieve specific documents from anywhere at any time.
3. **Data Security**: Implementing robust security measures and role-based access control protects sensitive academic records from unauthorized access, loss, or damage.
4. **Advanced Search Capabilities**: The system's powerful search functionality allows users to quickly find specific theses based on various criteria such as keywords, author names, and publication years, improving research efficiency.
5. **Scalability**: Designed to accommodate the growing volume of academic records, the system can scale seamlessly with the university's expanding needs without compromising performance.
6. **Environmental Sustainability**: By digitizing documents and reducing paper usage, the project contributes to environmental sustainability and resource conservation.
7. **Improved Data Integrity**: Automated data entry minimizes errors associated with manual transcription, ensuring the accuracy and reliability of academic records.
8. **Research Potential**: The digital repository enables data analysis and insights, unlocking new research opportunities and contributing to academic advancements.

## Future Work

1. **Integration with Advanced Data Analytics**: Future development could include integrating advanced data analytics tools to derive insights from the digitized records, supporting academic research and decision-making processes.
2. **Artificial Intelligence Enhancements**: Incorporating AI algorithms to enhance OCR accuracy and automate metadata tagging could further improve the efficiency and reliability of the system.
3. **Mobile Accessibility**: Developing mobile applications to provide on-the-go access to the academic records, enhancing the system's usability for students and faculty.
4. **Cloud-Based Solutions**: Transitioning to cloud-based storage and processing could offer improved scalability, data redundancy, and disaster recovery options.
5. **User Feedback Integration**: Continuously gathering user feedback to refine and enhance the system's features, ensuring it meets the evolving needs of the university community.
6. **Inter-University Collaboration**: Expanding the system to support inter-university collaborations, enabling shared access to academic records and fostering a broader research community.
7. **Enhanced Security Protocols**: Regularly updating and enhancing security measures to protect against emerging cyber threats and ensure the continued safety of academic records.
8. **Customization Options**: Offering customizable features to cater to the specific needs and workflows of different universities, ensuring the system is adaptable to various academic environments.

## Feasibility of the Project

The feasibility analysis of "Data Extraction of FYP title page" ensures that the project is viable and beneficial from multiple perspectives, including economic and technical aspects.

### Economic Feasibility

The economic feasibility of the "Data Extraction of FYP title page" project is promising due to the significant cost savings and financial benefits it offers. By automating the data entry process and transitioning from paper-based systems to a digital platform, the project eliminates the substantial labor and material costs associated with manual transcription and physical document storage. The initial investment in the development and deployment of the system is quickly offset by long-term savings from reduced error rates, enhanced efficiency, and improved accessibility of academic records. This efficiency not only streamlines administrative tasks but also allows university staff to focus on more critical responsibilities, thereby increasing overall productivity. Furthermore, the scalable design of the system ensures that it can grow alongside the institution without necessitating frequent, costly upgrades, providing sustained economic benefits as the volume of academic records expands.

### Technical Feasibility

The technical feasibility of the "Data Extraction of FYP title page" project is well supported by the availability of robust technologies and the expertise of the project team. The essential technologies for the system, including Optical Character Recognition (OCR), centralized databases, and secure data storage, are mature and widely accessible. Leveraging established tools such as Tesseract OCR, Python, Streamlit, and MySQL, the project can be developed and implemented effectively. The team members possess the necessary skills in backend handling, frontend application development, and documentation, ensuring that all technical requirements are adequately addressed. The system is designed for seamless integration with existing university databases, facilitating smooth data migration and adoption. Its scalable architecture guarantees efficient handling of large data volumes without performance degradation. Additionally, robust security measures, such as role-based access control, ensure the protection of sensitive academic records, while the user-friendly interface promotes ease of use for students, faculty, and staff, enhancing overall system reliability and adoption.

## Introduction

The proposed "Data Extraction of FYP title page" is set to transform the management of academic thesis and related documents within universities. By harnessing advanced technology, this system aims to digitize and streamline the entire process, providing a user-friendly, efficient, and secure method for handling large volumes of academic records. Utilizing Optical Character Recognition (OCR) for precise data conversion and maintaining a centralized database, the system ensures easy access, enhanced data security, and scalability. This innovative solution addresses the shortcomings of traditional paper-based methods, catering to the modern needs of educational institutions.

## Present System

This project involves the development of a web application tailored for the efficient management of student data. The primary goal of this application is to streamline the processes of storing, retrieving, deleting, and reporting on student information, with a strong focus on leveraging an SQL database to ensure data integrity and organization.

The application is designed to handle multiple facets of student-related information by categorizing it into distinct sections. These sections include titles of student projects, names of supervisors, academic sessions, and student names. Each of these categories is managed through separate data tables in an SQL database, which facilitates easy data manipulation and retrieval.

In the data storage and retrieval process, the application provides a structured approach to handling different types of information. For instance, the Title Section is responsible for managing the titles of student projects. Each project title is recorded as an individual entry in the SQL database, ensuring that titles are uniquely identified and easily accessible. Similarly, the Supervisor Section maintains records of supervisors who are assigned to oversee these projects. Each supervisor's name is stored in a dedicated table, allowing for straightforward linking of supervisors to their respective projects.

The Session Section logs academic sessions relevant to the projects, with each session stored as a separate record in its own table. This organization helps track the timeline and context in which each project is conducted. The Student Section keeps track of student names and their associated details. Each student’s information is entered into a specific table, ensuring that all related data is kept together and easily manageable.

The application also includes robust data management features. Users can add new records to the system through intuitive forms that allow for the input of data into the appropriate sections. Once data is entered, it is stored in the SQL database, organized into the respective tables for titles, supervisors, sessions, and students. This process ensures that all information is accurately recorded and categorized.

Retrieving data is made simple through the application’s search functionalities. Users can query the SQL database to find specific information based on various criteria, such as project titles, supervisor names, or student details. This ability to efficiently search and access data enhances the user experience and supports quick decision-making.

The application also provides functionality for deleting outdated or incorrect records. Users can remove entries from the SQL database as needed, ensuring that the data remains current and relevant. Additionally, the application allows for the updating of existing records. This feature enables users to correct or modify information as necessary, maintaining data accuracy and integrity.

One of the key features of the application is its capability to generate and export reports in PDF format. Users can create comprehensive reports summarizing the stored data, which can be downloaded and saved for offline use or printed for physical distribution. This report generation feature is particularly valuable for documentation and presentation purposes, providing a clear and organized way to share information with stakeholders.

The use of an SQL database is integral to the application’s design, providing a robust and reliable system for managing data. SQL databases are known for their ability to handle large volumes of data efficiently and support complex queries, which is essential for maintaining the diverse and detailed records involved in student data management. By organizing data into separate tables for each category, the application ensures that information is systematically stored and easily accessible.

Extracting text from thesis images using Optical Character Recognition (OCR) involves preparing high-quality images with proper resolution and contrast. After selecting an OCR tool Tesseract OCR, the software converts the images into editable text.

## Why We Need New System

The existing manual, paper-based system for managing academic records is plagued by inefficiencies and limitations that hinder productivity and accuracy. Manual data entry is labor-intensive and error-prone, diverting valuable human resources from more essential academic tasks. Physical document storage demands considerable space and resources, making accessibility and preservation challenging. Furthermore, the absence of a standardized digital database complicates the efficient search and retrieval of specific documents. These issues necessitate a new system that automates data entry, enhances accessibility, improves data security, and offers scalable solutions to meet the growing demands of academic institutions.

## Drawbacks and Limitations of The Present System

Some of the Drawbacks and Limitations of the present System are the following:

**1. Manual Data Entry Errors:** One of the most significant drawbacks of the present system for managing student data is the reliance on manual data entry. This traditional approach is highly susceptible to human error, which can lead to incorrect or inconsistent data. Errors may occur during the input process due to typographical mistakes, misinterpretation of handwritten information, or oversight of critical details. These inaccuracies can affect various aspects of data management, from student records to project details, potentially leading to incorrect decisions or evaluations based on flawed data. Furthermore, manual data entry can be time-consuming, especially when handling large volumes of information, which can impact the efficiency and productivity of administrative staff.

**2. Inefficient Data Retrieval**: Another significant issue with the present system is the inefficiency of data retrieval. Traditional methods often involve manually searching through physical records or navigating complex databases with limited search functionalities. This can be particularly problematic when dealing with extensive datasets, such as student records or project details. The lack of advanced search features, such as filtering or sorting options, can make it challenging to locate specific information quickly. As a result, staff may spend considerable time searching for and retrieving data, which can delay decision-making processes and hinder overall administrative efficiency.

**3. Lack of Centralized:** Access Information is scattered across multiple systems or spreadsheets, complicating the process of obtaining a unified view of student data. This decentralization makes data retrieval cumbersome and time-consuming, hindering efficient access to necessary information.

**4. Inefficient Reporting:** Generating reports manually is a slow and error-prone process, limiting the ability to extract meaningful insights and analytics from the data. The lack of automated reporting tools means that creating comprehensive reports requires significant effort and is prone to inaccuracies.

**5. Data Security Issues:** Storing sensitive student information in unsecured or poorly managed systems increases the risk of data breaches. Managing access control is also challenging, making it difficult to ensure that only authorized personnel have access to specific data, thereby compromising data security.

**6. Error Prone Processes:** Manual data entry processes are highly susceptible to human errors, resulting in incorrect data being stored. The absence of automated validation allows incorrect or incomplete data to enter the system easily, further compromising data quality.

1. **Limited Data Retrieval and Search Capabilities:** The present system has limited data retrieval and search capabilities, making searches for specific information slow and inefficient. This limitation hinders quick access to necessary data, affecting decision-making and overall productivity.

## 

## 2.5. Tools Used In The Project

**Front-End Development:** The graphical user interface (GUI) is developed, ensuring responsive design for mobile and desktop access.

**Back-End Development**: The system's core functionalities are developed, including image-to-text conversion, search algorithms, and database integration.

**Integration of SQL database:** The SQL database forms the backbone of the application, storing and managing all student-related data.

### 2.5.1. Software Tools Used

### Visual Studio Code Editor

**Description:** VS Code is a free, open-source code editor developed by Microsoft. It offers a wide range of features such as IntelliSense (code completion), debugging, Git integration, and a rich extension ecosystem.

**Key Features**: Integrated terminal, powerful extensions, customizable interface, and robust language support.

**Alternatives to VS Code**

**1. Sublime Text:**

**Description:** Sublime Text is a fast, lightweight, and highly customizable code editor known for its speed and simplicity.

**2. Notepad++:**

**Description**: Notepad++ is a free, open-source code editor for Windows, known for its simplicity and lightweight footprint.

**3. Eclipse:**

**Description:** Eclipse is a popular, open-source IDE primarily used for Java development but supports other languages through plugins.

**4. Atom**

**Description:** Atom is a free and open-source code editor developed by GitHub, designed to be highly hackable and customizable.

### Technologies and Language:

**Python**

* **Purpose**: Python is a high-level programming language used for backend programming.
* **Alternatives**: Some alternatives to Python for backend programming include:
  + JavaScript (Node.js)
  + Ruby
  + Java
  + PHP

**Streamlit**

* **Purpose**: Streamlit is a python library used for frontend design.
* **Open Source**: Yes, Streamlit is open source and available under the Apache 2.0 license.
* **Alternatives**: Some alternatives to Streamlit for building web applications include:
  + Dash (by Plotly)
  + Flask
  + Django
  + Shiny (for R)

**Tesseract OCR (Optical Character Recognition)**

* **Working of OCR:** In this project, OCR technology is employed to extract text from scanned images of FYP title pages, converting them into editable and searchable data. The process begins with capturing high-quality images of the title pages. Using an OCR tool like Tesseract, the software analyzes the images, segmenting them into lines and characters, and applies pattern recognition algorithms to identify the text. The extracted text is then stored in a centralized SQL database, ensuring accurate and efficient data management. This enables easy access, retrieval, and reporting of academic records, streamlining the overall management of thesis documents within the university.
* **Purpose**: Tesseract is used for character recognition in the image.
* **Open Source**: Yes, Tesseract is open source and available under the Apache 2.0 license.
* **Reference**: More details can be found on the official Tesseract GitHub page: [Tesseract GitHub](https://github.com/tesseract-ocr/tesseract).
* **Alternatives**: Some alternatives to Tesseract OCR include:
  + Google Cloud Vision OCR
  + ABBYY FineReader
  + Adobe Acrobat OCR
  + Amazon Textract

**MySQL**

* To develop the student data management web application, we utilize an SQL database with PHP on a localhost environment, facilitated by phpMyAdmin and XAMPP server. This setup involves installing XAMPP, which includes Apache (web server), MySQL (database server), and PHP. Once installed, start Apache and MySQL services via the XAMPP control panel. Access phpMyAdmin through http://localhost/phpmyadmin to create databases and tables for managing student projects, supervisors, academic sessions, and student information. Your PHP scripts will handle data operations like storing, retrieving, updating, and deleting records, ensuring data integrity and organization.
* **Purpose**: MySQL is a popular relational database management system used for storing and managing data. It is used in conjunction with PHP for web development.
* **Open Source**: Yes, MySQL is open source and available under the GNU General Public License (GPL).
* **Reference**: You can find more information on the official MySQL website: [mysql.com](https://www.mysql.com/).
* **Alternatives**: Some alternatives to MySQL include:
  + PostgreSQL
  + SQLite
  + MariaDB
  + Microsoft SQL Server

### Hardware Requirements:

* Processor: Intel dual core or above
* Processor Speed:1.0GHZ or above
* RAM: 2 GB RAM or above
* Hard Disk: 20 GB hard disk or above

## Overview of Proposed System

The "Data Extraction of FYP title page" aims to change how universities handle student theses and academic documents. It offers a complete and advanced system to tackle the current problems like making things faster, easier to access, more secure, and helping with research.

## Key Features of the Proposed System:

### Image-to-Text Conversion:

The core functionality of the system is the conversion of images, such as scanned thesis documents, into machine-readable text. Leveraging Optical Character Recognition (OCR) technology, this process ensures accurate and reliable data entry, eliminating the need for manual transcription.

### Centralized Database:

All academic records and thesis documents are stored in a centralized and secure database. This centralized approach ensures easy access, efficient management, and systematic organization of records.

### User-Friendly Interface:

The system features an intuitive and user-friendly interface accessible to students, faculty, and university staff. Users can easily search, retrieve, and browse academic records.

### Advanced Search Capabilities:

The search functionality allows users to perform advanced searches based on keywords, supervisor names, thesis titles, Session and more. This streamlined search process significantly enhances record retrieval.

### ****Report Generation and Export****:

Ability to generate comprehensive reports summarizing stored data. Export reports in PDF format for offline use or physical distribution.

### Scalability:

The system is designed to scale efficiently with the growing volume of academic records. It can accommodate the needs of expanding universities without compromising performance.

### Integration and Compatibility:

The system is designed to integrate with existing university databases, making the migration process seamless.

## Functional Requirements

Functional Requirements (FR) outline the specific functionalities and features that the "Data Extraction of FYP title page" must possess to meet the needs of its users and achieve its objectives. Here's a general overview of the FR for the system:

**Module 1: Image-to-Text Conversion**

* Implement the image-to-text conversion feature using Optical Character Recognition (OCR) technology.
* Develop a user interface for users to upload an image.
* Integrate OCR to analyze the image and extract text from it.
* Display the extracted text to the user.

**Module 2: Text Saving to Database**

* Create a database structure to store the extracted text records.
* Develop a user interface with a "Save" button.
* Implement the functionality to save the extracted text into the database when the user clicks the "Save" button.

**Module 3: Sidebar and Navigation**

* Design a sidebar menu for easy navigation within the app.
* Include sections searching records and viewing the total number of student records.

**Module 4: Search and Retrieve Records**

* Develop a search bar to allow users to search for specific records in the database.
* Implement the functionality to retrieve records that match the search query.
* Display the search results to the user, showing each record separately.

**Module 5: Record Management**

* Include options for viewing and deleting individual records.
* Add a "Delete" button to each record, allowing users to remove unwanted entries.

**Module 6: Total Record Count**

* Calculate and display the total number of student records in the database.
* Update the count dynamically as new records are added or existing records are deleted.

## Use Case Diagram:

A Use Case Diagram is a visual representation that outlines the interactions between users (actors) and a system to achieve specific goals. It captures the functional requirements of a system by depicting the various use cases, or tasks, that users can perform. The diagram helps in understanding the system's functionality from an end-user perspective and facilitates communication between stakeholders, such as developers, clients, and end-users.

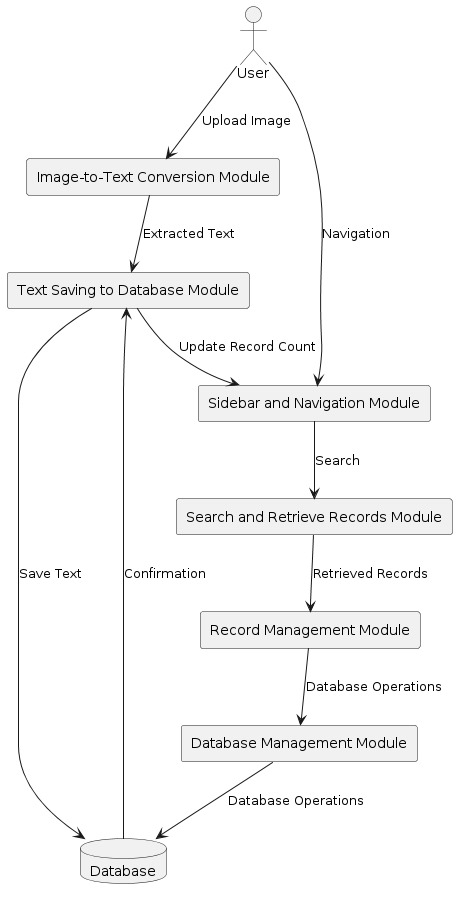


Figure 3.1: Use Case Diagram

**Use Case Diagram Description**

1. **Actor: User**
   * The user interacts with various modules in the system to perform specific actions.
2. **Image-to-Text Conversion Module**
   * **Action**: The user uploads an image.
   * **Process**: The module converts the uploaded image to text using Optical Character Recognition (OCR) technology.
   * **Output**: Extracted text.
3. **Text Saving to Database Module**
   * **Action**: The extracted text is saved to the database.
   * **Process**: The module provides an interface for saving the extracted text.
   * **Output**: Confirmation that the text has been saved to the database.
   * **Interaction**: Updates the record count after saving.
4. **Sidebar and Navigation Module**
   * **Action**: Provides navigation options for the user.
   * **Functionality**: Allows the user to navigate to different sections of the application.
   * **Interaction**: The user can navigate to the search and retrieve records module.
5. **Search and Retrieve Records Module**
   * **Action**: The user performs a search for specific records in the database.
   * **Process**: The module retrieves records that match the search query.
   * **Output**: Retrieved records are displayed to the user.
6. **Record Management Module**
   * **Action**: Provides options for viewing and managing individual records.
   * **Functionality**: The user can view, delete, or manage records.
   * **Interaction**: Interacts with the database management module to perform database operations.
7. **Database Management Module**
   * **Action**: Handles database operations such as saving, updating, and deleting records.
   * **Interaction**: Ensures that all operations on the database are performed correctly and securely.
8. **Database**
   * **Role**: Central storage for all the text records and metadata.
   * **Interaction**: The database is updated with new text records, and existing records can be retrieved or managed.

### 3.4.1 Flow of Actions

1. **Upload Image**: The user uploads an image to the Image-to-Text Conversion Module.
2. **Extract Text**: The module converts the image to text.
3. **Save Text**: The extracted text is sent to the Text Saving to Database Module and saved to the database.
4. **Update Record Count**: After saving, the record count is updated in the Sidebar and Navigation Module.
5. **Navigation**: The user navigates through the Sidebar and Navigation Module to access different functionalities.
6. **Search Records**: The user searches for specific records using the Search and Retrieve Records Module.
7. **Retrieve Records**: The module retrieves and displays records matching the search criteria.
8. **Manage Records**: The user can view or delete records through the Record Management Module.
9. **Database Operations**: All operations on records (save, update, delete) are handled by the Database Management Module, ensuring data integrity and security.

## Sequence Diagram

A Sequence Diagram is a type of interaction diagram in Unified Modeling Language (UML) that shows how objects interact in a particular scenario of a use case. It depicts the sequence of messages exchanged between various objects to carry out a function or process. Sequence Diagrams focus on the order of message flow between the actors and the system and among system components, illustrating how operations are carried out, what messages are sent, and in what order.

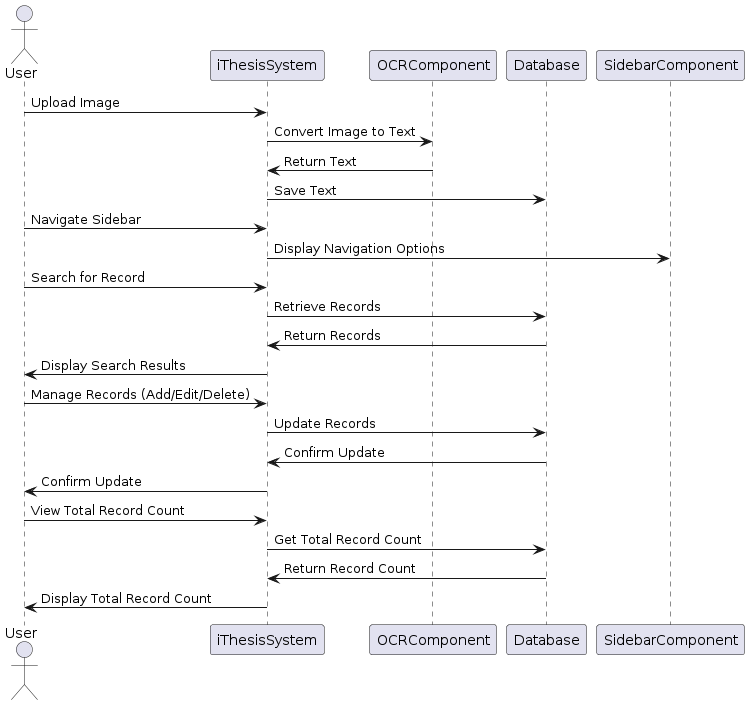


Figure 3.2: Sequence Diagram

**Use Case Diagram Description**

1. **Image-to-Text Conversion**:

* The user uploads an image.
* The system sends the image to the OCR component for conversion.
* The OCR component returns the converted text to the system.
* The system saves the text to the database.

1. **Sidebar and Navigation**:

* The user navigates using the sidebar.
* The system interacts with the sidebar component to display navigation options.

1. **Search and Retrieve Records**:

* The user initiates a search for records.
* The system queries the database for the requested records.
* The database returns the relevant records.
* The system displays the search results to the user.

1. **Record Management**:

* The user manages records by adding, editing, or deleting them.
* The system updates the records in the database.
* The database confirms the update.
* The system confirms the update to the user.

1. **Total Record Count**:

* The user requests to view the total record count.
* The system queries the database for the total record count.
* The database returns the count.
* The system displays the total record count to the user.

## Entity-Relationship Diagram:

An Entity-Relationship Diagram (ERD) is a type of diagram used in database design to visually represent the relationships between data entities within a system. ERDs depict the data structure of a system, showing entities (which represent data objects) and their relationships. Entities are typically represented as rectangles, relationships as diamonds or lines, and attributes as ovals. ERDs are essential in designing and understanding the database architecture, ensuring data integrity, and defining how data is stored and accessed.

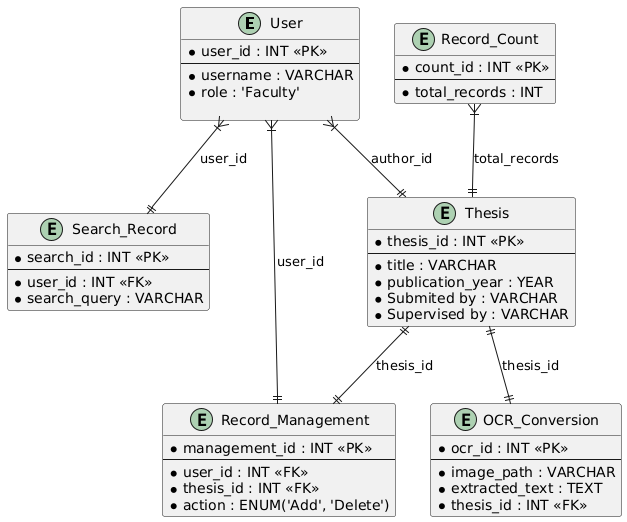


Figure 3.3: Entity-Relationship Diagram

**Use Case Diagram Description**

* **Thesis Document**: Represents the thesis documents stored in the system. Each thesis has a unique ThesisID and contains attributes such as Title, Author, PublicationYear, and Content.
* **Search History**: Represents the search history of users. Each search entry has a unique SearchID and contains the search Query performed by a user. It also stores the UserID of the user who performed the search.
* **Record Management**: Represents the actions performed on thesis records by users. Each action entry has a unique RecordID and contains details about the action (e.g., upload, delete). It also stores the ThesisID of the thesis record involved in the action.

The relationships depicted in the ERD indicate how the entities are connected:

* Users upload thesis documents (Thesis Document) to the system.
* Users perform searches, which are logged in the Search History.
* Users manage thesis records, leading to actions recorded in the Record Management entity.

## Non-Functional Requirements:

* **Usability:** The user interface should be intuitive and straightforward, allowing users to navigate and perform tasks without requiring extensive training.
* **Reliability:** The application should be available 99.9% of the time, ensuring minimal downtime. Ensuring that all data entered into the system is accurate and remains consistent over time. The application should gracefully handle errors and provide meaningful feedback to users, ensuring that data is not lost or corrupted.
* **Scalability:** The SQL database should be designed to handle growth in the volume of data without performance loss. The application architecture should support scaling out (adding more servers) and scaling up (enhancing server capacity) as needed.
* **Responsiveness:** The application should be fully responsive, working seamlessly on various devices, including mobile phones, tablets, laptops, and desktops. The user interface should adapt to different screen sizes and resolutions, providing an optimal viewing and interaction experience on all devices.

## Introduction

A test plan is a comprehensive document that outlines the testing strategy, resources, schedule, and scope for a software project. It acts as a roadmap, ensuring systematic evaluation of the application to identify potential issues early, enhance product quality, and meet specified requirements. This plan fosters clear communication among stakeholders, developers, and testers, promoting transparency and accountability. By carefully planning the testing phase, organizations can reduce risks, lower costs, and deliver a reliable and user-friendly product.

System Testing involves testing the complete and integrated software application to verify that it meets all specified requirements, ensuring the system functions correctly as a whole. This is conducted after integration testing to identify any issues arising from component interactions.

## Unit Test

Unit Testing focuses on testing individual components or units of a software application in isolation to ensure they function correctly. It is usually performed by developers during the coding phase.

Table 4.1: Upload Image Module

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Test Steps | Expected Results | Pass/Fail |
| TC-01 | |  | | --- | | Verify image upload functionality | | |  | | --- | | 1. Open the upload interface.  2. Upload a valid image file. | | |  | | --- | | Image is uploaded successfully and displayed in the interface. | | Pass |

Table 4.2: Conversion of Image

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Test Steps | Expected Results | Pass/Fail |
| TC-02 | Verify OCR conversion of uploaded image | 1. Upload a valid image.  2. Trigger the OCR conversion. | Extracted text is displayed correctly. | Pass |

Table 4.3: Save Extracted Data in Database

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Test Steps | Expected Results | Pass/Fail |
| TC-03 | |  | | --- | | Verify saving extracted text to the database | | |  | | --- | | 1. Upload image and extract text.  2. Click the "Save" button. | | |  | | --- | | Extracted text is saved in the database with a unique ID. | | Pass |

Table 4.4: Side Bar Navigation Module

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Test Steps | Expected Results | Pass/Fail |
| TC-04 | |  | | --- | | Verify sidebar navigation | | |  |  |  | | --- | --- | --- | | |  | | --- | | 1. Enter a query  in the search bar.  2. Execute the search. |  |  | | --- | |  | | | |  |  |  | | --- | --- | --- | | |  | | --- | | Relevant records matching the query are displayed. |  |  | | --- | |  | | | Pass |

Table 4.5: Search Module

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Test Steps | Expected Results | Pass/Fail |
| TC-05 | Verify the effectiveness of the Image-to-Text Conversion Module | 1. Search for a specific word in the extracted text | The highlighted text is displayed in a clear and readable format. | Pass |

Table 4.6: Fetching Record Module

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Test Steps | Expected Results | Pass/Fail |
| TC-06 | Verify retrieval of specific records | * + - 1. Perform a search.       2. Select a specific record from results. | Selected record details are displayed correctly. | Pass |

Table 4.7: Delete Record Module

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Test Steps | Expected Results | Pass/Fail |
| TC-07 | Verify deletion of a record | 1. Search for a record.  2. Click the "Delete" button on the record. | Record is removed from the database and no longer displayed. | Pass |

Table 4.8: Total Record Module

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Test Steps | Expected Results | Pass/Fail |
| TC-08 | Verify total record count update | 1. Add a new record.  2. Delete an existing record. | Total record count is updated dynamically. | Pass |

## System Test

The system test table provides a structured overview of the system's functionalities and corresponding test cases. It outlines each test case's ID, title, description, preconditions, test steps, expected output, and pass/fail criteria. This table serves as a comprehensive reference for systematically verifying the system's behavior and ensuring that it meets the specified requirements.

Table 4.11 "Data Extraction of FYP title page" provides a concise overview of the main system test cases, their descriptions and expected results.

Table 4.9: System Test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Case | Test Objective | Description | Expected Results | Pass/Fail |
| ST-1 | End-to-End Workflow Test | Verify the complete workflow from image upload, text conversion, database save, search, and record management, including navigation. | |  | | --- | | Users can upload an image, convert to text, save to the database, search for records, and manage records seamlessly. |  |  | | --- | |  | | Pass |
| ST-2 | User Access and Navigation Test | Verify user navigation, access to various sections, and role-specific actions. | |  | | --- | | Users can navigate to all sections, perform role-specific actions, and access functionalities without issues. |  |  | | --- | |  | | Pass |
| ST-3 | Data Integrity and Security Test | Verify data integrity, security, and integration with existing university databases. | |  | | --- | | Data remains accurate, secure, and only accessible to authorized users. University database integration functions correctly. |  |  | | --- | |  | | Pass |
| ST-4 | Performance and Analytics Test | Verify system performance under load, and accuracy of reporting and analytics features. | |  | | --- | | System performs well under load, generates accurate reports and analytics, and provides functional help and support resources. |  |  | | --- | |  | | Pass |

## Conclusion

The "Data Extraction of FYP title page" project represents a pivotal advancement in the digital management and archiving of academic theses within university settings. By transitioning from traditional paper-based systems to a sophisticated digital solution, "Data Extraction of FYP title page" addresses numerous challenges and limitations inherent in manual data entry and physical document storage.

Key features of "Data Extraction of FYP title page" such as image-to-text conversion using Optical Character Recognition (OCR), a centralized and secure database, a user-friendly interface, advanced search capabilities, and robust document security measures, collectively provide a streamlined, efficient, and secure approach to handling academic records. The system's scalability ensures it can accommodate the growing volume of academic records as universities expand, while its integration capabilities allow seamless interaction with existing university databases.

The development and implementation phases involved rigorous testing—both unit and system testing—to ensure the functionality, accuracy, performance, and security of each component and the overall system. These tests confirmed the system's readiness for deployment and its ability to meet the outlined objectives.

Automating the digitization and archiving processes through "Data Extraction of FYP title page" not only reduces the resource intensiveness and error-proneness associated with manual data entry but also significantly enhances the accessibility and usability of academic records. This enhancement facilitates efficient search, retrieval, and analysis of theses by researchers, students, and faculty members, thereby unlocking new research opportunities and insights.

Moreover, "Data Extraction of FYP title page" contributes to environmental sustainability by diminishing the reliance on paper and physical storage spaces. This digital transformation aligns with the broader trend of leveraging technology to improve efficiency and sustainability in academic and administrative processes.

## First Interface

This initial interface welcomes the user with a title and provides options to browse and upload files. The navigation bar includes search functionalities with multiple checkbox. for searching by Name, Title, Session, Supervisor name.

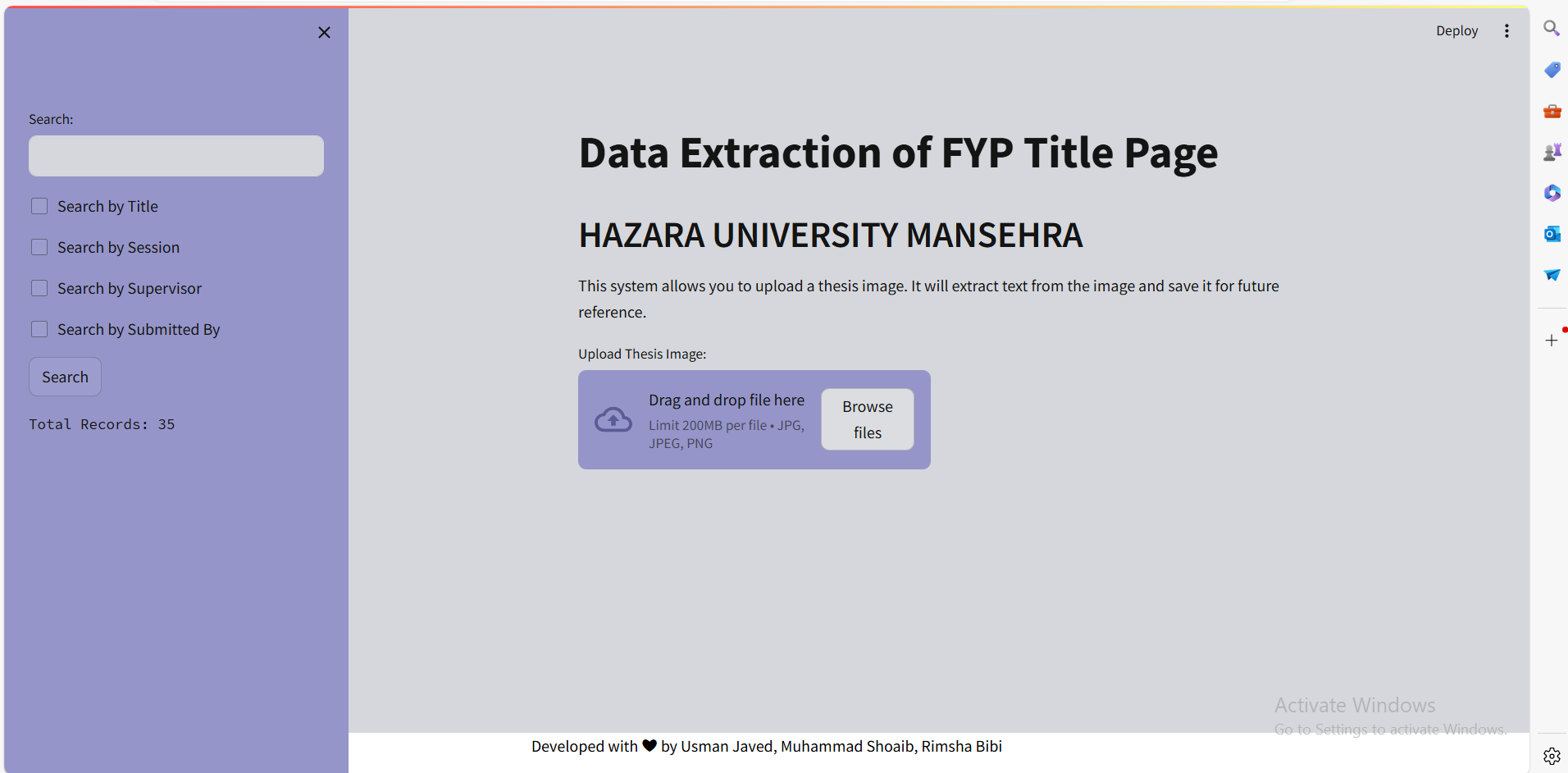


Figure 5.1: First Interface

## Browse File

The user uploads an image, which is processed by the Image-to-Text Conversion module. The extracted text is displayed on the interface, allowing the user to review and save it to the database.

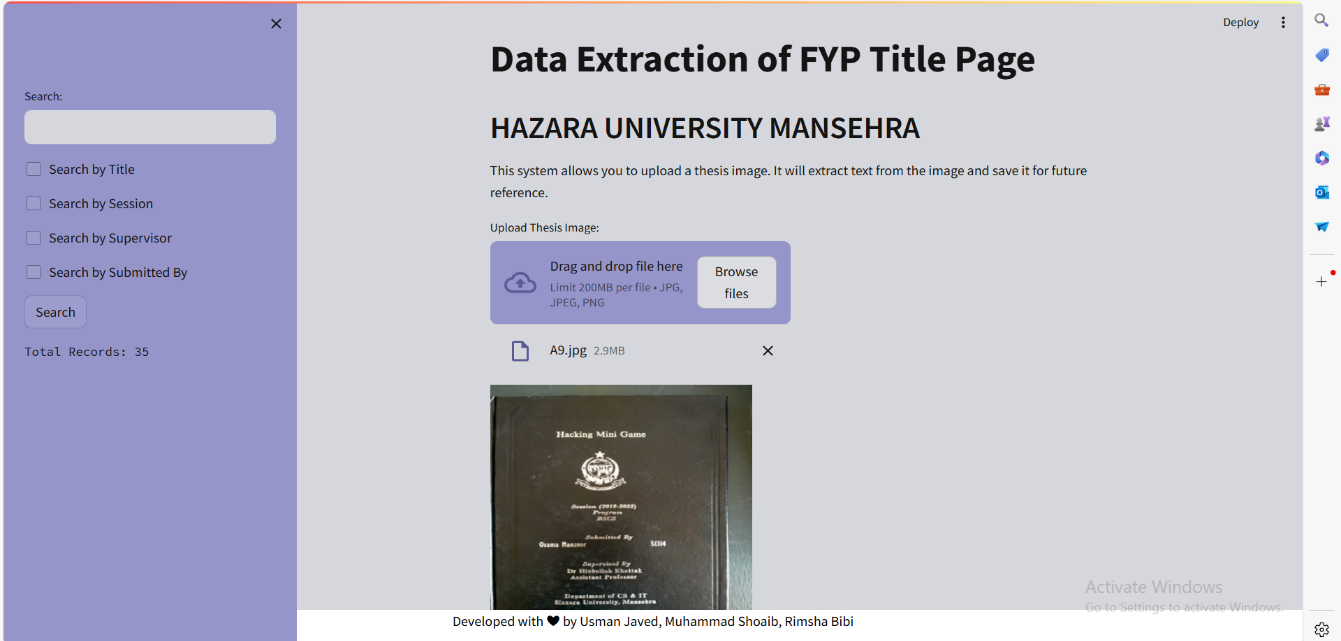


Figure 5.2: Browse File

* 1. **Extracted Data from Image**

This screenshot showing that data is Extracted from images using Optical Character Recognition (OCR).

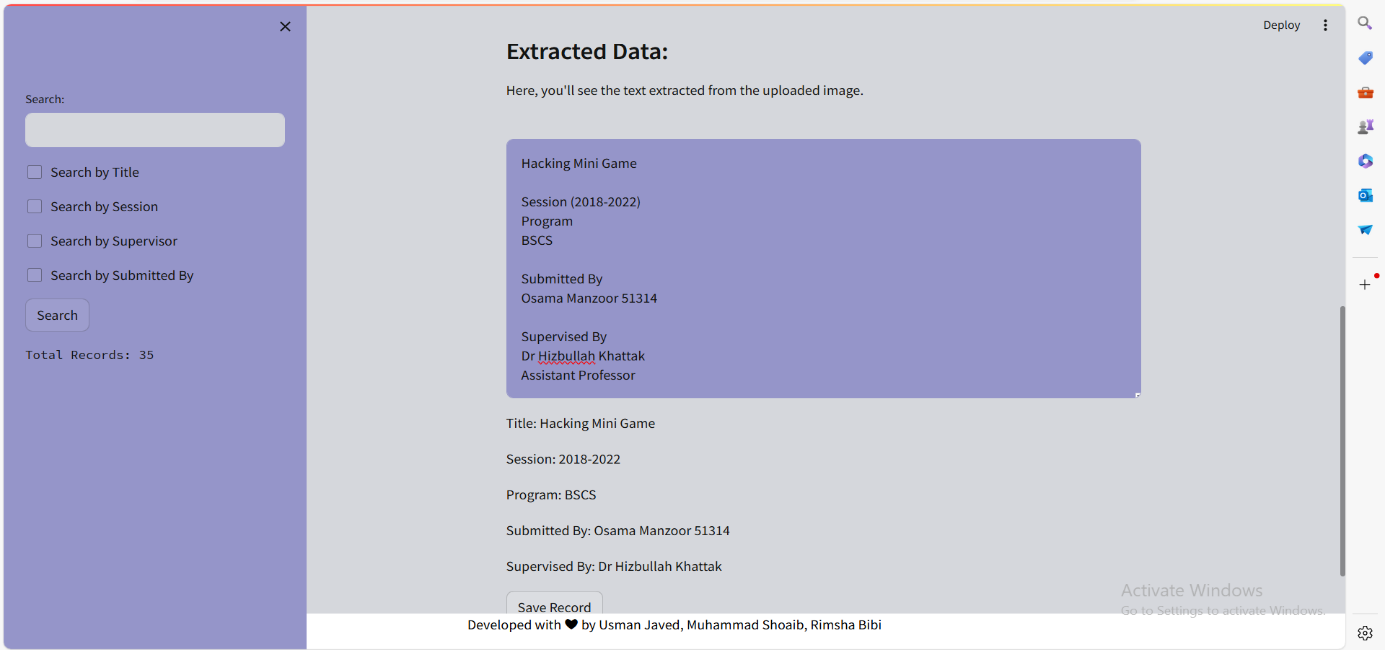


Figure 5.3: Extracted Data from Image

## Saving Record

The “save record” button functionality is displayed in the last which allows the user to save the extracted text from the image into the database when you click that button it will save the record into the database

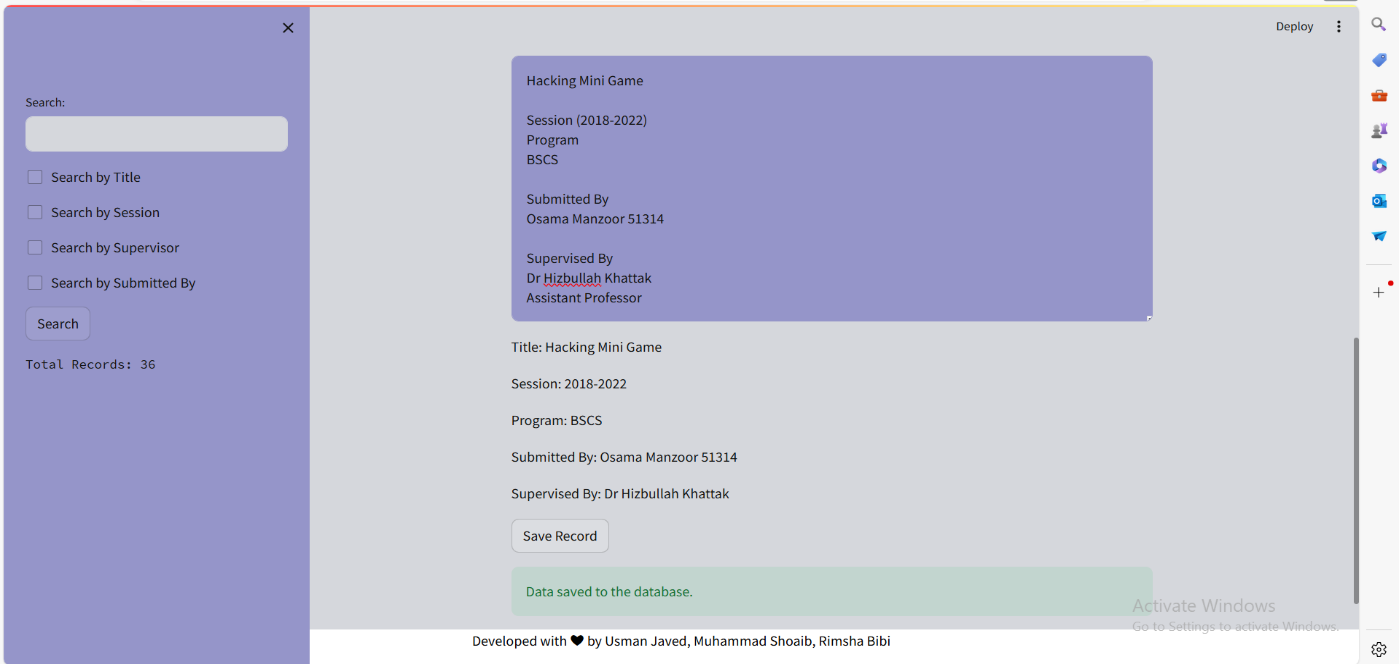


Figure 5.4: Saving Record

## Search Results by Title

This search functionality allows users to retrieve academic records by specifying the title, making it easier to locate documents from a particular academic period.

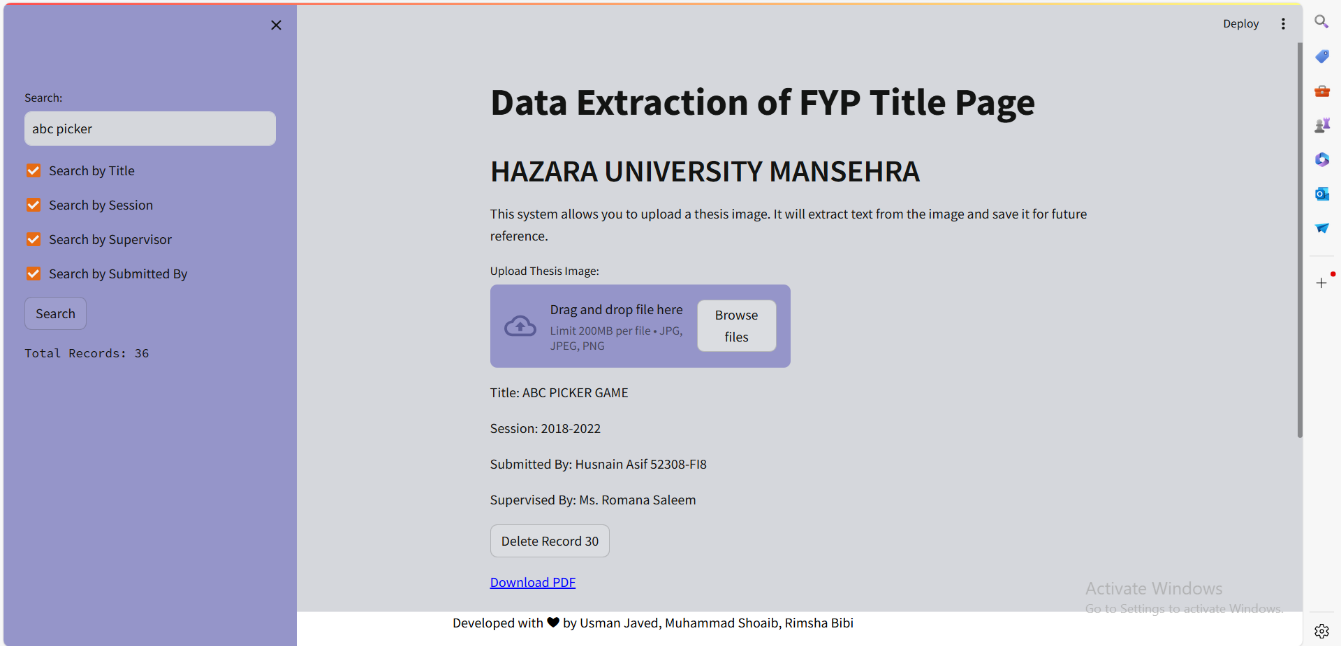


Figure 5.5: Search by Title

## Search Results by Session

This search functionality allows users to retrieve academic records by specifying the session, making it easier to locate documents from a particular academic period.



Figure 5.6: Search by Session

## Search Results by Supervisor

This search feature enables users to find theses and academic records associated with a specific supervisor, facilitating quick access to documents supervised by particular faculty members.

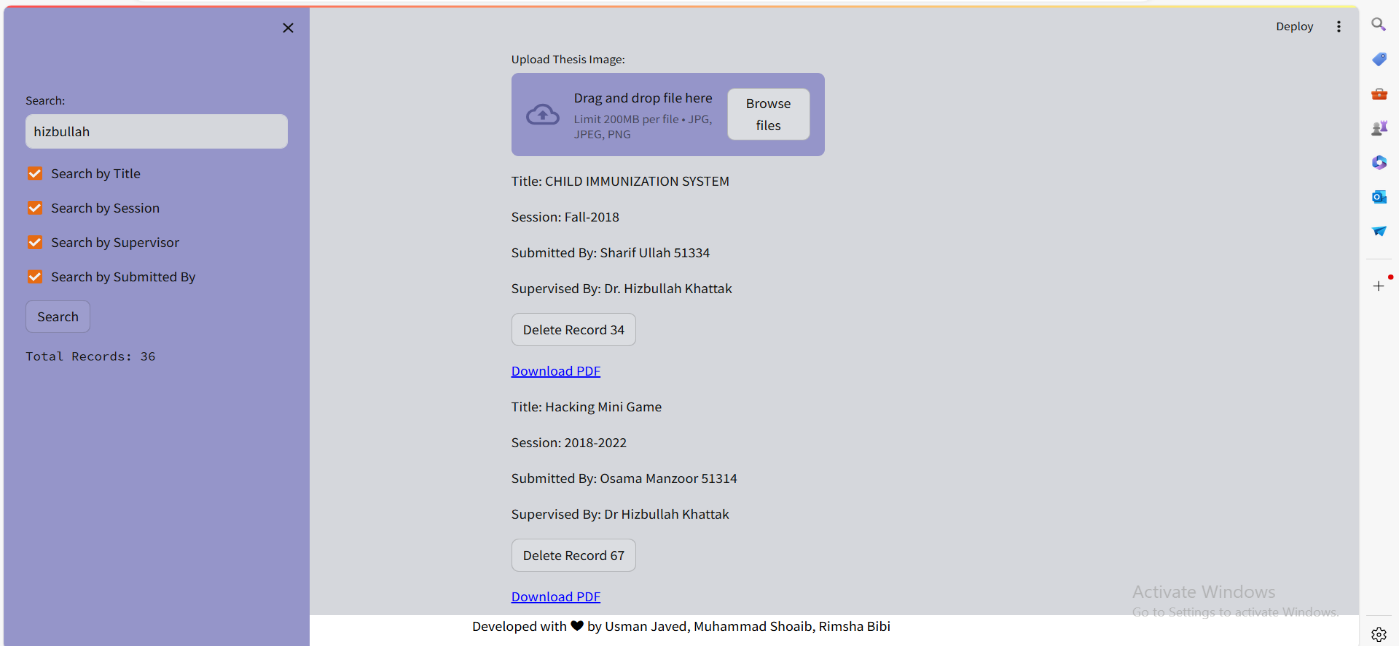
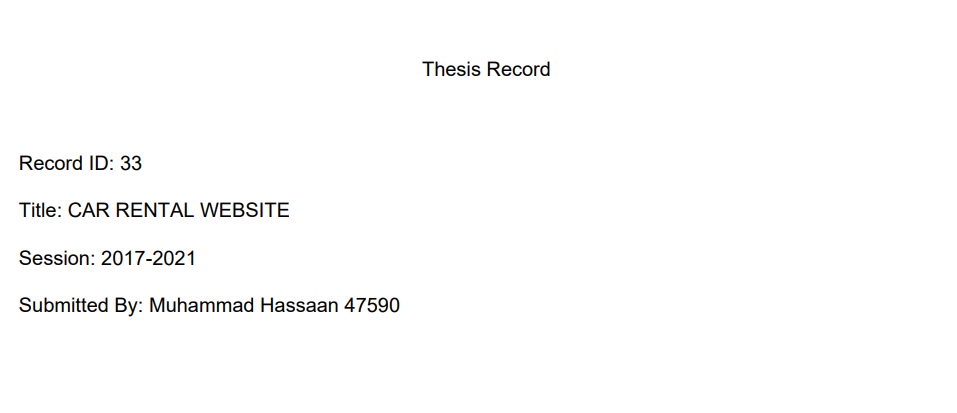
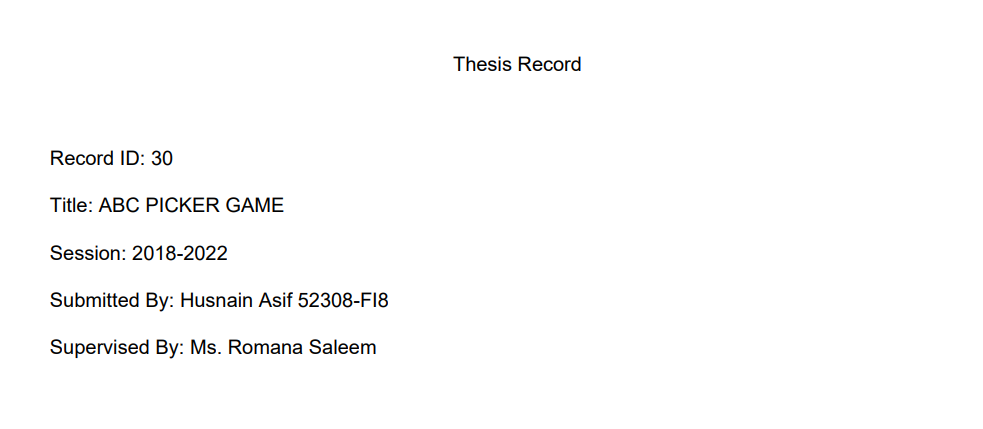


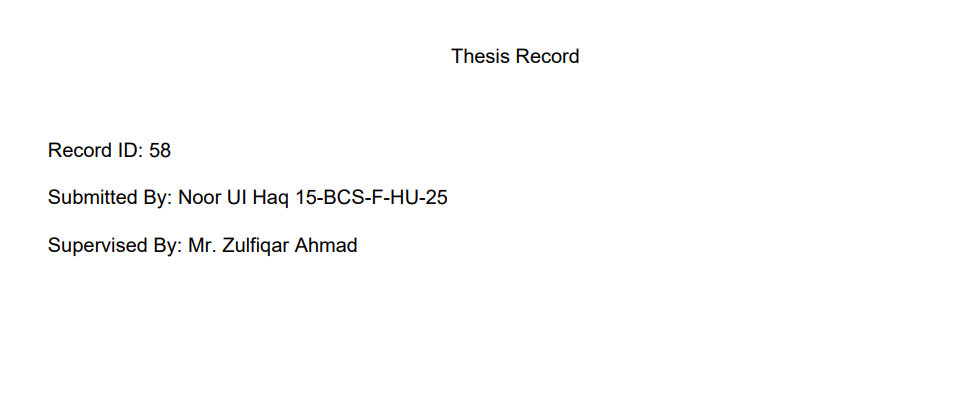
Figure 5.7: Search by Supervisor Name

## Download PDF

To download and save a PDF report in the web application, users select specific criteria such as an academic session through the user interface. The application then queries the SQL database to retrieve the relevant data. Using a PHP library FPDF, the application generates a PDF report summarizing the fetched data. This PDF report is automatically made available for download. Users can save the report for offline use or print it for physical distribution. This functionality ensures that all necessary information is documented and easily accessible in a professional format.

****





**6.1. Reference manual**

**6.1.1. Figure 5.1 Initial Interface**

The initial interface welcomes users with a title and provides options to:

* Browse files
* Upload files

Search using the navigation bar, which includes checkboxes for searching by:

* Name
* Title
* Session
* Supervisor name

**6.1.2. Figure 5.2 Image Upload and Processing**

Users can upload an image, which is then processed by the Image-to-Text Conversion module.

**6.1.3. Figure 5.3 Extracted Text Display**

The extracted text is displayed on the interface, allowing users to:

* Review the text
* Save it to the database

**6.1.4. Figure 5.4 Saving Extracted Text**

The "Save" button allows users to save the extracted text from the uploaded image into the database.

**6.1.5. Figure 5.5 Search by Title**

The search functionality enables users to retrieve academic records by specifying the title, making it easier to locate documents from a particular academic period.

**6.1.6. Figure 5.6 Search by Session**

The search functionality also allows users to retrieve academic records by specifying the session, making it easier to locate documents from a particular academic period.

**6.1.7. Figure 5.7 Search by Supervisor**

This search feature enables users to find theses and academic records associated with a specific supervisor, facilitating quick access to documents supervised by particular faculty members.

**6.2. Future scope**

* **User Authentication:** Add user login and role-based access control
* **Image Editing:** Allow users to edit extracted text and images
* **Record Versioning:** Maintain version history for each record
* **Advanced Search:** Implement fuzzy search, faceted search, and filtering
* **Integration with Other Systems:** Integrate with existing databases, CRM, or ERP systems
* **Security and Compliance:** Enhance security measures and ensure compliance with regulations
* **Mobile Optimization:** Ensure a seamless user experience on mobile devices
* **AI-powered Features:** Explore AI-powered features like auto-tagging, entity recognition, and sentiment analysis

## 

## REFERENCES

1. Adams, R., & Weis, D. (2019). Enhancing Digital Records Management: A Case Study of Integrating Electronic Theses and Dissertations. The Journal of Academic Librarianship, 45(6), 102042. [DOI: 10.1016/j.acalib.2019.102042]
2. Guercio, M., Candela, L., & Manghi, P. (2020). Assessing the FAIRness of ETDs in Institutional Repositories: The CORE-ETD Service. In Proceedings of the International Conference on Dublin Core and Metadata Applications (Vol. 2020, pp. 111-115). DCMI.
3. Metson, G. S., & Brody, T. (2018). Archiving the Digital Scholarly Record: A Case Study of Scholarly Journal Archiving. D-Lib Magazine, 24(7/8). [DOI: 10.1045/july2018-metson]
4. Schöpfel, J., & Prost, H. (2013). Degrees of openness: Access restrictions in electronic theses and dissertations. Grey Journal (TGJ), 9(1), 31-44.
5. Thomas, W. K. (2019). A Systematic Review of Electronic Theses and Dissertations Repositories. College & Research Libraries, 80(2), 197-222. [DOI: 10.5860/crl.80.2.197]
6. Todman, J., Smith, L., & Sparks, S. (2016). Curating the Digital Scholarly Record: A Roundtable Discussion. Proceedings of the Association for Information Science and Technology, 53(1), 1-4. [DOI: 10.1002/pra2.2016.145053010010]
7. Yan, Y. (2020). Building a Digital Archive of Student Theses and Dissertations: A Case Study from University of Houston Libraries. Information Technology and Libraries, 39(1), 5-13. [DOI: 10.6017/ital.v39i1.11891]
8. Zhang, L. (2018). Electronic Theses and Dissertations (ETDs): A Review of Open Access Issues and Challenges. The Electronic Library, 36(4), 658-677. [DOI: 10.1108/EL-12-2017-0255]
9. Zhang, Y., Wu, J., & Ma, J. (2018). A Digital Theses and Dissertations (ETDs) System for Modern Universities. Journal of Convergence Information Technology, 13(2), 94-104.
10. Zhou, Y., Zhao, D., & Yang, Y. (2017). Research on the Application of Electronic Theses and Dissertations (ETDs) in Higher Education Institutions. Journal of Information, 38(10), 190-193.