Digitisation of scanned reviews from zbMATH Open

Bela Gipp, Ivan Pluzhnikov, Moritz Schubotz and Olaf Teschke

This note reports on the digitisation of scanned reviews from zbMATH volumes 1–529. We describe the procedure, techniques, and achievements, and release the dataset of images, corresponding text, and structured information for the community.

1 Introduction

In an earlier column [1], we discussed the challenge and possible approaches to convert scanned mathematical content into LaTeX, with a focus on the more than 250,000 pages of reviews from the volumes 1–529 of *Zentralblatt für Mathematik und ihre Grenzgebiete*. This corpus plays a specific role since it bridges the gap between the reviews of the *Jahrbuch über die Fortschritte der Mathematik* from 1868 to 1942, whose 181,613 reviews were mostly (OCR-supported) manually retyped in a project 1998–2002 [10], and the digitally-born zbMATH reviews since 1984. Though the scans were human-readable, they do not serve the requirements of a modern information ecosystem: they are not easily searchable, analysable, or processable; in particular, they cannot be automatically integrated to or interlinked with other services.

Especially, the transition of zbMATH into an open service [4, 5] makes it desirable to have these resources available as open resources. Even just within the service, features like the recently introduced entity linking [3] would benefit much from retrodigitisation.

2 LaTeX transition and review assignments

The digitisation of the scanned zbMATH volumes involved two essential components: Pure LaTeX transformation and review assignment through proper segmentation. During the last years, AI tools have improved significantly. For the translation of scans into LaTeX, Mathpix[1] has evolved into the state-of-the-art tool. As already indicated in [1], it performs quite well even for relatively low-quality scans, and has still improved much in the last years. Tests confirmed its capability; in fact, Mathpix was able to convert 99.1% of the pages into further processable LaTeX, outperforming other tools significantly.

However, the task did not stop here: Distributed among the 254,444 pages are 823,280 entries – mostly with single reviews, but also both title-only indexed items and joint reviews for multiple documents. Several short reviews may appear on one page, as well as long multi-page-reviews, especially for books. This creates a complex $n : m$-assignment task. To make things worse, the typesetting style (including placement of metadata and reviewer signatures) changes frequently with publication years. Experiments with traditional procedural algorithms for the first volumes indicated that it may require more than five person years of coding to assign the raw OCR output reliably – which would not just create costs of almost the magnitude of manual typing, but also postpone a release significantly.

3 LLM-based document splitting

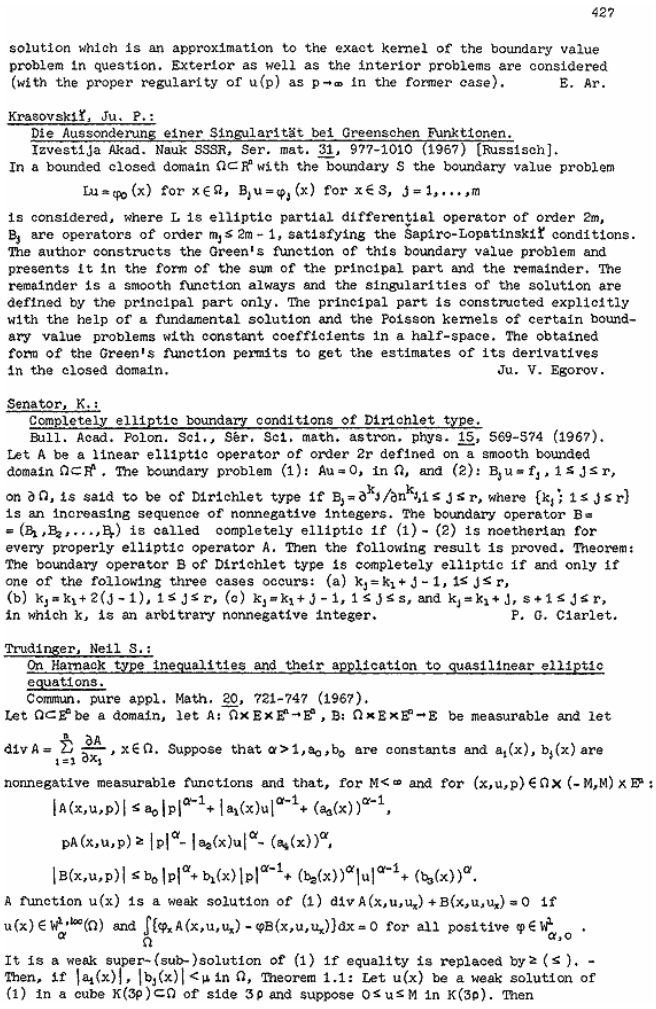
Fortunately, modern artificial intelligence methods offer more efficient alternatives. Large language models (LLMs) have evolved as powerful and flexible tools for such tasks.

The core pipeline processed the OCR rows sequentially, constructing LLM prompts that including the current and next document titles and sources, along with OCR-derived text. These prompts were passed to a generative LLM to extract the correct document boundaries. Special handling was applied for final documents and exclusions.

Due to hardware constraints, three instruction-tuned open-source LLMs were selected based on performance benchmarks: CalmExperiment-7B-slerp, Maxine-7B, and Mistral-Passthrough-10B. These models were further improved via fine-tuning using the LoRa (low-rank adaptation) optimisation technique [11].

To further improve accuracy, a majority voting ensemble was introduced: if two models agree, their output is selected; otherwise, the best-performing model (CalmExperiment) is used. This approach reduces individual model errors through consensus.

[1] https://mathpix.com



427

solution which is an approximation to the exact kernel of the boundary value problem in question. Exterior as well as the interior problems are considered (with the proper regularity of u(p) as p→∞ in the former case). E. Ar.

Krasovskiĭ, Ju. P.:
Die Aussonderung einer Singularität bei Greenschen Funktionen.
Izvestija Akad. Nauk SSSR, Ser. mat. 31, 977-1010 (1967) [Russisch].
In a bounded closed domain $\Omega \subset R^n$ with the boundary S the boundary value problem

$$Lu = \varphi_0(x) \text{ for } x \in \Omega,\ B_j u = \varphi_j(x) \text{ for } x \in S,\ j = 1, \ldots, m$$

is considered, where L is elliptic partial differential operator of order 2m, B_j are operators of order $m_j \leq 2m-1$, satisfying the Šapiro-Lopatinskiĭ conditions. The author constructs the Green's function of this boundary value problem and presents it in the form of the sum of the principal part and the remainder. The remainder is a smooth function always and the singularities of the solution are defined by the principal part only. The principal part is constructed explicitly with the help of a fundamental solution and the Poisson kernels of certain boundary value problems with constant coefficients in a half-space. The obtained form of the Green's function permits to get the estimates of its derivatives in the closed domain. Ju. V. Egorov.

Senator, K.:
Completely elliptic boundary conditions of Dirichlet type.
Bull. Acad. Polon. Sci., Sér. Sci. math. astron. phys. 15, 569-574 (1967).
Let A be a linear elliptic operator of order 2r defined on a smooth bounded domain $\Omega \subset R^n$. The boundary problem (1): $Au = 0$, in Ω, and (2): $B_j u = f_j$, $1 \leq j \leq r$, on $\partial\Omega$, is said to be of Dirichlet type if $B_j = \partial^{k_j}/\partial n^{k_j}$, $1 \leq j \leq r$, where $\{k_j;\ 1 \leq j \leq r\}$ is an increasing sequence of nonnegative integers. The boundary operator $B = (B_1, B_2, \ldots, B_r)$ is called completely elliptic if (1) - (2) is noetherian for every properly elliptic operator A. Then the following result is proved. Theorem: The boundary operator B of Dirichlet type is completely elliptic if and only if one of the following three cases occurs: (a) $k_j = k_1 + j - 1$, $1 \leq j \leq r$, (b) $k_j = k_1 + 2(j-1)$, $1 \leq j \leq r$, (c) $k_j = k_1 + j - 1$, $1 \leq j \leq s$, and $k_j = k_1 + j$, $s+1 \leq j \leq r$, in which k_1 is an arbitrary nonnegative integer. P. G. Ciarlet.

Trudinger, Neil S.:
On Harnack type inequalities and their application to quasilinear elliptic equations.
Commun. pure appl. Math. 20, 721-747 (1967).
Let $\Omega \subset E^n$ be a domain, let $A: \Omega \times E \times E^n \to E^n$, $B: \Omega \times E \times E^n \to E$ be measurable and let

$$\operatorname{div} A = \sum_{i=1}^{n} \frac{\partial A}{\partial x_i},\ x \in \Omega.$$

Suppose that $\alpha > 1$, a_0, b_0 are constants and $a_i(x)$, $b_i(x)$ are nonnegative measurable functions and that, for $M < \infty$ and for $(x,u,p) \in \Omega \times (-M, M) \times E^n$:

$$|A(x,u,p)| \leq a_0 |p|^{\alpha-1} + |a_1(x)u|^{\alpha-1} + (a_3(x))^{\alpha-1},$$

$$pA(x,u,p) \geq |p|^{\alpha} - |a_2(x)u|^{\alpha} - (a_4(x))^{\alpha},$$

$$|B(x,u,p)| \leq b_0 |p|^{\alpha} + b_1(x)|p|^{\alpha-1} + (b_2(x))^{\alpha}|u|^{\alpha-1} + (b_3(x))^{\alpha}.$$

A function u(x) is a weak solution of (1) $\operatorname{div} A(x,u,u_x) + B(x,u,u_x) = 0$ if $u(x) \in W^{1,\infty}_{\alpha}(\Omega)$ and $\int_{\Omega} \{\varphi_x A(x,u,u_x) - \varphi B(x,u,u_x)\}dx = 0$ for all positive $\varphi \in W^{1}_{\alpha,0}$.

It is a weak super-(sub-)solution of (1) if equality is replaced by $\geq$ ($\leq$). - Then, if $|a_i(x)|$, $|b_i(x)| < \mu$ in Ω, Theorem 1.1: Let u(x) be a weak solution of (1) in a cube $K(3\rho) \subset \Omega$ of side 3ρ and suppose $0 \leq u \leq M$ in $K(3\rho)$. Then

A typical scan from Vol. 153.

Several baselines were evaluated, including regular expressions, a computer vision model (DIT), and ChatGPT-4o (with and without few-shot prompting). Results showed that non-fine-tuned LLMs failed completely, while fine-tuned models performed strongly with an accuracy of 93–97% [8]. The Majority Voting approach achieved the highest accuracy of 97.5%, significantly outperforming all baselines.

4 Error analysis

The analysis of the output identified four main error sources: OCR formatting issues, ambiguity in similar titles, hallucinations in LLM outputs, and confusion between metadata and content. Attempts to predict errors using OCR confidence scores and token probabilities were unsuccessful, indicating these are not reliable indicators.

To map extracted documents back to their original LaTeX sources, exact string matching (with preprocessing to handle OCR spacing issues) was used to determine start and end indices. This method achieved high accuracy (98.4%) on 90.6% of a test dataset. However, this also indicated the need to exclude a significant fraction of about 10% of hallucinated output.

Finally, the approach was scaled to the full OCR output of the more than 250,000 pages, successfully assigning 662,720 (about 89%) of the reviews to the correct metadata. Beyond this, the approach also corrected OCR errors and removed metadata misplaced by the OCR in the review texts.

5 Processing the data into zbMATH Open

At the beginning of 2026, these reviews had been processed into the zbMATH Open production system, and made internally visible to the editors. In the following weeks, further systematic errors have been identified and corrected. This included about 10k formal LaTeX errors, but also semantic errors mostly resulting from poor print quality (e.g., it turned out that 1, l, and i were hardly distinguishable in some volumes). So far, these corrections are far from being completed, the main remaining tasks being the extraction and identification of reviewer signatures and in-text citations.[2] Both are especially affected by the varying style specifics and could not yet be automatically identified with sufficient precision; but the rapid technological developments will likely make this much easier in the near future. Likewise, the documents which had to be excluded due to LLM hallucinations can be likely addressed by newer versions soon.

However, overall the corpus converged to a state deemed as sufficient to be searched and displayed. They have been publicly released since March 2026, with a remark indicating their status and a feedback option for possible corrections. Gradually, they are checked and receive eventually a flag of an editorial approval (so far, in the course of the testing, this has been done for 31,152 documents). Only then, they will be also available through the zbMATH Open APIs [2, 9], via our standard CC-BY-SA 4.0 licence. However, for research purposes, the full data of the automated processing are already available – the Mathpix results at [7] and the data after the LLM split at [6].

6 Conclusions

Transforming more than 250,000 scanned pages with a high accuracy into LaTeX demonstrates a useful specific application of now standard AI technology tailored to a longstanding challenge.

[2] It might be noted that both issues also still mainly remain unsolved for the manually typed JFM volumes!

Trudinger, N. S.

On Harnack type inequalities and their application to quasilinear elliptic equations. (English)

Zbl 0153.42703

Commun. Pure Appl. Math. 20, 721-747 (1967).

Let $\Omega \subset E^n$ be a domain, let $A : \bar{\Omega} \times E \times E^n \to E^n, B : \Omega \times E \times E^n \to E$ be measurable and let $\operatorname{div} A = \sum_{i=1}^n \frac{\partial A}{\partial x_i}, x \in \Omega$. Suppose that $\alpha > 1, a_0, b_0$ are constants and $a_1(x), b_1(x)$ are nonnegative measurable functions and that, for $M < \infty$ and for $(x, u, p) \in \Omega \times (-M, M) \times E^n$:

$$|A(x,u,p)| \le a_0|p|^{\alpha-1} + |a_1(x)u|^{\alpha-1} + (a_3(x))^{\alpha-1},$$
$$pA(x,u,p) \ge |p|^\alpha - |a_2(x)u|^\alpha - (a_4(x))^\alpha,$$
$$|B(x,u,p)| \le b_0|p|^\alpha + b_1(x)|p|^{\alpha-1} + (b_2(x))^\alpha|u|^{\alpha-1} + (b_3(x))^\alpha.$$

A function $u(x)$ is a weak solution of (1) $\operatorname{div} A(x,u,u_x) + B(x,u,u_x) = 0$ if $u(x) \in W^2_\alpha, loc(\Omega)$ and $\int_\Omega \{\varphi_x A(x,u,u_x) - \varphi B(x,u,u_x)\}\, dx = 0$ for all positive $\varphi \in W^2_{\alpha,0}$. It is a weak super-(sub-) solution of (1) if equality is replaced by $\geq (\leq)$. Then, if $|a_1(x)|, |b_y(x)| < \mu$ in Ω, Theorem 1.1: Let $u(x)$ be a weak solution of (1) in a cube $K(3\rho) \subset \Omega$ of side 3ρ and suppose $0 \le u \le M$ in $K(3\rho)$. Then

$$\operatorname*{ess\ max}_{K(\rho)} u(x) \le C \operatorname*{ess\ min}_{K(\rho)} u(x), C = C(\alpha, n, a_0, b_0 M, \mu\rho).$$

Theorem 1.2: Let $u(x)$ be a weak super-solution of (1) in K and $0 \le u < M$ in $K(3\rho)$. Then $p^{-n/\gamma}\|u(x)\|_{\gamma,K(2\rho)} \le C \min_{K(\rho)} u(x)$ for $\gamma < n(\alpha-1)/(n-\alpha)$ when $\alpha \le n$, and $\gamma \le \infty$ when $\alpha > n$. Theorem 1.3: Let $u(x)$ be a weak sub-solution of (1) in $K, 0 \le u < M$ in $K(3\rho)$. Then ess $\max u(x) \le Cp^{-n/\gamma}\|u(x)\|_{\gamma,K}$ (2ρ). These results are now applied to obtain the Hölder estimate in Theorem 2.2: Let $u(x)$ be a locally bounded weak solution of (1) in Ω. Then if $u(x)$ is locally Holder continuous in Ω (up to a set of measure zero) and if $K = K(\rho_0) \subset \Omega$ then

$$\operatorname*{osc}_K u \le c\left(\frac{\rho}{\rho_0}\right)^\delta \left\{\sup_K |u| + \mu\rho(\mu\rho)^{\alpha-1}\right\}$$

for all $p < p_0$, where $\delta > 0$ and $C = C(\alpha, n, a_0, b_0, \sup_K u)$. Theorem 2.3 embodies a similar result for a type of quasi-elliptic equation. Next, for the uniformly elliptic equation $a^{1j}(x, u_x)\, u_{x_1 x_j} = 0$ with $a^{1j}(x,p) \in C^1(\Omega \times E^n)$ and

$$(1+|p|)a^{10}_p(x,p) + |a^{10}_x(x,p)| \le \lambda^{-1} \text{ for } (x,p) \in \Omega \times E^n$$

we have the following solution of the Dirichlet problem. Theorem 3.1: Let $\partial\Omega$ satisfy an exterior sphere condition and let $\varphi(x) \in C^\circ(\partial\Omega)$. Then there is a $u \in C^\circ(\bar{\Omega}) \cap C^2(\Omega)$ such that $a^{1j}(x, u_x)\, u_{x_1 x_j} = 0, u = \varphi$ in $\partial\Omega$. The paper concludes with some estimates of the solution of (1) at the boundary of Ω and with generalizations of the preceding when the functions a_1, b_1 are no longer bounded but subject only to certain integrability conditions.

This review text is based on a scanned copy of the printed version. It was converted to LaTeX using MathPix and a specifically developed LLM to assign the text parts to the metadata. It may contain errors, misassignments, or gaps; in particular, the reviewer signature has not yet been extracted reliably in general. If you notice any errors, please report them directly to our editorial team via the Contact Form.

Digitised review in zbMATH Open.

Compared to previous approaches, it obtained the goal with significantly reduced costs: Less than 4% of projected costs of traditional OCR/retyping approaches, consisting mainly of licence costs and support for the second author's thesis project at the University of Göttingen.[3] To the best of our knowledge, this might be the largest project of this type so far; and although the technology is already very advanced, human oversight could still provide valuable feedback and corrections.

References

[1] M. Beck, I. Beckenbach, T. Hartmann, M. Schubotz and O. Teschke, Transforming scanned zbMATH volumes to LaTeX: planning the next level digitisation. *Eur. Math. Soc. Newsl.* **117**, 49–52 (2020)

[2] M. Fuhrmann and F. Müller, A REST API for zbMATH Open access. *Eur. Math. Soc. Mag.* **130**, 63–65 (2023)

[3] M. Fuhrmann, P. Scharpf and M. Schubotz, Entity linking for zbMATH open. *Eur. Math. Soc. Mag.* **134**, 61–63 (2024)

[4] K. Hulek and O. Teschke, The transition of zbMATH towards an open information platform for mathematics. *Eur. Math. Soc. Newsl.* **116**, 44–47 (2020)

[3] These extra costs could actually be covered due to reviewers donating their compensation to the further development of zbMATH Open, to whom we are grateful!

[5] K. Hulek and O. Teschke, The transition of zbMATH towards an open information platform for mathematics II: A two-year progress report. *Eur. Math. Soc. Mag.* **125**, 44–47 (2022)

[6] I. Pluzhnikov, A. Satpute and M. Schubotz, Document texts extraction of zbMATH Open volumes (1–529). URL https://doi.org/10.5281/zenodo.14496349 (2024)

[7] I. Pluzhnikov, A. Satpute and M. Schubotz, Retrodigitalization of zbMATH Open volumes (1–529): Mathpix 2024 OCR results. URL https://doi.org/10.5281/zenodo.14253145 (2024)

[8] I. Pluzhnikov, A. Satpute, M. Schubotz, O. Teschke and B. Gipp, LLM-supported document separation for printed reviews from zbMATH Open. arXiv:2604.00554v1 (2026)

[9] M. Schubotz and O. Teschke, zbMATH Open: towards standardized machine interfaces to expose bibliographic metadata. *Eur. Math. Soc. Mag.* **119**, 50–53 (2021)

[10] O. Teschke, The "Jahrbuch über die Fortschritte der Mathematik" as a part of zbMATH Open. *Eur. Math. Soc. Mag.* **122**, 62–64 (2021)

[11] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, LoRa: low-rank adaptation of large language models. arXiv:2106.09685v2 (2021)

Bela Gipp studied computer science at the University of Magdeburg and UC Berkeley. After completing his PhD, he held postdoctoral positions at UC Berkeley and the National Institute of Informatics in Tokyo. He later served as a professor at the University of Konstanz and the University of Wuppertal before joining the University of Göttingen, where he is chair of scientific information analytics and serves as director of research at the Göttingen State and University Library (SUB). His research focuses on methods in natural language processing and blockchain technology.

gipp@uni-goettingen.de

Ivan Pluzhnikov finished his master's thesis, "LLM-supported document separation for printed reviews from zbMATH Open," in the research group of Bela Gipp, Georg August University of Göttingen, Germany. He plans to support FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Berlin, Germany in integrating his research results into the information service zbMATH Open.

ivanpluzhnikov@proton.me

Moritz Schubotz studied physics and computer science at the University of Wuppertal and TU Berlin. After his PhD, he was a postdoc at the University of Konstanz and the National Institute of Informatics in Tokyo before joining FIZ Karlsruhe in 2019. At FIZ Karlsruhe, he researches methods to improve zbMATH Open and related research infrastructure.

moritz.schubotz@fiz-karlsruhe.de

Olaf Teschke studied mathematics at Humboldt University Berlin, and completed his PhD in algebraic geometry there. He moved to FIZ Karlsruhe in 2008, and has been working there since 2009 as the head of the Mathematics Department and managing editor of zbMATH (including a short intermediate term as editor-in-chief). He has also served in several EMS committee positions. His main occupation is information infrastructure for mathematics.

olaf.teschke@fiz-karlsruhe.de