End-to-End Open Digital Archiving Framework for Regional Language Literary Heritage: Lessons from Kaniyam Foundation

Authors: Anwar Latiff, Kaniyam Foundation (கணியம் அறக்கட்டளை)

Venue: Web Archiving and Digital Libraries Workshop (WADL 2026), held in conjunction with ACM/IEEE Joint Conference on Digital Libraries (JCDL 2026)

Abstract

Preserving regional language heritage requires integrated technical workflows and open-access frameworks across the entire digital resource life cycle. This paper presents the end-to-end digital archiving pipeline established by Kaniyam Foundation for Tamil literary and historical texts. We detail our practical workflow spanning physical acquisition, image post-processing, optical character recognition (OCR) refinement, public domain dissemination via the Internet Archive, and structured metadata cataloging on Wikisource and Wikidata. We highlight how the strategic application of Creative Commons licensing (CC BY-SA 4.0) alongside accessibility-first ebook standards ensures long-term preservation, legal clarity, and discoverability for under-represented languages in global digital library infrastructures.

1. Introduction

Digital libraries and web archiving initiatives have vastly expanded access to global knowledge, yet non-Western and regional language literary heritage remains severely under-represented in open-access digital repositories. Rare manuscripts, historical print editions, and community literary archives in languages like Tamil face risks of physical deterioration and digital erasure due to the lack of standardized, low-resource digital preservation pipelines.

Kaniyam Foundation (கணியம் அறக்கட்டளை) was established to address this gap by constructing an end-to-end, community-driven digital archiving ecosystem. This paper outlines our operational framework—from initial artifact acquisition to global web archiving and semantic web integration—offering a reproducible model for regional digital library projects.

2. The Digital Resource Life Cycle Pipeline

Our archiving framework structures the digital preservation process into five sequential phases, balancing image quality, textual accuracy, legal sustainability, and global interoperability.

Lifecycle Phase	Key Tools & Technologies	Primary Objectives & Outputs
1. Physical Capture & Rights Management	CZUR Scanners, CC BY-SA 4.0 Deeds	Non-destructive scanning, legal permission acquisition, rights declarations.
2. Image Post-Processing	ScanTailor Advanced, IrfanView	DPI normalization, page splitting, skew correction, background cleaning.
3. Text Extraction & Formatting	Tesseract OCR, Regex, LibreOffice Writer, Sigil	Tamil Unicode conversion, structured semantic HTML/ePub generation, WCAG accessibility.
4. Web Archiving & Storage	Internet Archive API, FFmpeg, Python Scripts	Long-term raw archival storage (PDF/EPUB), media capture, public domain access.
5. Semantic Web & Proofreading	Wikisource, Wikidata, Tamil 99 Keyboard	Crowdsourced proofreading, linked open data metadata (DDC, author relations, P31/P50 mapping).

2.1 Physical Acquisition and Rights Clearance

The preservation lifecycle begins with acquiring physical access to rare print editions, palm-leaf manuscripts, and legacy collections (such as the works of Salai Ilanthiraiyan and A. Sivalingam archives). To guarantee public availability without legal encumbrance, Kaniyam Foundation establishes formal permission deeds securing Creative Commons Attribution-ShareAlike 4.0

(CC BY-SA 4.0) licensing or verifying public domain status prior to digitization.

2.2 Image Processing and Quality Standardization

Physical scans are processed through ScanTailor Advanced to normalize output parameters: 300+ DPI resolution, automated page boundary detection, de-skewing, and margin alignment. These steps eliminate optical artifacts that impair downstream optical character recognition (OCR) engines for non-Latin script layouts.

2.3 Text Processing, OCR, and Accessibility

Processed images undergo OCR pipeline processing customized for Tamil script rendering. Raw text output is refined using standard Tamil Unicode character mappings and regular expression (Regex) cleanup routines in tools like Notepad++ and LibreOffice Writer. Document formatting adheres to Web Content Accessibility Guidelines (WCAG) and EPUB 3 standards, using semantic heading structures (<h1>, <h2>) to support screen readers and low-vision readers.

2.4 Web Archiving and Repository Ingestion

Archival artifacts are ingested into public repositories using automated upload pipelines. High-resolution PDFs and EPUB files are archived on the Internet Archive with full text-search capabilities. For audiovisual heritage (e.g., historical radio broadcasts), automated capture workflows using FFmpeg and VLC are deployed for stream archival.

2.5 Linked Open Data and Community Proofreading

To ensure long-term discoverability, digital texts are integrated into open web knowledge networks:

- **Tamil Wikisource:** Provides a web-based, crowdsourced platform for line-by-line proofreading against original page scans.
- **Wikidatification:** Structured metadata entries are created on Wikidata, linking digitized texts to author profiles, Dewey Decimal Classification (DDC) numbers, publication dates, and institutional repository identifiers.

3. Challenges and Key Lessons

Operating a regional digital archiving pipeline presents distinct technological and resource challenges:

- **Script Complexity and OCR Accuracy:** Traditional OCR solutions often struggle with

conjunct Tamil typography and vintage print fonts, necessitating human-in-the-loop verification via crowdsourced platforms like Wikisource.

- **Open Licensing Advocacy:** Establishing open access requires proactive engagement with copyright holders, heirs, and regional institutions to adopt Creative Commons framework agreements.
- **Cross-Platform Synchronization:** Maintaining metadata consistency across local catalogs, Internet Archive records, and Wikidata items requires structured automation scripts.

4. Conclusion and Future Work

Kaniyam Foundation's end-to-end archiving framework demonstrates that high-quality, legally clear, and accessible digital preservation of regional languages is feasible using open-source tools and community networks. Future work includes expanding automated metadata ingestion pipelines, integrating AI-assisted Tamil OCR workflows, and advocating for broader adoption of open digital library standards across regional institutions.

References

- Creative Commons. (2026). *Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)*.
- Internet Archive. (2026). *Digital Library of Free & Borrowable Books*. archive.org.
- Wikidata Contributors. (2026). *Wikidata: Property definitions for literary and archival works*. wikidata.org.
- World Wide Web Consortium (W3C). (2026). *EPUB 3.3 Accessibility Requirements*.