Hadith Dataset Project: Strategy and Project Plan

Objective

The goal of this project is to collect datasets in PDF, image, text, and JSON formats for the most reliable Arabic dictionaries and Isma ur Rijal books. To ensure reliability, we will consult subject matter experts from institutions such as Darul Uloom Amjadiya, Darul-Uloom Korangi, and Binori Town. The expert-recommended books will then be sourced in digital format, ensuring necessary permissions for use in an open-source project.

Project Plan

Phase 1: Expert Consultation

1. Identify Experts
    - Shortlist at least 10 scholars or subject matter experts from well-known religious institutions.
    - Scholars should have expertise in Arabic linguistics, Hadith sciences, or Islamic history.
2. Contact Experts
    - Modes of communication: Call, email, or visit in person.
    - Objective: Obtain a list of the most reliable Arabic dictionaries and Isma ur Rijal books.
3. Document Expert Opinions
    - Create a Markdown file recording the following details:
        - Name of expert
        - Academic qualifications (degree, specialization)
        - Institutional affiliation
        - List of recommended books

Phase 2: Data Collection

4. Find PDFs of Recommended Books
    - Search archive.org and other credible online repositories for clear, high-quality PDFs.
    - If PDFs are not available online, explore alternative sources, including libraries or publishers.
5. Secure Permission for Use
    - Contact publishers or copyright holders to obtain written permission for using PDFs, images, and text in an open-source project.

    - If necessary, draft a formal request letter explaining the project's purpose and scope.

Phase 3: Dataset Structuring

6. Convert Data into Structured Formats
    - Extract text from PDFs using OCR (Optical Character Recognition) tools.
    - Convert extracted text into structured JSON format for easy indexing and searchability.
    - Organize images and text efficiently for integration into the open-source project.
7. Quality Assurance & Validation
    - Verify dataset accuracy through cross-referencing with printed editions.
    - Ensure proper metadata tagging (e.g., book name, author, year, topic classification).

Expected Outcomes

✅ A well-documented list of reliable books backed by expert validation.
✅ A dataset including PDFs, images, text, and structured JSON files.
✅ Legal permissions ensuring the open-source usability of the dataset.
✅ A foundation for future Hadith and Arabic linguistics projects.

Next Steps

- Begin expert outreach and compile recommendations.
- Simultaneously search for PDFs and draft permission request letters.
- Once permissions are obtained, initiate data extraction and structuring.