Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

System for secure document digitization and maintenance.

Group No. 15

Members
Sujoy Dev-17102036
Rashmi Shetty-17104070
Priya Naik-17104021

Project Guide and Coguide
Dr. Sameer Nanivadekar, Prof. Kiran B. Deshpande

Contents

ABSTRACT

- One of the challenges faced by every corporate industry is maintenance of records, mainly non-digitized type, i.e. hard copies and prints
- While a few of these documents are digitized, the are some which are very crucial in nature and require high level of maintenance, such as, ration card, marks sheet, or etc. Furthermore there are billing invoices.
- Storing these essential documents on a server can be a solution but, it requires manual labour and verification, which might lead to the misplacement of data during the process.
- Our application can act as a maintenance provider by extracting essential data from these documents and storing them in data structures.
- These data structures can be later used for performing semantic searches on the documents, which will have been already tagged by our software application.
- Semantic searches will be extremely useful while fetching the required documents using only important tags such as <title>,<document_holders_name>,<date>, etc. This will be useful for quickly tracking the documents and the data
- This system will reduce the amount of issues of traditional document maintenance which involves manual verification of data by the human eye, since once the data is extracted it can be verified and tested against a set of rules
- These rules will help us in locating document discrepancies.

INTRODUCTION

- In this project we present the modelling, implementation and development of a for secure document digitization and maintenance of documents.

- The proposed software application will be used to train and test the model by supplying images/pdfs (converted to images),

- Through this project we will provide semantic search facilities for images by extracting keyphrases/data via Natural Language Processing.

- Furthermore a cross platform application can be built on the same technology for scanning documents and extracting data from them and storing them for later use, on the go.

PROBLEM DEFINITION

- To create system for secure digitization and maintenance of documents. Institutions/users must be able to upload documents(expense bills etc.) and the important data from these documents must be extracted and stored for semantic search later. If the data is not extracted from the documents the users must be able to select the data themselves. Store everything in a centralized repository. Send the users insights about their document ,i.e., expiration, validity, etc.

OBJECTIVES

- Reduce document maintenance and information retrieval efforts from documents such as invoices, purchases orders, maintenance records, etc. by developing an ecosystem by using **machine learning, Computer Vision** for document data identification, extraction and validation.
- Gain necessary information from the documents and store the data as document format for easy storage, matching and verification of data.
- Filter through numerous, large document sets within a matter of seconds by entering the keywords to be searched only, such as <date>, <ownerInfo>.
- Update the knowledge-base regularly to gain newer information from the documents.
- Platform to train and test with newer document types.
- Reduce paper dependency and go digital.
- Create a cross platform desktop application for packaging the technology

LITERATURE REVIEW -1

- **Title**: *Document Specific Supervised Keyphrase Extraction With Strong Semantic Relations*
- **Authors**: Huiting Liu ,Wang Peng Zhao, Xindong Wu
- **Findings**: Keywords can be extracted via natural language processing after the text has been extracted from the documents. Keyphrases provide semantic metadata that summarize and characterize documents.
- **Advantages:** Semantic relation can be extracted by NLP application
- **Disadvantages:** No document summarization
- **Publication**: IEEE - 22nd October 2019

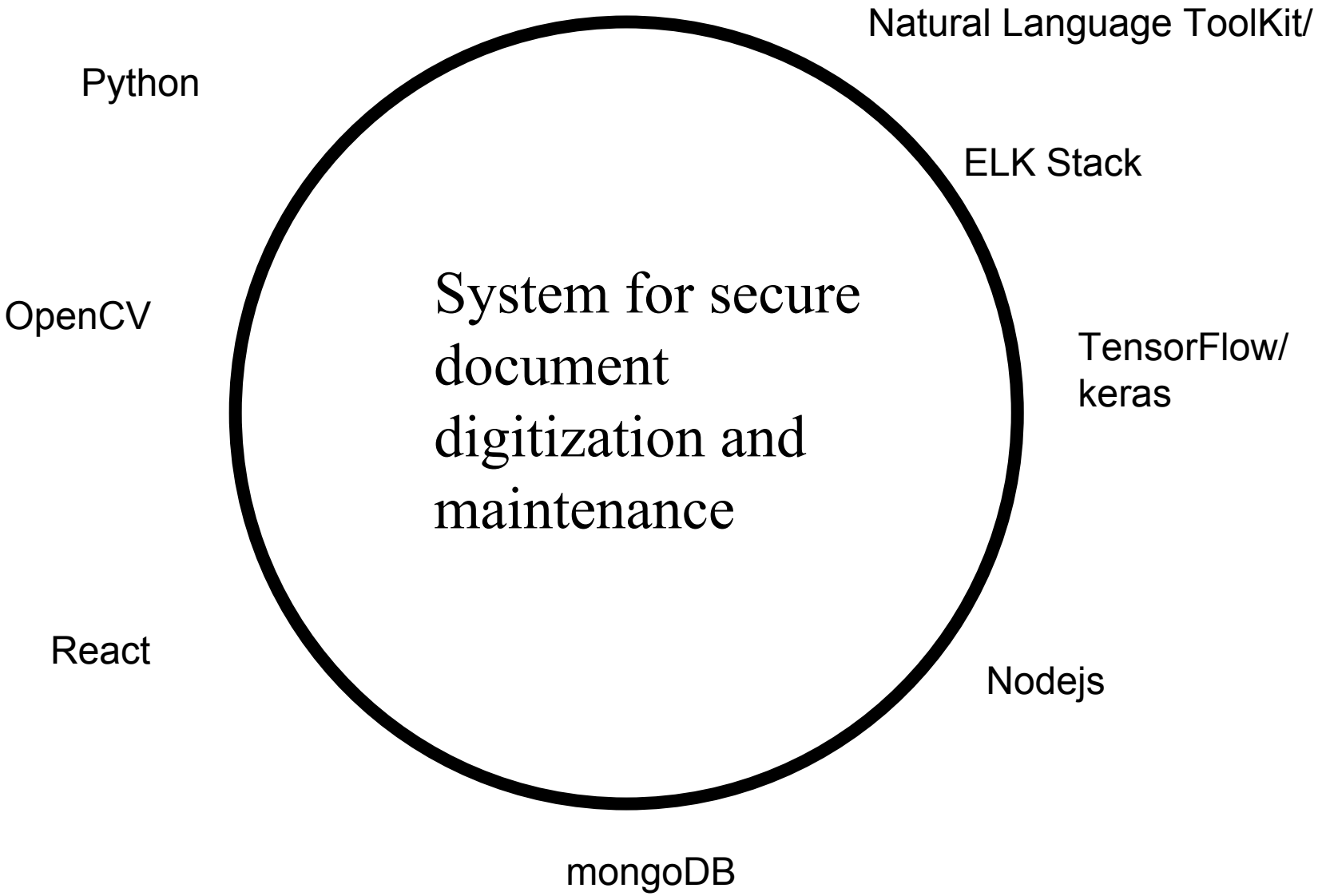
LITERATURE REVIEW -2

- **Title**:*Keyword Extraction Through Contextual Semantic Analysis of Documents*
- **Authors:** Terry Ruas, William Grosky
- **Findings:** Traditional approaches make use of techniques that rely on analyzing just the syntactic aspect of texts, ignoring the meaning they convey and more importantly, the semantic effect of oneone word over another. Hence we ought to use the following two approaches, the first approach extends the concept of Word Sense Disambiguation (WSD), and the second approach enhances the theory behind traditional lexical chains. These applied techniques also consider distinct levels of abstraction with respect to the meanings of words, in addition to the context in which they appear.
- **Advantages:** Word sense disambiguation and traditional lexical chains and give meaning to the context
- **Disadvantages:** Synset provided by the different algorithms(BSD & FLC) contaminate the score mediaWiki categories
- **Publication**: 2017 Conference: the 9th International Conference Research Gate

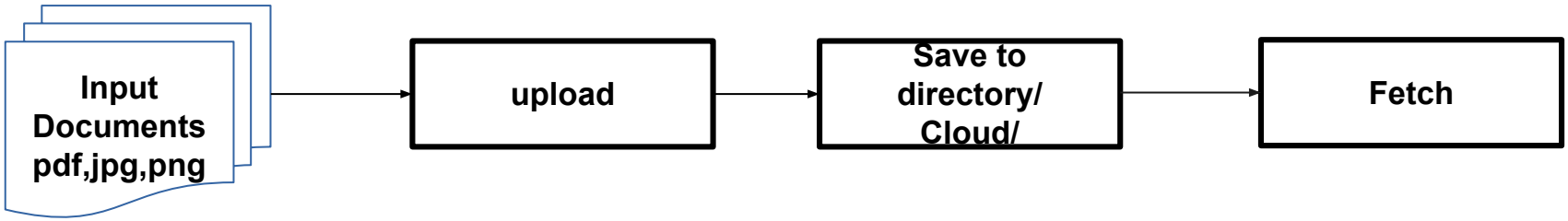
LITERATURE REVIEW -3

- **Title**: *A System for Handwritten and Printed Text Classification*
- **Authors:** Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala
- **Findings:** Various features described in her are extracted from these word images and these are used to analyze the differences between machine print and handwritten text. There is no clear separation between machine printed and handwritten text density values, but pixel density can be used to augment the classification efficiency.
- **Advantages:** Can classify between machine printed and handwritten alphabets with 98.6% accuracy
- **Disadvantages:** Only english language support for classification
- **Publication**: 2017 UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim)

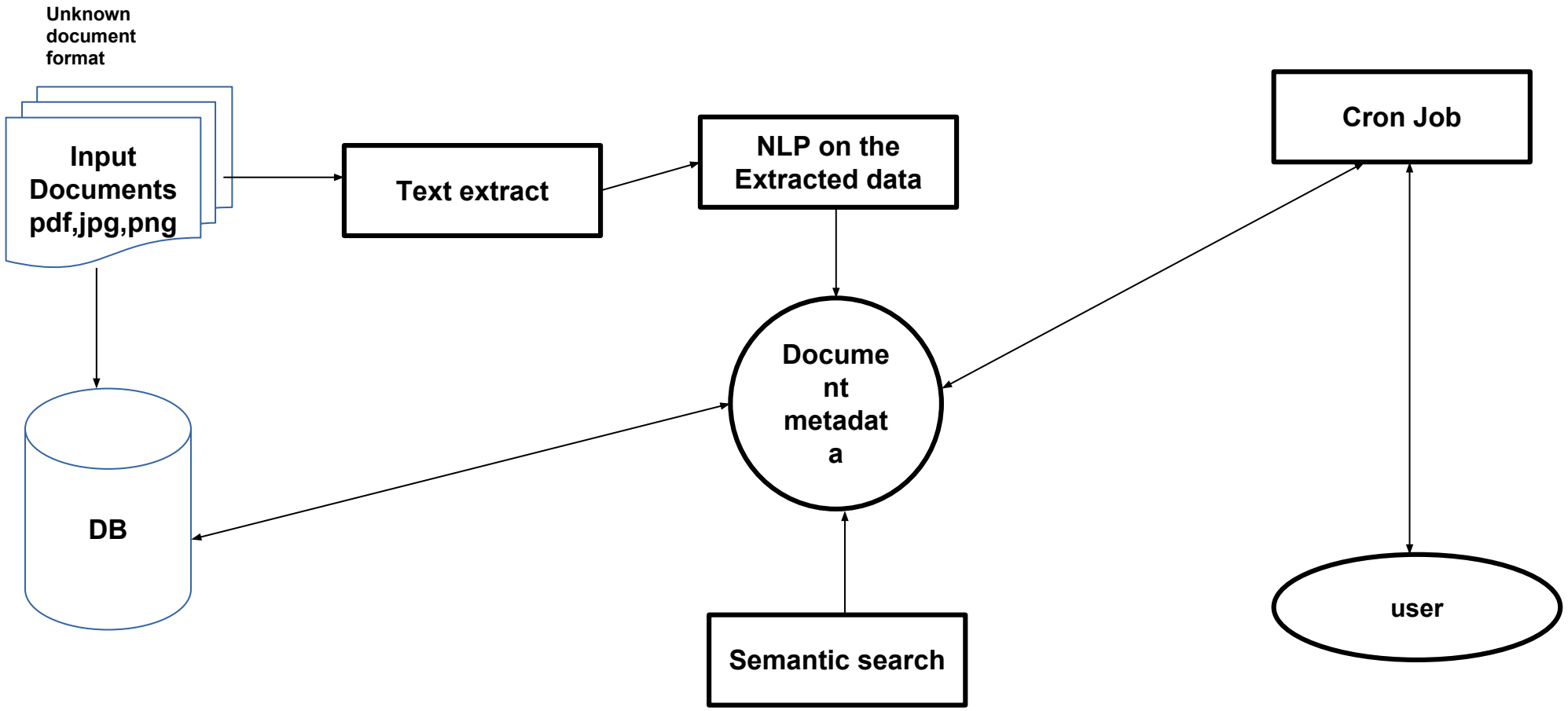
TECHNOLOGY STACK



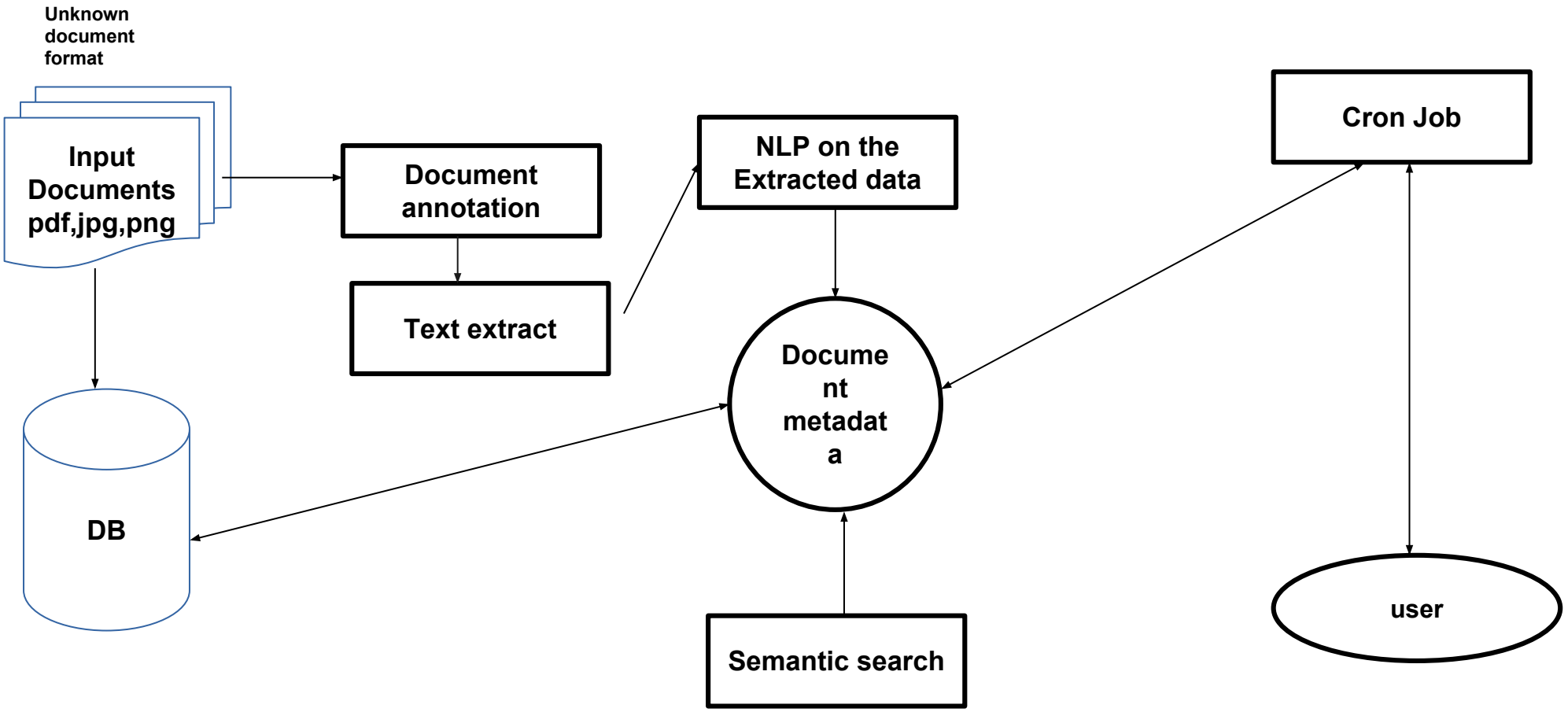
EXISTING SYSTEM WORKING/ARCHITECTURE



PROPOSED SYSTEM WORKING/ARCHITECTURE



PROPOSED SYSTEM WORKING/ARCHITECTURE



Prototype Design

https://www.figma.com/proto/hYiyfAvpHyFXRFP7kgH5nS/Docs?node-id=0%3A1&scaling=min-zoom

Plan Of Paper Publication

- IEEE 2021 6th International Conference for Convergence in Technology

- Introducing a notification system to inform users of data collected from the images by using fault tolerant messaging queues like kafka, rabbitMQ.

- Users will be provided access to an online annotation tool which will help them extract extra information from their docs using named entity recognition and at the same time will be used to save the images with the annotation for training the model at regular intervals.

- Elk stack to real time log/data analysis, Dynamic TensorFlow/Keras training capabilities to improve the model, ie, it will work as an opensource project where others can contribute by annotating their custom document formats

FUTURE SCOPE

- Secure sensitive information/document by custom encoding/decoding.

- Further this concept can be implemented into an mobile expense tracking application, where in the user can scan any type of invoice and it can be digitized.

CONCLUSION

- Thus with the application of effective text extraction from pdf, image files and by performing some Natural Language Processing on the extracted data we can extract key-pairs, keywords, keyphrases from the non digitized documents perform digitization with the new and refined data. This data can be pushed to a central database for global access. This will be used to provide users with reliable insights. Through this approach we will achieve true digitization of documents.

REFERENCES

[1] *https://ieeexplore.ieee.org/document/8879476*
Huiting Liu ,Wang Peng Zhao, Xindong Wu. *Document Specific Supervised Keyphrase Extraction With Strong* Semantic *Relations. IEEE - 22nd October 2019*

[2]*https://www.researchgate.net/publication/324074602_Keyword_Extraction_Through_Contextual_Semantic_Analysis_of_Documents*
Terry Raus, William Grosky. Keyword Extraction Through Contextual Semantic Analysis of Documents. 2017 Conference: the 9th International Conference Research Gate

[3]*https://ieeexplore.ieee.org/document/8359046*
Bala Mallikarjunarao Garlapati, Srinivasa Rao Chalamala. *A System for Handwritten and Printed Text Classification. 2017 UKSim-AMSS 19th International Conference on Computer Modelling & Simulation (UKSim)*

Thank You...!!