perplexity

1. Introduction

This project focuses on **digitizing unstructured documents** (PDFs or scanned uploads) and empowering users through translation, voice assistance, and an interactive chatbot. Users can upload documents, have them digitized and structured, and interactively learn about any word or sentence by clicking on it—triggering a right-side panel with explanations, translations, or voice playback. The solution will be delivered as a **web dashboard** for maximum accessibility and feature richness. A Chrome extension is considered for future scope.

2. Dataset Information

a) Dataset Introduction

We will use open-source OCR, translation, and language understanding datasets to train and evaluate the system's ability to extract, translate, and explain document content. These datasets are essential for robust digitization and user education features.

b) Data Card

Field	Value
Size	10,000+ documents, 100,000+ sentences (varies by dataset)
Format	PDF, scanned images (JPG/PNG), text (TXT/JSON)
Data Types	Document images, extracted text, translations, audio
Languages	English, Hindi, Telugu, Spanish
Domains	Financial, legal, general

c) Data Sources

- **OCR:** BhasaAnuvaad, IIIT-ILOCR
- **Translation:** BhasaAnuvaad, CVSS
- **Language Understanding:** SQuAD, IndicNLP

d) Data Rights and Privacy

- All datasets are open-source and used for research/non-commercial purposes, with no PII included.
- User-uploaded documents will be processed securely, with encryption and access controls.

- The system will comply with GDPR and similar regulations by anonymizing data, obtaining user consent, and allowing data deletion on request.

3. Data Planning and Splits

- **Loading:** Scripts to ingest PDFs/images, convert to text using OCR.
- **Preprocessing:** Clean extracted text, segment sentences, detect language, and align with translations.
- **Managing Data:** Use version control (e.g., DVC), store metadata (language, domain, source).
- **Splitting:** 80% training, 10% validation, 10% test, stratified by language and document type.

4. GitHub Repository

- **Link:** (To be created, e.g., github.com/your-org/doc-digitizer)
- **Folder Structure:**
  - `/data/` — Data scripts and sample datasets
  - `/ocr/` — OCR and preprocessing code
  - `/translation/` — Translation and explanation modules
  - `/voice/` — Voice synthesis and playback
  - `/dashboard/` — Web dashboard code
  - `/docs/` — Documentation and diagrams
  - `/tests/` — Unit/integration tests
  - `README.md` — Overview, install, usage, contribution

5. Project Scope

a) Problems

- Unstructured documents are hard to search, analyze, or understand.
- Language barriers and complex terms hinder user comprehension.
- Lack of interactive, accessible tools for document education.

b) Current Solutions

- Basic OCR tools extract text but lack translation or explanation features.
- Some web apps offer translation, but not interactive explanations or voice assistance.
- No seamless browser-based solution for real-time document education.

c) Proposed Solutions

- Digitize and structure documents using OCR and NLP.
- Provide instant translation, voice playback, and chatbot explanations for any selected text.
- Deliver as a web dashboard for easy access and integration (Chrome extension in future scope).

6. Current Approach Flowchart and Bottleneck Detection

Flowchart:

1. User uploads PDF/scanned document
2. OCR extracts text and structure
3. Text is segmented into sentences/words
4. User clicks a word/sentence → right panel pops up
5. Panel shows translation, explanation, and voice playback
6. Chatbot available for further questions

Bottlenecks:

- OCR errors with poor-quality scans
- Translation/explanation accuracy for domain-specific terms
- Real-time performance for large documents
- UI responsiveness and accessibility

Improvements:

- Use advanced OCR (e.g., Tesseract, BhasaAnuvaad)
- Fine-tune translation/explanation models on domain data
- Optimize UI for speed and accessibility

7. Metrics Objectives and Business Goals

- **Metrics:** OCR accuracy, translation BLEU score, user engagement (clicks, time spent), response latency, user satisfaction surveys
- **Objectives:**
  - Achieve >90% OCR accuracy on test set
  - BLEU >25 for translations
  - <2s latency for panel pop-up
  - 80%+ positive user feedback
- **Business Goals:**
  - Reduce manual document review time

  - Increase user understanding and satisfaction
  - Enable compliance with multilingual and accessibility standards

8. Failure Analysis

- **Risks:**
  - Poor OCR/translation for low-quality or handwritten documents
  - Privacy breaches if user data is mishandled
  - UI/UX issues causing user frustration
- **Mitigation:**
  - Use robust models and regular retraining
  - Encrypt and access-control all user data
  - Iterative UI testing and feedback

9. Deployment Infrastructure

- **Infrastructure:**
  - Cloud-based OCR/NLP/translation services (e.g., AWS, GCP)
  - Secure storage for user documents
  - Web dashboard (React/Vue)
  - RESTful APIs for backend processing
- **Supported Platforms:**
  - All major browsers (dashboard)
- **Flowchart:**
  - User → Dashboard → Backend API → OCR/NLP/Translation → Output Panel

10. Monitoring Plan

- Monitor OCR/translation accuracy, API latency, error rates, and user interactions
- Log and audit access to user documents for security
- Collect user feedback for continuous improvement

11. Success and Acceptance Criteria

- 90% OCR accuracy and BLEU >25 on test data
- <2s response time for panel pop-up
- 80%+ user satisfaction in surveys
- No major security/privacy incidents

12. Timeline Planning (12 Weeks)

Week	Milestone/Task
1	Project scoping, requirements, team setup
2	Dataset acquisition, initial data audit
3	Data pipeline: loading, preprocessing scripts
4	OCR module development & testing
5	Translation & explanation module setup
6	Voice assistance module setup
7	Dashboard UI/UX design, clickable text prototype
8	Integrate OCR, translation, voice with dashboard
9	Chatbot integration, right-panel pop-up logic
10	End-to-end testing, bug fixes, accessibility review
11	User testing, feedback, final optimizations
12	Deployment, documentation, project wrap-up

- **Buffer:** 2-3 days per major task for risk mitigation and review.
- **Chrome extension:** Planned for future scope after dashboard MVP.

13. Additional Information

- **UI/UX:** The dashboard will feature a document viewer with clickable text. When a user clicks a word or sentence, a right-side panel will display translation, explanation, and voice playback. The chatbot will be accessible from the panel for further queries.
- **Accessibility:** Voice assistance and screen reader compatibility will be prioritized.
- **Scalability:** The backend will be designed to handle multiple concurrent users and large documents efficiently.
- **Documentation:** Comprehensive user and developer documentation will be maintained in the repository.

Summary:

- Deliver a robust web dashboard for document digitization and user education in 12 weeks.
- Prioritize core features (digitization, translation, voice, chatbot, interactive panel).
- Plan Chrome extension as a future enhancement after MVP delivery.

⁂

1. https://www.census.gov/library/visualizations/interactive/decennial-census-microfilm-scanning-progress.html
2. https://bmiimaging.com/blog/paper/paper-scanning-timelime/

3. https://www.gwbhs.org/documents/2012/11/si-644-digitization-project-plan.pdf/
4. https://www.youtube.com/watch?v=3dC1CCmh7q8
5. https://bp-ms.co.uk/7-steps-to-a-successful-document-scanning-project/
6. https://www.youtube.com/watch?v=RX40s6GJJHo
7. https://recordsmanagement.tab.com/document-imaging/how-to-plan-a-successful-document-imaging-project-part-2/
8. https://www.reddit.com/r/Archivists/comments/192orgy/equipment_recommendations_for_digitization_project/
9. https://www.ifla.org/files/assets/preservation-and-conservation/publications/digitization-projects-guidelines.pdf