National Archives of Finland handwriting recognition for court records

IMPLEMENTED 2022–2025 · GOVERNMENT AND PUBLIC SERVICES · FINLAND · BAND C · PARTLY PROVED
VERSION 1.0 · PUBLISHED 25 NOVEMBER 2026 · UPDATED 25 NOVEMBER 2026 · FACTS CHECKED 25 NOVEMBER 2026

A national archive trained its own AI handwriting model for Finnish and Swedish and used it to make millions of scanned court record pages searchable by full text.

ABOUT THIS SYNOPSIS

This is MarLuc's summary and grading of publicly available sources, for information only. It is not investment, legal, financial or professional advice, and should not be relied on for any decision. Readers must validate it against the original sources and, where appropriate, challenge its conclusions.

Facts and figures are as reported by the sources listed below, which may be incomplete, self-reported or out of date. Check the original source before relying on any figure.

The impact band, proof marker and SEAM view are MarLuc's opinion. The organisation is not a MarLuc client and has not reviewed or endorsed this synopsis. Names and trademarks belong to their owners.

NOTICE WORDING LAST UPDATED 16 NOVEMBER 2026

AT A GLANCE

AI used: Transformer-based handwriting recognition model trained by the archive.

Implemented: 2022–2025. Grade: band C, Process change, Partly proved.

Headline result, as reported: Archive's own handwriting model made over three million court record images full-text searchable in 2022; its later multicentury model scored a character error rate of 2.9 on the archive's tests

1 · The case

The facts, as reported by the 4 sources listed at the end. Where a source is silent, the field says "Not reported".

Organisation

National Archives of Finland (Kansallisarkisto). Finland's national archive. Its holdings include court records, church records and war diaries, per its sources. Staff numbers: Not reported. The archive says it has 240 shelf kilometres of material to digitise. Canvas block: 4. Operations and process. Also 3. Customers and proposition.

Business challenge

The challenge. Handwritten historical documents are hard to search. The archive says its strategic goal is to turn historically significant material into data that is more searchable and usable for digital research.

Who implemented it

In-house (In-house): Developed and trained the handwriting model.
READ-COOP (Software vendor): Cooperative that maintains Transkribus; co-delivered the service and processing.
Microsoft (Model or cloud provider): Source of the TrOCR architecture the model is fine-tuned from.

Technology (SEAM pillar E)

SEAM types: T5 Custom / fine-tuned models. Deployment: Not reported. Vendors and products: Microsoft, READ-COOP, TrOCR, YOLOv8, Transkribus.

Where it was built in: Used to process scanned court records and other records into text for the Finnish Court Records Search Service; the model is applied to large sets after fine-tuning on a small portion.

What changed

Before: researchers searched the court records by key word spotting. After: the service offers full-text and phrase search across over three million images, mostly 1809 to 1870 and in Swedish. The archive also uses the model on other large sets such as war diaries. Staff time and cost effects: Not reported. [1] [2] [3] [4]

What it cost

Not reported

What it saved

Not reported. By the archive, one multicentury model works as well as century-specific ones, so it does not plan to maintain separate models. Researchers used the model to extract over six million entries from about 200,000 images of church migration records. [1] [2] [3] [4]

Governance and control

The archive measured accuracy with character error rate on a held-out test set and trains on human transcriptions. Its project partners include national archives in Estonia and Latvia. Accountable owner, human review of recognised text and standards: Not reported.

S2 · Accountable owner · Not reported
Not reported.

A1 · Safe use · Not reported
Not reported.

A2 · Human in the loop · Not reported
Human transcribers create the training data; checking of recognised text before users rely on it is not reported. [2] [3]

A3 · Monitoring and assurance · Claimed
The archive measures character error rate on held-out test data by century and language. [2]

A4 · Vendor AI · Not reported
The model is fine-tuned from Microsoft's TrOCR and the 2022 processing was done with READ-COOP; how the archive assesses supplier models is not reported. [1] [2]

A5 · Agent controls · Not reported
Not reported; the system recognises text and is not described as an agent.

A6 · Standards and regulation · Not reported
Not reported.

What worked, what did not, what is unproven

WORKED

– The multicentury model reached a character error rate of 2.9 per 100 characters on the archive's own test set.

– One model handled Swedish and Finnish without separate language classification, which the archive called unexpected.

– Independent researchers used the archive's model as part of a pipeline that produced a dataset of over six million entries.

DID NOT

– The highest character error rate, 5.5, was on 17th-century pages, which the archive links to scanning quality of very old paper.

– On church migration table cells, independent researchers found the model read 49.9% of test lines exactly; lines a human annotator could not read were excluded.

– The archive says more training data would improve accuracy.

UNPROVEN

– Cost, staff time saved and how often researchers use the full-text search: Not reported.

– Accuracy on material beyond the archive's own test set, such as the 20th-century Russian records now being trained for.

– Whether recognised text is checked before users rely on it.

Timeline

Mar 2022 Archive adds over 1.5 million new court record images with automated text recognition, taking the service to over three million images [1]

Mar 2022 Search changes from key word spotting to full-text search [1]

2024 Archive reports a multicentury Finnish and Swedish model trained on 1600s to 1900s handwriting [2]

2025 University researchers reuse the archive's model to extract six million entries from church migration records [4]

Nov 2025 ArchXAI project with the Estonian and Latvian archives trains a Russian handwriting model; over 1,700 pages transcribed so far [3]

2 · Grade and SEAM assessment

MarLuc's opinion, made on the evidence above. How cases are graded: marluc.co.uk/case-studies/how-we-rate

Level of change and impact

Claimed: Not applicable. Grade: C, Process change. Proof: Partly proved. Level of change: Process change.

Process change (Level 2, band C). Records are now text-recognised and searched in full text instead of by key word spotting.

Graded C. The archive redesigned how handwritten records reach users: scan, recognise text with its own model, then offer full-text search. Results are the archive's own, with one independent reuse. Cost, time saved and use by researchers are not reported.

The SEAM view and lessons learnt

The archive built a model for its own material and put it into a public search service, so the result is easy to attribute to the AI. The accuracy figures are the archive's own; one outside reuse on a different kind of record read about half of test lines exactly. Before copying the approach, a leadership team should ask how much human transcription it needs to train a model on its own handwriting, and how it will tell users that recognised text can contain errors.

1. Train on your own records and test on pages the model has not seen.
2. Report accuracy by period and type of record, not as one number.
3. Tell users that recognised text can contain errors.

Evidence

Score 2 of 4. Partly independent: the service and accuracy figures are the archive's own; a University of Helsinki and University of Turku preprint reused the model and reported its own measurements.

Evaluation method. Self-reported archive tests, with one independent reuse: character error rates on a held-out 10% test set from the archive's poster, and an independent preprint that applied the model to church records.

Sources disagree. The archive's poster reports a character error rate of 2.9 errors per 100 characters on its court-record test set. The independent preprint reports 0.19 and exact matches on 49.9% of lines for the same model on church migration table cells; it does not state the unit, and the tests differ, so the figures are not directly comparable.

Status. Ongoing: the archive's ArchXAI project, running 2025 to 2028, is extending handwriting recognition to Russian; the court records service said in 2022 that more judgment books would be added.

3 · Original sources

Read these in full. To challenge a fact or a grade, use the contact form at marluc.co.uk or the link on https://marluc.co.uk/case-studies/national-archives-finland-court-records-htr

[1] National Archives of Finland: The Finnish Court Records Search Service, you can now use full-text search with over three million hand-written pages (press release, 21 March 2022)
https://kansallisarkisto.fi/en/-/the-finnish-court-records-search-service-update-you-can-now-use-full-text-search-with-over-three-million-hand-written-pages

[2] National Archives of Finland: Handwritten Text Recognition model for historical documents (poster)
https://kansallisarkisto.fi/documents/141232930/153230445/poster_joensuu_202411_final+1.pdf/4d75cc58-d36d-7fac-bbb5-91e5e547ea43/poster_joensuu_202411_final+1.pdf?t=1736927900122

[3] Central Baltic Programme: ArchXAI project develops AI-powered handwritten text recognition for 19th and 20th century Russian (20 November 2025)
https://centralbaltic.eu/project-news/archxai-project-develops-ai-powered-handwritten-text-recognition-for-19th-and-20th-century-russian/

[4] arXiv (University of Helsinki and University of Turku): Creating a Historical Migration Dataset from Finnish Church Records, 1800-1920
https://arxiv.org/pdf/2506.07960

National Archives of Finland (Kansallisarkisto) website: https://kansallisarkisto.fi/en

Version 1.0 · Published 25 November 2026 · Updated 25 November 2026 · Facts checked against sources 25 November 2026 · Downloaded 26 November 2026. This PDF may be out of date: the latest version, and a way to report an error or a better source, is at https://marluc.co.uk/case-studies/national-archives-finland-court-records-htr

VERSION HISTORY

v1.0 25 November 2026 First published