https://doi.org/10.52027/18294685-aa.2.25-12

ALVARD GRIGORYAN, AREN YEGHIAZARYAN, DAVIT ISPIRYAN, DAVIT MELIKSETYAN, DAVIT NAVASARDYAN, DAVIT NAZLUXANYAN, HAKOB SHAHNAZARYAN, LEVON ANANIKYAN, MARI ASLANYAN, MARIA KATVALYAN, MILENA ARSHAKYAN, SONA KHOSROVYAN, SUREN SAGHATELYAN, VAHE KARAPETYAN, VALERI SEVYAN
TUMO Armenia

AUTOMATED QUALITY ASSESSMENT AND RESTORATION OF SCANNED ARMENIAN NEWSPAPERS

Բանալի բառեր՝ Հայկական թերթեր, փաստաթղթերի վերականգնում, որակի գնահատում, պատկերների բարելավում, թվայնացում, համակարգչային տեսողություն, OCR (օպտիկական նշանների ճանաչում), թվային պահպանություն։
Keywords: Armenian newspapers, document restoration, quality assessment, image enhancement, digitization, computer vision, OCR, digital preservation.

Introduction

Cultural heritage institutions have, over the past decades, engaged in large-scale digitization programs to preserve their collections and make them widely accessible online. Digitization reduces the handling of fragile originals and democratizes access to historical materials, enabling researchers and the general public to explore archives remotely. Beyond its technical dimension, it has also become a form of patrimonialization, transforming digital reproductions into heritage objects in their own right and reshaping curatorial practices [Bermès, 2020]. Within this context, artificial intelligence (AI) has emerged as a key component of digitization and curation workflows, supporting tasks such as optical character recognition (OCR), metadata extraction, image classification, and semantic linking of related documents across large digital collections [Rehm et al., 2020]. Large-scale initiatives such as NewsEye[4], the e-NDP project, or the Mekhitarist Archives of Venice have demonstrated the potential of end-to-end AI pipelines for heritage materials, combining document layout analysis, handwriting recognition, linguistic annotation, and scholarly indexing [Jolivet, Terriel & Canteaut, 2025; Vidal-Gorène, 2023]. These integrated approaches have fostered the development of "augmented" digital libraries, enabling more effective navigation, search, and analysis of extensive text and image corpora [Gasparini & Kautonen, 2022][5].

[4] https://www.newseye.eu/
[5] Within the scope of Armenian studies, examples of AI-augmented document versions include the Mekhitarist catalogue of Venice, fully extracted through AI with semantic understanding of its content (see: https://catalog.mechitar.org/), and

However, deploying AI systems in real library environments raises challenges that go beyond experimental performance metrics. One major issue lies in the heterogeneity of digitized materials, especially when collections span several decades of scanning technologies, file formats, and imaging standards. Such variation significantly affects model robustness, making it difficult to generalize across entire archival datasets. For instance, OCR accuracy often drops sharply on degraded documents, even when models are trained on large, generic corpora, notably due to segmentation and layout analysis issues. Institutions thus face a strategic question: should they re-digitize their holdings to benefit from modern imaging standards and clean outputs, or invest in automated post-processing methods to enhance or to restore existing scans? The first approach is costly and time-consuming, while the second requires reliable computer vision models capable of handling degraded, noisy, or misoriented images.

Historical newspapers exemplify this challenge: they often display uneven illumination, torn pages, scanning blur, or skewed alignment, in addition to extremely large formats. Manual correction, though possible, is impractical at scale. Building on previous developments in document layout analysis and AI-assisted curation, the present work focuses on automating post-processing and quality assessment in historical newspaper digitization. Its goal is to improve the uniformity and usability of existing scans through lightweight procedures for page detection, orientation correction, and basic enhancement. Experiments are conducted on digitized Armenian newspaper collections from the National Library of Armenia, characterized by variable image quality depending on the source and scanning period[6].

Recent research on document restoration increasingly focuses on generative methods that attempt to reconstruct the missing or damaged parts of historical documents in order to improve OCR or HTR results. Generative Adversarial Networks (GANs) and diffusion-based models have shown impressive visual capabilities in recreating plausible document layouts or even predicting the original appearance of degraded materials [Vidal-Gorène & Camps, 2024; Ranjan & Ravinder, 2024; Yang *et al.*, 2025; Zhang *et al.*, 2024]. However, such models may introduce artefacts or hallucinated content, raising concerns about the authenticity of the restored information [Vidal-Gorène, 2025; Yang et al., 2025]. For example, GAN-generated manuscript pages can convincingly imitate the visual style of the source while producing meaningless or incorrect text sequences, and diffusion models like DiffHDR can generate entire missing characters or blocks [Vidal-Gorène & Camps, 2024; Yang *et al.*, 2025].

In contrast, this work adopts a more operational perspective, relying on supervised computer-vision techniques such as classification, object detection, and semantic segmentation, focusing only on the «damaged» material. These models have proved effective for structural and quality analysis of digitized documents, where identifying page boundaries, orientation, or degradation patterns is often more relevant than full image generation. The proposed approach therefore aims to standardize existing scans through automated detection, cropping, and perspective correction, facilitating downstream tasks such as OCR and metadata extraction while avoiding the cost and complexity of large-scale re-digitization. In this perspective, AI serves as a practical tool for improving the consistency and usability of heterogeneous digitization outputs, particularly in long-term national programs where equipment and standards have evolved over time [Gasparini & Kautonen, 2022].

---

the Dulaurier collection at the French National Library [Vidal-Gorène *et al.*, 2025] (see: https://github.com/calfa-co/datalab-dulaurier).

[6] https://tert.nla.am/

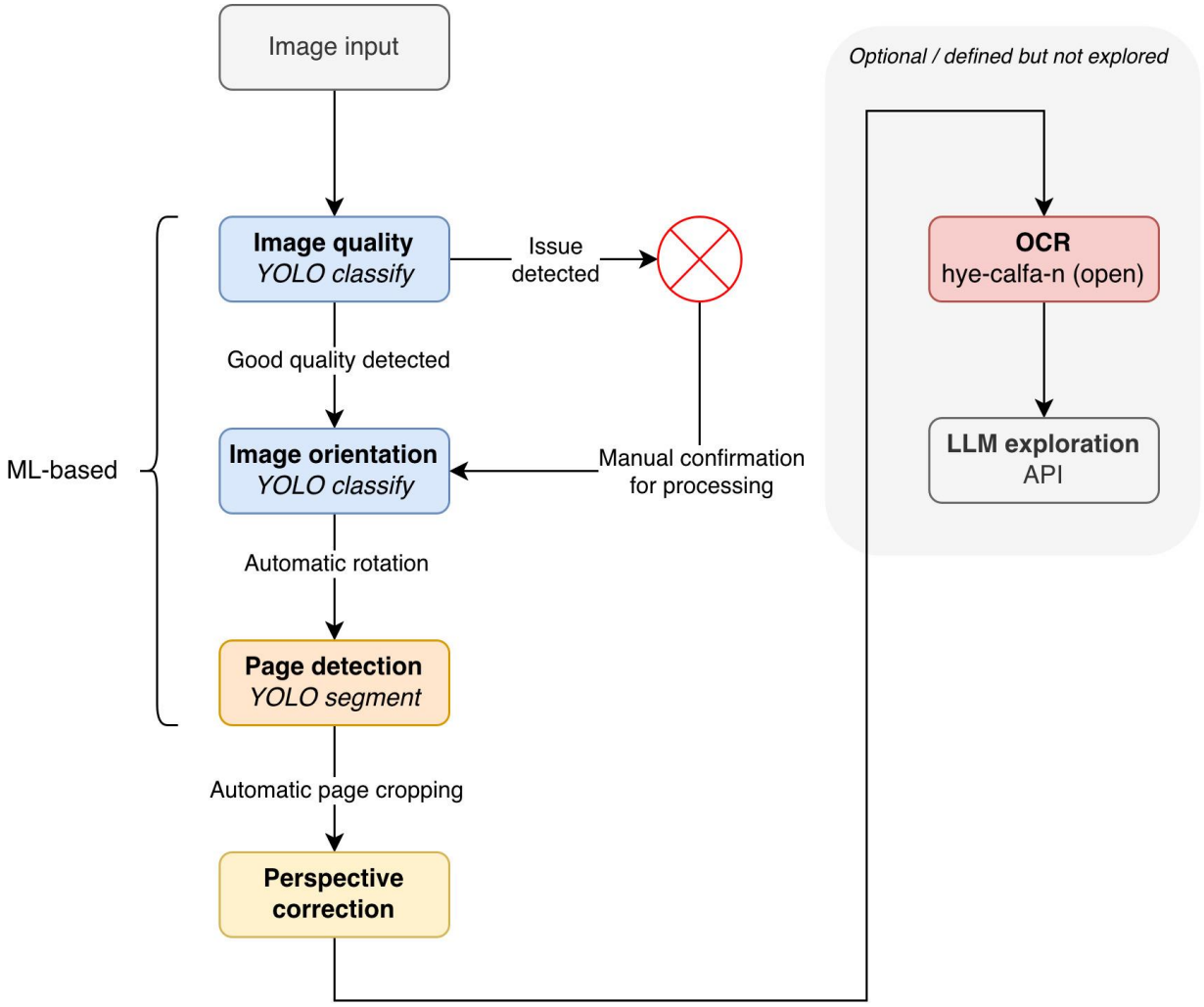


Figure 1: Overview of the document preprocessing pipeline

The process begins with image quality assessment and orientation classification using YOLO-based models. If the image is of sufficient quality, the system automatically detects and crops document pages before applying perspective correction using OpenCV. Optional modules, defined but not evaluated in this study, include Optical Character Recognition (OCR) using hye-calfa-n open-source model, and text exploitation through Large Language

Methodology and Dataset

The proposed processing pipeline takes a raw digitized image and sequentially processes it through several classification and detection stages. Its main objectives are to assess the quality of the digitization, verify and correct image orientation, detect and crop one or several pages within the frame, and apply a perspective correction to produce a clean, properly aligned image ready for Optical Character Recognition (OCR).

Enhancement operations such as brightness or contrast adjustment, advanced curvature correction, or shadow removal are not included in the current workflow. However, the pipeline's modular design allows such stages to be integrated easily when needed. The workflow was also extended experimentally with OCR and Large Language Model (LLM) modules (via Mistral or OpenAI APIs) for topic modeling, summarization, and article classification, although these analytical components were not formally evaluated in this study.

Three YOLOv11n models were trained to perform the core tasks: image quality classification, orientation classification, and page segmentation.

• **Image quality classification**: A YOLOv11n classify model with two classes (good quality, bad quality) was trained to assess the legibility of scans. Images were labeled as bad quality when exhibiting visual defects—such as shadows, excessive curvature, or severe physical degradation, that hindered human readability and OCR accuracy.

• **Orientation classification**: A second YOLOv11n classify model was trained with four classes corresponding to 0°, 90°, 180°, and 270° text orientations. Training data were augmented by rotating original images to ensure balanced representation across all classes.

• **Page detection**: A YOLOv11n segment model was trained with two classes (PageLeft, PageRight), each annotated as a polygon enclosing the visible page area. When only a single page was present, it was labeled PageLeft by convention.

All YOLO models were trained using an input image size of 1536×1536 pixels, chosen to preserve small textual details and ensure readability within high-resolution newspaper scans. Internal YOLO data augmentation techniques were activated to improve generalization, including scale variation, mosaic composition, and mixup blending, while horizontal or vertical flipping was deliberately disabled to maintain textual consistency. A dropout rate of 30% was applied during training to reduce overfitting and improve robustness across heterogeneous input conditions.

Once the page polygons were detected, each region was cropped and rectified using a perspective transformation computed from the polygon corner coordinates. Since YOLO segmentation outputs unordered polygon points, the detected vertices (x_i, y_i) were first rearranged into a consistent order (top-left, top-right, bottom-right, bottom-left) based on the sum and difference of their coordinates. This ordering ensures a coherent mapping between the source and destination planes before computing the transformation. The homography matrix M is then estimated from four pairs of corresponding points between the detected polygon corners (x_i, y_i) and their destination coordinates (x_i', y_i'):

$$\begin{bmatrix} x_i' \\ y_i' \\ 1 \end{bmatrix} \sim M \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}, i = 1, \dots, 4.$$

The transformation is defined by a 3×3 projective matrix M:

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = M \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, where M = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}$$

After normalization, the pixel coordinates in the rectified image are obtained as:

$$x' = \frac{m_{11}x + m_{12}y + m_{13}}{m_{31}x + m_{32}y + m_{33}}, y' = \frac{m_{21}x + m_{22}y + m_{23}}{m_{31}x + m_{32}y + m_{33}}$$

This mapping is implemented in OpenCV using the function cv2.getPerspectiveTransform() to compute M from the ordered point pairs, and cv2.warpPerspective() to apply the transformation. The resulting warped image is projected into a normalized rectangular plane that preserves the geometry of the page, minimizes perspective distortion, and produces a top-down view suitable for OCR processing.

For the OCR stage, we employed the open-source hye-calfa-n transcription model developed by Calfa and based on the Tesseract engine [Calfa, 2025]. This model automatically infers the reading order and performs text transcription. It extends Tesseract's default capabilities by incorporating data representative of 20th-century Armenian print production, including damaged and low-quality

documents, and can be run with the pytesseract python package. Supporting Classical, Western, and Eastern Armenian, it significantly reduces character and word error rates while remaining lightweight and easily deployable in large-scale heritage digitization workflows. For higher-level text exploitation tasks, the pipeline can connect to Mistral or OpenAI APIs (requiring an API key)[7]. In the case of historical newspapers, however, OCR remains particularly challenging due to the complexity of page layout analysis and the ambiguity of reading order across multi-column or irregular structures, an ongoing research problem in document analysis and digital humanities [*Karmanov et al.*, 2025]. The pipeline is designed with modularity and accessibility in mind, enabling seamless integration within a web-based application for large-scale or user-interactive processing. The overall architecture of this workflow is illustrated in Figure 1.

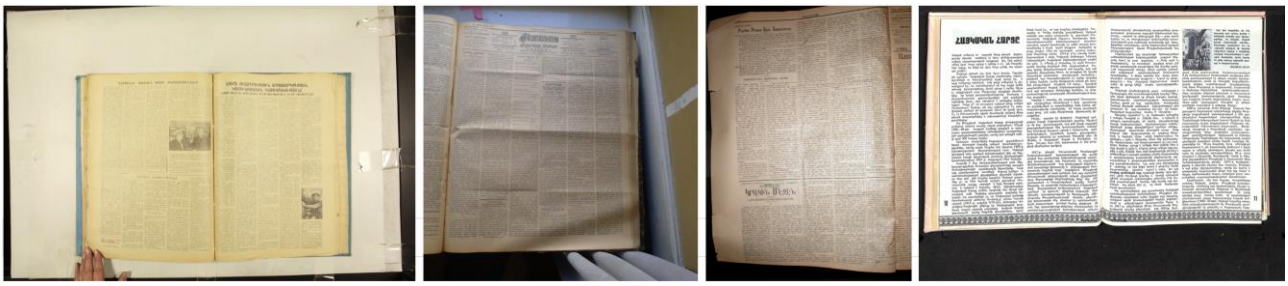


Figure 2: Dataset overview

To train the models, we assembled a representative dataset of 450 real images and 3,000 synthetic variants. The real dataset includes 300 archival scans from the National Library of Armenia, selected to reflect the diversity of digitization conditions across decades. The collection comprises scans of varying resolutions and quality, including shadowed, curved, or oriented pages, as well as both single and double-page layouts. The scans originate from a wide range of historical newspapers and time periods (e.g., Ժամանակ, Constantinople, 1915; Հայրենիք, Boston, 1938; Յառաջ, Paris, 1977), thus covering multiple printing and digitization standards. To complement this corpus, we added 150 smartphone photographs of contemporary newspapers, intentionally captured under uncontrolled conditions—non-orthogonal angles, variable lighting, and inconsistent viewing distances—to simulate real-world acquisition scenarios. This combination ensured a balanced dataset between high-quality and degraded samples.

For the classification tasks (image quality and orientation), only the 450 real images were used, evenly distributed across the target classes to ensure balanced training. In contrast, for the page detection and segmentation task, additional 1,500 synthetic samples were generated using the Albumentations library to simulate realistic degradation scenarios.

The synthetic dataset was produced by applying a probabilistic composition of photometric and geometric transformations, including brightness and contrast shifts (RandomBrightness Contrast), color perturbations (RGBShift, HueSaturationValue), and environmental effects such as shadows, fog, or rain (RandomShadow, RandomFog, RandomRain). Structural distortions such as ElasticTransform, OpticalDistortion, and Perspective were also introduced to emulate paper

[7] The generative LLM-based solution was chosen for its integration flexibility and its ability to handle document understanding tasks efficiently, thanks to its very large context window, particularly suitable for press materials. In the long term, the development of Armenian-specific embeddings could enable the use of smaller and more specialized models, following initiatives such as Metric-AI's (see: https://huggingface.co/Metric-AI/armenian-text-embeddings-1).

curvature, scanning skew, and camera-induced deformation. This augmentation pipeline enabled the generation of visually diverse and plausible degraded images, thereby improving the robustness of page detection models. The resulting dataset provides an equilibrium between real-world diversity and synthetic variability, ensuring exposure to a broad range of document conditions. All annotations have been made using the LabelStudio interface.

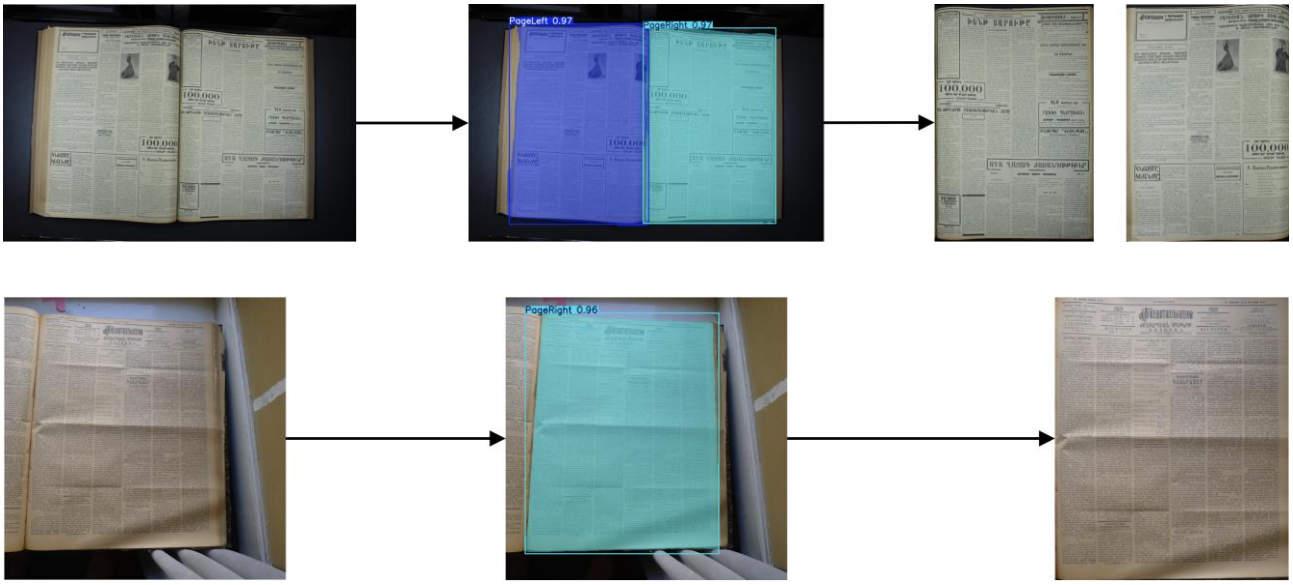
Task / Model	Images	Annotation Type	# Annotations
Image Quality Classification	450	Binary labels (good quality / bad quality)	450 (120 good quality, 300 bad quality, only random 120 samples used for training)
Orientation Classification	448	4 classes (0°, 90°, 180°, 270°)	450 (112 for each class)
Page Detection	1,950	Polygons (PageLeft, PageRight)	3211 (1621 for PageLeft and 1590 for PageRight)

Results and Discussion

The proposed models were evaluated on an independent validation set corresponding to 10% of the dataset, randomly selected and excluded from training. Image quality classification achieved a high level of reliability, with 97% of low-quality images correctly identified as bad quality and 98% of high-quality scans correctly classified as good quality. Minor confusion occurred mainly in borderline cases, such as slightly blurred but still legible pages. These misclassifications had no significant impact on downstream processing, as the subsequent steps remain robust to moderate degradation.

For orientation classification, the model achieved near-perfect performance, with less than 1% confusion between opposite orientations (0° vs 180° and 90° vs 270°). It generalizes well across the four rotation classes, confirming the adequacy of the dataset size and class balance. The page detection and segmentation model reached an average precision of 0.98, recall of 0.97, and a mean Average Precision (mAP) of 97.1%. Most segmentation errors occurred in challenging cases involving severe curvature or strong shadowing, while standard scans and smartphone photographs were accurately processed. The model successfully distinguishes PageLeft and PageRight regions, performing consistently across both single- and double-page layouts.

To assess domain robustness, two models were trained: one using only NLA-scanned data, and another including the 150 smartphone photographs. This setup enabled an in-domain (scanned documents) versus out-of-domain (non-scanned photographs) comparison. The results showed no statistically significant difference, with performance variations remaining within ±1% across all metrics. This indicates that the models generalize effectively to heterogeneous acquisition conditions, including less standardized photographic inputs.



The results obtained in this study confirm that a modular AI-driven pipeline can efficiently automate key preprocessing stages in historical document digitization. Despite its relatively small training corpus, the system demonstrates stable performance across heterogeneous inputs, validating the feasibility of large-scale deployment within cultural heritage workflows.

However, several limitations and biases should be acknowledged. The most important concerns the empirical definition of image quality: the binary distinction between good and bad quality remains subjective, as it relies on human perception of legibility rather than quantifiable measures. This introduces potential inconsistencies in labeling, particularly for borderline cases. Future work should explore the integration of objective indicators, such as contrast, sharpness, or OCR confidence metrics, to make quality assessment more reproducible. A second limitation relates to the restricted dataset size. With fewer than five hundred real samples, the models were trained on a relatively small corpus for deep learning applications. While data augmentation and synthetic transformations helped enrich the diversity of samples, they cannot fully capture the texture complexity, lighting variations, or degradation patterns of genuine historical materials. Increasing both the volume and heterogeneity of the dataset would directly enhance the model's generalization and stability.

The full implementation is open-sourced and available on GitHub. The repository is organized as follows:

- main.py orchestrates the full inference workflow and user interface logic;
- config.py defines model paths, thresholds, and runtime parameters;
- models/ contains trained YOLOv11n weights (quality.pt, classify_rotation.pt, PageDetectNewspaper_best.pt);
- modules/ implements the functional components: classification.py (quality/orientation tasks), detection.py (page segmentation), ocr.py and llm.py (optional text processing), utils.py (image I/O and transformations), and load_models.py (dynamic model loading).

Code availability

Models, inference code and web app are available on Github: https://github.com/CVidalG/workshop-TUMO2025 All models were trained and tested in Python 3.10 using PyTorch 2.3 on a single NVIDIA T4 GPU, with an average inference time below one second per page. The models can also be executed efficiently on CPU.

Acknowledgments

This paper presents the outcomes of the workshop *Preserving the Past with AI*, held at TUMO Yerevan and led by Calfa[8] (Chahan Vidal-Gorène and Baptiste Queuche), in collaboration with the National Library of Armenia. We extend our sincere gratitude to the National Library of Armenia for providing high-resolution document samples used for model training and evaluation, as well as for the valuable discussions that helped shape the project's objectives and use cases. Special thanks are also due to the TUMO teams for their support in organizing and facilitating the workshop.

ԱՄՓՈՓՈՒՄ

Սույն հետազոտությունը ցույց է տալիս պատմական հայկական թերթերի ավտոմատ նախամշակման համար նախատեսված թեթև և մոդուլային արհեստական բանականության (ԱԲ) հոսքագծի կիրառելիությունն ու արդյունավետությունը։ Դասակարգման, հատվածավորման և երկրաչափական ուղղման մեթոդների համադրումը միասնական համակարգում զգալիորեն բարձրացնում է թվայնացված, ձևով ու որակով տարբեր արխիվների օգտագործելիությունը՝ միաժամանակ պահպանելով դրանց վավերականությունն ու բնօրինակ տեսքը։

Աշխատանքը իրականացվել է ԹՈՒՄՈ Երևանում անցկացված երկշաբաթյա ուսանողական աշխատարանի շրջանակում, ուստի այն պետք է դիտարկել որպես կրթական բնույթի նախագիծ, այլ ոչ թե որպես ամբողջությամբ պատրաստ արտադրական լուծում։ Ստացված հոսքագիծը և մոդելները լիարժեք աշխատունակ են, սակայն ինստիտուցիոնալ թվայնացման գործընթացներում կիրառելուց առաջ անհրաժեշտ են հետագա զարգացումներ և լայնածավալ փորձարկումներ։

Նախագիծը հաստատում է բաց ԱԲ գործիքների մեծ ներուժը մշակութային ժառանգության պահպանման ոլորտում և ցույց է տալիս, որ նույնիսկ սահմանափակ ռեսուրսներով հնարավոր է հասնել բարձր արդյունավետության։ Հետագա աշխատանքները կուղղվեն տվյալների շտեմարանի ընդլայնմանը, որակի օբյեկտիվ չափորոշիչների կատարելագործմանը, ինչպես նաև հոսքագծի ինտեգրմանը Հայաստանի ազգային գրադարանի թվային ենթակառուցվածքներում՝ արտադրական մասշտաբով կիրառման նպատակով։

[8] https://calfa.fr

REFERENCES

1. Bermès, E. Le numérique en bibliothèque: naissance d'un patrimoine. L'exemple de la Bibliothèque nationale de France (1997–2019), Paris, École nationale des chartes, 2020.
2. Calfa. hye-calfa-n: Open-Source OCR Model for Armenian, 2025, [Electronic Resource] URL: https://github.com/calfa-co/hye-tesseract; 02.11.2025.
3. Gasparini, A.; Kautonen, H. Understanding Artificial Intelligence in Research Libraries – Extensive Literature Review // LIBER Quarterly: Journal of European Research Libraries, 2022, Vol. 32 (1), pp. 1–36, DOI: 10.53377/lq.10934; 02.11.2025.
4. Jolivet, V.; Terriel, L.; Canteaut, O. From Manuscript to Data: An Integrated Pipeline for Handwriting Recognition, Editing, and Indexing. // Journal of Data Mining and Digital Humanities, 2025, [Electronic Resource] URL: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://hal.science/hal-05117289/document, 02.11.2025.
5. Ranjan, A.; Ravinder, M. ROBDD-TrOCRBERTa: A Novel Robust-Optimized Blurred Document Text Deblurring and Completion with DCGAN-TrOCR and DistilRoBERTa. // International Journal of Information Technology, 2024, Vol. 16, № 7, pp. 4611–4619.
6. Rehm, G.; Bourgonje, P.; Hegele, S.; Kintzel, F.; Schneider, J. M.; Ostendorff, M.; Zaczynska, K.; Berger, A.; Grill, S.; Räuchle, S.; et al. QURATOR: Innovative Technologies for Content and Data Curation // arXiv preprint. arXiv:2004.12195, 2020, pp. 1–10. [Electronic Resource] URL: https://arxiv.org/abs/2004.12195, 02.11.2025.
7. Karmanov, I.; Deshmukh, A. S.; Vögtle, L.; Fischer, P.; Chumachenko, K.; Roman, Timo, Seppänen, Jarno, Parmar, Jupinder, Jennings, Joseph, Tao, Andrew, et al. Éclair – Extracting Content and Layout with Integrated Reading Order for Documents, 2025, // arXiv preprint arXiv:2502.04223, [Electronic Resource] URL: https://arxiv.org/abs/2502.04223, 02.11.2025.
8. Vidal-Gorène, Ch. OCR / HTR Technologies and Armenian Heritage Preservation. // Bulletin of Armenian Libraries, 2023, Vol. 6, № 1, pp. 61–65. [Electronic Resource] URL: https://journal.nla.am/index.php/banber/article/view/20, 02.11.2025.
9. Vidal-Gorène, Ch.; Camps, J.-B. Image-to-Image Translation Approach for Page Layout Analysis and Artificial Generation of Historical Manuscripts // Proceedings of the International Conference on Document Analysis and Recognition, 2024, pp. 140–158.
10. Vidal-Gorène, Ch.; Camps, J.-B. Traiter des manuscrits endommagés par le feu grâce à l'augmentation de données par intelligence artificielle: le cas du ms. L.II.14 de Turin (T) // Studi Francesi, Torino, Rosenberg & Sellier, 2025, Vol. 206, pp. 377–383, [Electronic Resource] URL: https://enc.hal.science/hal-05184196, 02.11.2025.
11. Vidal-Gorène, Ch.; Decours-Perez, A.; Kasparian, A.; Tanelian, A.; Ohanian, A. Armenian HTR: State of the Art, Transcription Guidelines and Good Practices, 2025, [Electronic Resource] URL: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://enc.hal.science/hal-05021697/document, 02.12.2025.
12. Yang, Zh.; Peng, D.; Shi, Y.; Zhang, Y.; Liu, Ch.; Jin, L. Predicting the Original Appearance of Damaged Historical Documents. // Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, № 9, pp. 9382–9390.
13. Zhang, J.; Peng, D.; Liu, Ch.; Zhang, P.; Jin, L. DocRes: A Generalist Model Toward Unifying Document Image Restoration Tasks // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1–10, [Electronic Resource] URL:

https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_DocRes_A_Generalist_Model_Toward_Unifying_Document_Image_Restoration_Tasks_CVPR_2024_paper.html, 02.11.2025.