Nationaal Archief handwriting recognition of VOC and notarial records

IMPLEMENTED 2019–2025 · GOVERNMENT AND PUBLIC SERVICES · NETHERLANDS · BAND C · CLAIMED ONLY

VERSION 1.0 · PUBLISHED 25 NOVEMBER 2026 · UPDATED 25 NOVEMBER 2026 · FACTS CHECKED 25 NOVEMBER 2026

A national archive trained handwriting recognition models on hand-typed transcriptions to make about two million scans of VOC and notarial records searchable, then extended the approach to a 30 million scan project.

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Facts and figures are as reported by the sources listed below, which may be incomplete, self-reported or out of date.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: Handwriting recognition models trained on hand-typed transcriptions, with AI to mark names and places.

Implemented: 2019–2025. Grade: band C, Process change, Claimed only.

Headline result, as reported: Archive says its model reads over 90% of characters correctly; about 2 million scans transcribed and over a million searchable on a temporary project site due to close in April 2025

1 · The case

The facts, as reported by the 3 sources listed at the end. Where a source is silent, the field says "Not reported".

Organisation

Nationaal Archief (National Archives of the Netherlands), with Noord-Hollands Archief. National archive of the Netherlands, a public body. Staff numbers: Not reported. The project partnered with Noord-Hollands Archief and other regional historical centres. Canvas block: 4. Operations and process. Also 3. Customers and proposition.

Business challenge

The challenge. The archive says many handwritten archives are held but few people can still read them, and researchers want to search the scans more easily. Its project covers 17th and 18th century records of the Dutch East India Company (VOC) and 19th century notarial archives.

How it was addressed. From early 2019 the archive had documents typed by hand, then trained handwriting recognition models in Transkribus, including one named IJsberg that is freely available. The trained models transcribed about two million scans automatically. Later, a further million West India Company scans were transcribed without extra training. In September 2019 the archive put out a tender to three software developers to build prototypes of a search tool. Names and places in the texts are marked automatically, and Vitec (formerly Picturae), Aincient and Sioux Technologies built the transcription management system and enriched transcriptions with AI. [1] [2] [3]

Who implemented it

In-house (In-house): Ran the project and produced the training transcriptions.
READ-COOP (Transkribus) (Software vendor): Platform used to train and run the handwriting recognition models.
Vitec (formerly Picturae), Aincient and Sioux Technologies (Systems integrator): Built the transcription management system and enriched transcriptions with AI.
KNAW Humanities Cluster (University or research lab): Developed the Loghi transcription software with the archive's help.

Technology (SEAM pillar E)

SEAM types: T5 Custom / fine-tuned models. Deployment: Not reported. Vendors and products: READ-COOP, Transkribus, IJsberg model, Loghi.

Where it was built in: Scans of VOC and notarial records were transcribed in Transkribus; transcriptions were then loaded into a search website and enriched with AI. The archive has integrated Loghi into its technical infrastructure.

What changed

Before: researchers browsed scans of handwritten documents. After: over a million transcriptions could be searched on the project website, part of them enriched with corrections, modernised spelling and entities. Staff time and cost effects: Not reported. [1] [2] [3]

What it cost

Not reported

What it saved

Not reported. The archive says full-text search of the transcriptions gives a very large number of results and that finding the right information is not easy. [1] [2] [3]

Governance and control

The archive trained models on documents typed by hand and published the models and training data under open licences on Transkribus and Zenodo. Accountable owner, review of output and a standards framework: Not reported.

S2 · Accountable owner · Claimed
The archive's digitisation project leader is named as the contact for the programme. [2]

A1 · Safe use · Not reported
Not reported.

A2 · Human in the loop · Not reported
Staff typed transcriptions by hand to train the models; human checking of machine output before publication is not reported. Part of the searchable transcriptions were enriched with corrections; who made them is not stated. [1] [2]

A3 · Monitoring and assurance · Claimed
The archive reports accuracy of more than 90% of characters against its hand-typed set. [1] [3]

A4 · Vendor AI · Claimed
The archive put out a tender to three software developers to build competing prototypes before choosing one; Vitec, Aincient and Sioux Technologies later built the transcription management system and enriched transcriptions with AI. [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

– By the archive, more than 90% of characters are recognised correctly.

– A further million West India Company scans were transcribed automatically without extra ground truth or training.

– The archive released the transcriptions and training data as open data and the IJsberg model for others to use.

DID NOT

– The archive says full-text search brings transcription errors, spelling variants, old language and layout recognition errors that make finding the right information hard.

– The Zenodo record says the improved PyLaia model is not publicly available.

UNPROVEN

– Independent accuracy of the models, especially on records unlike those used for training.

– Cost, staff time saved and how many researchers use the transcriptions.

– How the 30 million scan Oorlog voor de Rechter project will perform at that scale.

Timeline

2019 Archive starts large-scale handwriting recognition with the project De ijsberg zichtbaar maken [1]

Feb 2020 Archive reports a model recognising over 90% of characters correctly, ready to transcribe two million scans [1]

May 2021 Archive launches the project website zoekintranscripties.nl [2]

May 2024 Version 8.1 of the open dataset published on Zenodo; the record says 1 million extra West India Company scans were transcribed without extra training [3]

Apr 2025 Project website due to close on 1 April 2025, by the archive's page [2]

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: Claimed only. Level of change: Process change.

Process change (Level 2, band C). Scans became machine-readable text through a repeatable pipeline; accuracy and use are the archive's own account.

Graded C. The archive changed how scans are made searchable, from page images to machine-read text. The accuracy figure is the archive's own and no independent evaluation was found, so the marker is claimed only. Cost, use and effect on researchers are not reported.

The SEAM view and lessons learnt

The accuracy and scale figures are the archive's own and the AI's part is easy to separate, because the output is a set of text files from a known process. The sources do not say what the work cost or how researchers use it. Before copying the approach, a leadership team should ask how it will handle errors in machine-read text and who will keep a search service running once project funding ends.

1. Publish training data and models so others can check accuracy.
2. Plan how users will find the right result among a very large number of matches.
3. Decide in advance who runs a project website when the project ends.

Evidence

Score 1 of 4. Self-reported: the accuracy and volume figures come from the archive's own news item, page and open dataset.

Evaluation method. Self-reported archive figures: the accuracy figure is from the archive's own news item, with open data that others could use to check it.

Sources disagree. The page counts differ between the archive's sources: the dataset record describes a plan to transcribe 2 million pages and a dataset title mentioning 3,000,000 HTR pages, while the project page says more than one million transcriptions are searchable and one and a half million are on Zenodo. The 2020 news item says 7,706 documents were typed by hand to develop and test models; the dataset record says about 6,000 scans were transcribed to train the model. The sources count different stages and datasets.

Status. Ongoing: the archive's page says it is transcribing 30 million scans in Oorlog voor de Rechter and has integrated the Loghi transcription software into its technical infrastructure; the project website was due to close on 1 April 2025.

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/nationaal-archief-ijsberg-handwriting-recognition

[1] Nationaal Archief: Succesvolle resultaten automatische handschriftherkenning (news, 4 February 2020)
https://www.nationaalarchief.nl/archiveren/nieuws/succesvolle-resultaten-automatische-handschriftherkenning

[2] Nationaal Archief: Handschriftherkenning (DataLab page)
https://www.nationaalarchief.nl/onderzoeken/datalab-nationaal-archief/handschriftherkenning

[3] Zenodo (Liesbeth Keijser, National Archive Netherlands): 6000 ground truth of VOC and notarial deeds, 3.000.000 HTR of VOC, WIC and notarial deeds
https://zenodo.org/records/11209325

Nationaal Archief (National Archives of the Netherlands), with Noord-Hollands Archief website:
https://www.nationaalarchief.nl

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/nationaal-archief-ijsberg-handwriting-recognition

VERSION HISTORY

v1.0 25 November 2026 First published