

Automated Extraction of Analog Macroseismic Questionnaires using a Multimodal LLM

Anna Bardroff¹, Andrea Brüstle², Thomas Wiesehöfer², Stefan Stange²

¹ Johann-Wolfgang-Goethe-Universität Frankfurt am Main, Institut für Geowissenschaften, Frankfurt am Main

² Landeserdbebendienst Baden-Württemberg, Landesamt für Geologie, Rohstoffe und Bergbau, Regierungspräsidium Freiburg, Freiburg

1. Introduction

The State Seismological Survey of Baden-Württemberg (Landeserdbebendienst, LED) holds an extensive collection of analogue paper questionnaires from historical macroseismic surveys of past earthquakes. Those questionnaires are predominantly filled in in handwriting and contain macroseismic observations at the municipal level (Gemeindefragebögen, cf. Fig. 1).

Fragebogen zur Beobachtung von Erdbeben

Am 5.12.2004 ereignete sich ein Erdbeben bei Waldkirch im Landkreis Emmendingen. Mit dieser Anfrage erhalten wir die Auswertungen der durch das Erdbeben verursachten Bodenveränderungen. Die Auswertungen sollen sich auf einen Gürtelschnitt ihrer Gemeinden ausdehnen. Wenn das nicht möglich ist, geben Sie bitte Ihre persönliche Einschätzung wieder.

Bitte zurücksenden, zusammen mit dem Formular an: led@lgrb-bw.de oder per Post an: Landesamt für Geologie, Rohstoffe und Bergbau (LGRB), Baden-Württemberg, Erdbaubereich, Albertstraße 5, 79104 Freiburg i. Br.

1. Beobachtungsort: Gemeinde: Donauversingen, Kreis: Schwarzwald-Baar-Kreis

2. Zeitpunkt der Erschütterung: Tag: 5.12.2004, Uhrzeit: 2.52 Uhr

3. Wie viele Einwohner Ihrer Gemeinde haben das Erdbeben gespürt? ☐ Einzelne, ☐ Wenige (ca. 5 %), ☐ Viele (ca. 50 %), ☒ Die Meisten (ca. 75 %), ☐ Alle (ca. 100 %), ☐ Unbekannt

4. Welche Art von Erschütterungen traten auf? (mehrere Antworten möglich)

5. Welche geologische Wirkung wurde beim Erdbeben beobachtet?

Fig. 1 Example of a questionnaire for the earthquake in Waldkirch on 5 December 2004

3. Workflow (cf. Fig. 2)

- Scanning
All questionnaires for a single event are collected and scanned on a standard multifunction copier with an automatic document feeder.
- Automated Processing
 - Splitting and Alignment Correction
The resulting PDF file is split into individual image files for each questionnaire. Optionally, the images can be aligned and rectified using the Scale-Invariant Feature Transform (SIFT) algorithm. This is not needed for the digitisation, but can help with the manual quality control.
 - Digitisation and Conversion to a Structured Output Format
For each image file, a fine-tuned multi-modal Large Language Model (LLM) is prompted to transfer the included information into a specified output format. The individual records are combined into a single output file.
- Manual Quality Control
The output data can be efficiently verified against the original scans and corrected if necessary using a custom-developed quality control tool.

4. Large Language Model (LLM)

We use an open-source Large Language Model (*Qwen3-VL-32B-Instruct* by Alibaba). The model is operated on premise and without any data transfer to “the cloud”, ensuring data security in cases where, for example, personal data is included in the questionnaires.

After digitising and manually correcting the first 300 questionnaires using our workflow, we fine-tuned the LLM with this verified data set. This retraining is crucial for the accuracy of the digitisation.

5. Quality Control Tool

Since 100% accuracy cannot (yet) be achieved with automatic digitisation, we have developed a web-based tool for manual quality control that displays the extracted data in the original format of the questionnaire, where it can also directly be corrected if necessary.

A hotkey can be used to switch between the scan and the digitised data, so that any differences are immediately apparent. Additionally, when moving the mouse over the digitised form, the corresponding section of the scan appears in a kind of “magnifying glass”, so that the original data remain visible at all times during editing (cf. Fig. 3).

Qualitätsprüfung

1. Beobachtungsort: Gemeinde: Donauversingen, Kreis: Schwarzwald-Baar-Kreis

2. Zeitpunkt der Erschütterung: Tag: 5.12.2004, Uhrzeit: 2.52 Uhr

3. Wie viele Einwohner Ihrer Gemeinde haben das Erdbeben gespürt? ☐ Einzelne, ☐ Wenige (ca. 5 %), ☐ Viele (ca. 50 %), ☒ Die Meisten (ca. 75 %), ☐ Alle (ca. 100 %), ☐ Unbekannt

4. Welche Art von Erschütterungen traten auf? (mehrere Antworten möglich)

5. Welche geologische Wirkung wurde beim Erdbeben beobachtet?

Fig. 3 Screenshots of the quality control tool, showing the original scan (left) and the editable view of the extracted data (right)

6. Conclusion / Outlook

After fine-tuning the LLM, the pre-verification accuracy—which varies with handwriting quality—is above 95 %, and any necessary corrections can be made swiftly with the help of the quality control tool. In this way, the workflow reduces the time required for digitisation from weeks to hours. Transfer of the system to other scopes of application, for instance handwritten bulletins, should be feasible.

2. AI Pipeline

We therefore developed an AI-based pipeline (cf. Fig. 2) that automates the bulk of the work and converts the macroseismic observations from the analogue forms into a structured JSON format for subsequent evaluation.

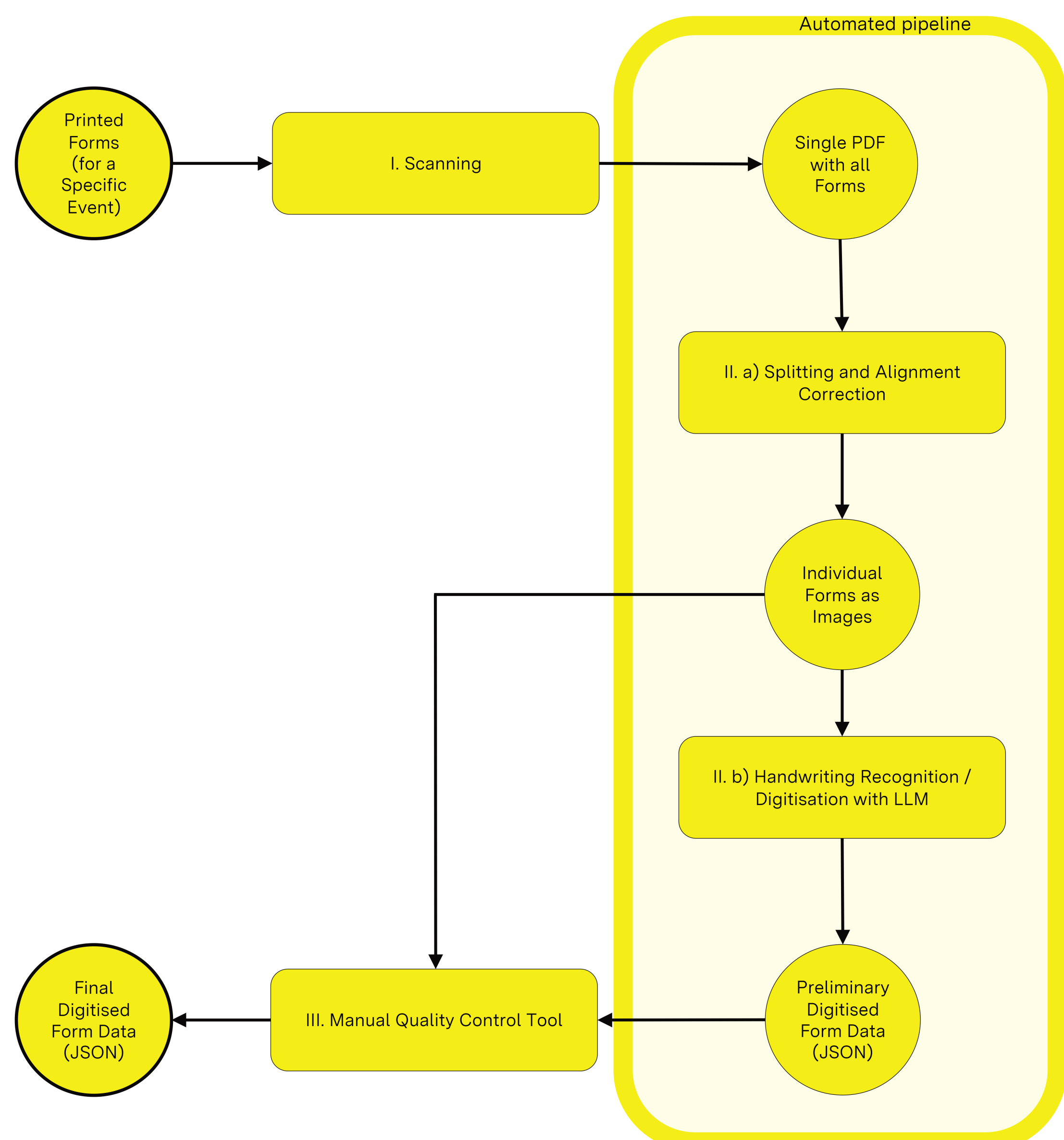


Fig. 2 Workflow for digitisation, with the automated pipeline shown on the right-hand side

