Transforming Scanned zbMATH Volumes to \LaTeX:
Planning the Next Level Digitisation

Marco Beck (University of Wuppertal, Germany), Isabel Beekenbach (FIZ Karlsruhe, Germany), Thomas Hartmann (FIZ Karlsruhe, Germany), Moritz Schubotz (FIZ Karlsruhe, Germany) and Olaf Teschke (FIZ Karlsruhe, Germany)

1 \LaTeX conversion as the next essential step for math digitisation

Since the advent of the internet, mathematicians have pursued the vision of a comprehensive, open and accessible digital collection of mathematical resources. The International Mathematical Union (IMU) supports this goal under the brand of the Global Digital Mathematics Library, and the diverse activities are fostered by the International Mathematics Knowledge Trust (IMKT). Obviously, this aim is not likely to be achieved in the near future – several technical and legal obstacles need to be overcome. The extent of the mathematics literature alone has been estimated at > 120 million pages of very diverse status [3]. However, the first technical step of digitisation – namely, the scanning of the existing mathematics literature – has been mostly successful due to public and private efforts: about 60% of the pages with mathematics literature are now available in some digital form [3]. However, scanned files have many limitations – although they can be read by most humans, their content is neither easily searchable nor machine-processable (e.g., for content analysis). Since mathematical content is intrinsically linked to formulæ, it would be highly desirable to have \LaTeX sources available. Unfortunately, this is not even the case for most digitally born documents (with only a few exceptions, most notably the arXiv) – currently, less than 3% of all maths pages are available as \LaTeX [3]. Transforming scans to \LaTeX is still challenging and costly. When the *Jahrbuch über die Fortschritte der Mathematik* was digitised, it required massive investments into typesetting formulæ (as well as correcting OCR errors). During the last years, MathOCR technology has made considerable advances, but it is still far from being seamlessly applicable in a scalable way. We will report here on the current state of the art for MathOCR with a view towards transforming the zbMATH volumes 1–529 (currently mostly just available as scans) into \LaTeX. Due to the diversity of content and formulæ in zbMATH, as well as of the types through the decades, we believe that this could serve as a meaningful representative model for the full corpus of mathematical literature.

2 Hoards of scanned reviews in zbMATH

Currently, zbMATH, and the aforementioned volumes in particular, includes more than 800,000 reviews and abstracts that exist as scanned images alone. Those items are distributed over 250,000 pages. For today’s zbMATH users, the usage experience for those items is not satisfying. For example, the fonts are hard to read, one cannot search for text, and copy and pasting of text is not possible. Moreover, the text is inaccessible for people with disabilities and also for information retrieval systems. Consequently, those reviews do not occur in recommendations and are not considered while scanning new articles for plagiarism. To improve this situation, we have manually transcribed about 15,000 abstracts over the past years. These reviews are now available to zbMATH users in the known digital form. Based on this experience, we estimate that the effort for manual retro-digitisation is immense. Outsourcing the \LaTeX processing part would cost about half a million Euro, and will take several years, and is thus infeasible. However, recent advancements in computer science, in particular deep learning, might drastically reduce the effort. In this paper, we will discuss our plans to use modern deep learning approaches to digitise the rest of the scanned images with reduced manual effort and discuss the challenges we foresee in this context.

3 Ingredients for digitisation

In this section, we will elaborate on the building blocks to retro-digitise past reviews before we discuss an example in the next section.

What is digitisation about?
The exponentially increased options for storing, transmitting, processing, linking, interpreting and reproducing information through digitisation have also set in motion fundamental transformation processes in science. Digitisation enables an open culture of innovation, in which data, information and ideas can be freely introduced and exchanged. If scientific literature can be digitally accessed from anywhere, the question arises as to whether printed research literature can be retrospectively digitised and made usable. For example, publications can be recorded using a scanner so that its software generates an image file in which the image is displayed as a raster graphic. A disadvantage of this form of digitalisation is that the quality of the image file depends on the scanning hardware and the paper of the original document, and often due to the book form, the text lines are not displayed straight. On the other hand, the font size and line spacing of printed or scanned articles are usually minimised due to the number of pages, cf. Figure 1. Therefore, digitised publications are often associated with poor readability and very limited further processing. Research literature in digital form is computer-generated liter-
Assess conversion quality

After converting the scanned images to \( \text{LaTeX} \) code, we have to assess the quality of our results. A usual measure in OCR would be the character error rate (CER), which makes sense for plain text. However, in a mathematical context it is not clear how to measure the similarity of two given formulæ in \( \text{LaTeX} \). We do not need exact character matches as long as the semantic meaning of a formula stays unchanged.

In the literature, several metrics are used to evaluate mathematical formula recognition algorithms. One possible metric is the recognition rate for complete expressions or individual symbols. This metric states whether two expressions or symbols are the same or different. There are also more refined metrics which look at different kinds of errors and weigh them or compare the symbol layout trees of mathematical expressions.

Sain, Dasgupta, Abhishek and Garain develop in [5] a method that compares the structure of two mathematical formulæ given in MathML. They convert the given formulæ into ordered trees and measure their similarity by the tree edit distance of the associate trees.

Another approach captures both the individual symbols of an expression and its structure as a bipartite graph [9]. Then different metrics are defined to measure the similarity of two bipartite graphs.

In our setting, we need metrics that directly compare \( \text{LaTeX} \) strings. We investigate the metrics mentioned above and compare them. If they turn out to be unsuitable for our purpose, we develop a new metric.

New search opportunities and TDM tools in copyright review

Our planned encompassing digital development of zbMATH opens up extensive research and use opportunities for scientists. The new conditions, introduced in 2019 and 2020, of the EU Copyright Directive and the German Copyright Law opens up extensive research and use opportunities for scientists. This new approach to copyright law would be the character error rate (CER), which makes sense for plain text. However, in a mathematical context it is not clear how to measure the similarity of two given formulæ in \( \text{LaTeX} \). We do not need exact character matches as long as the semantic meaning of a formula stays unchanged.

In the literature, several metrics are used to evaluate mathematical formulæ recognition algorithms. One possible metric is the recognition rate for complete expressions or individual symbols. This metric states whether two expressions or symbols are the same or different. There are also more refined metrics which look at different kinds of errors and weigh them or compare the symbol layout trees of mathematical expressions.

Sain, Dasgupta, Abhishek and Garain develop in [5] a method that compares the structure of two mathematical formulæ given in MathML. They convert the given formulæ into ordered trees and measure their similarity by the tree edit distance of the associate trees.

Another approach captures both the individual symbols of an expression and its structure as a bipartite graph [9]. Then different metrics are defined to measure the similarity of two bipartite graphs.

In our setting, we need metrics that directly compare \( \text{LaTeX} \) strings. We investigate the metrics mentioned above and compare them. If they turn out to be unsuitable for our purpose, we develop a new metric.
Regarding the mathematical formulæ, the human reviewer made a typographical error in the summation in line 3 which was used instead of the correct identifier $n$. While both MathOCR solutions recognised this correctly, Infy recognised an instead of an $\subseteq$ sign in line 3. Both Mathpix and Infy did not correctly recognise lim sup$_{n \to \infty}$ in line 20. They did split limit and sup as separate operators. Additionally, the very last line of the formula \( \subseteq \) was correctly recognised by both systems, but was falsely manually transcoded as $\subseteq$. In this example, Infy had several issues with sub and superscripts

\[
\begin{align*}
a_1 & \to a_2 \\
b_1 & \to b_2 \\
a_3 & \to a_4 \\
b_5 & \to b_3 \\
l^3 & \to l^1.
\end{align*}
\]

Besides these optical differences there are several differences in the \LaTeX{} code that generate identical or very similar output. This is different delimiter expectations discussed below, a different encoding of spaces, different ways of switching between maths and text mode, and different encodings of dot arrows and mathematical operators such as the sum sign (cf. line 21, Mathpix). To fully normalise these issues, either \LaTeX{} grammar parser (like texvc) included in mathoid [6] is required, or the \LaTeX{} code needs to be converted to MathML to simplify comparison by using prebuilt APIs such as the math tools [2].

1. http://www.infinityproject.org
2. https://mathpix.com

Figure 2. Comparison of different approaches to infer \LaTeX{} code from a scanned example image

project coordinators will be provided with legal information and recommendations to be taken into account in the context of new usages such as TDM.

4 An example

After having introduced the basic ingredients in the last section, we will now discuss the conversion quality based on the example shown in Figure 1. Figure 2 compares the output of Infy project 1 (left), the manual conversion (middle), and Mathpix 2 (right) for the select example. In the image, mistakes are highlighted in red, whereas alternative formatting is highlighted in orange. The begin and start delimiter such as $\langle \rangle$, \langle, \rangle as well as irrelevant grouping braces \{, \} and \left\langle \right\rangle combinations of brackets are shown in gray. Unfortunately, the output of Mathpix did not compile for this example, as the left and right brackets were not balanced. In particular, the opening left bracket in line 20 ends with an invisible right bracket in line 21 which is not in maths mode. Other than that the more excessive use of the left-right version of brackets by Mathpix did not create a visual difference for the selected example. For the text recognition, Mathpix recognised an additional s in the word ‘series’ line 1, and missed a space in ‘as its’ in line 21. This was also incorrectly recognised by Infy. Infy has spelling issues for the word ‘infinite sequence’ in line 12 and ‘deduces’ line 19. Moreover, Infy recognised the word ‘From’ as formula \Gamma (line 19).

Listing 1: Infy

\begin{verbatim}
1 \text{the author shows that the series} \\
2 \displaystyle \sum_{n=1}^{\infty} \Gamma(n) (1/n) \infty \text{is Euler’s function, and the sum of divisors}.
3 \text{The proof depends on a general lemma 1 on irrational series, and on these properties}:
4 \text{there are only $O(x)$ integers $n$ satisfying \( \frac{\varphi(n)}{n} \approx \frac{1}{x} \)}.
5 \text{Finally assume there exists a constant $c>0$ as \( \frac{m_i}{k} \infty \text{a constant).}
6 \text{Lemma 1 is obtained as a special case of the more general lemma 4:}
7 \text{Let there exist an infinite sequence of integers $\geq 0$ such that $a_k \leq k^s$ and $\varphi(k)$ irrational.}
8 \text{When $i_1$ and $i_2$ are consecutive suffixes of $\Gamma(n)$, then there is a $k$ with \( i_1 < k < i_2 \)}.
9 \text{Under these conditions all series \( \sum_{k=1}^{m_i} \frac{a_k + b_k}{t^k} \) are irrational.}
10 \text{The author deduces Theorem 2:}
11 \text{Let \( \{a_{\alpha}\} \text{ and } \{b_{\lambda}\} \text{ be two infinite sequences of integers } \geq 0 \text{ such that } a_k \leq k^s }.
12 \text{The author shows that the series \( \sum_{n=1}^{\infty} t^{-\varphi(n) -\varphi (n)} \) is Euler’s function, and the sum of divisors.}\n\end{verbatim}

Listing 2: Manual transcript

\begin{verbatim}
1 \text{the author shows that the series} \\
2 \displaystyle \sum_{n=1}^{\infty} \Gamma(n) (1/n) \infty \text{is Euler’s function, and the sum of divisors}.
3 \text{The proof depends on a general lemma 1 on irrational series, and on these properties}:
4 \text{there are only $O(x)$ integers $n$ satisfying \( \frac{\varphi(n)}{n} \approx \frac{1}{x} \)}.
5 \text{Finally assume there exists a constant $c>0$ as \( \frac{m_i}{k} \infty \text{a constant).}
6 \text{Lemma 1 is obtained as a special case of the more general lemma 4:}
7 \text{Let there exist an infinite sequence of integers $\geq 0$ such that $a_k \leq k^s$ and $\varphi(k)$ irrational.}
8 \text{When $i_1$ and $i_2$ are consecutive suffixes of $\Gamma(n)$, then there is a $k$ with \( i_1 < k < i_2 \)}.
9 \text{Under these conditions all series \( \sum_{k=1}^{m_i} \frac{a_k + b_k}{t^k} \) are irrational.}
10 \text{The author deduces Theorem 2:}
11 \text{Let \( \{a_{\alpha}\} \text{ and } \{b_{\lambda}\} \text{ be two infinite sequences of integers } \geq 0 \text{ such that } a_k \leq k^s }.
12 \text{The author shows that the series \( \sum_{n=1}^{\infty} t^{-\varphi(n) -\varphi (n)} \) is Euler’s function, and the sum of divisors.}\n\end{verbatim}

Listing 3: Mathpix

\begin{verbatim}
1 \text{the author shows that the series} \\
2 \displaystyle \sum_{n=1}^{\infty} \Gamma(n) (1/n) \infty \text{is Euler’s function, and the sum of divisors}.
3 \text{The proof depends on a general lemma 1 on irrational series, and on these properties}:
4 \text{there are only $O(x)$ integers $n$ satisfying \( \frac{\varphi(n)}{n} \approx \frac{1}{x} \)}.
5 \text{Finally assume there exists a constant $c>0$ as \( \frac{m_i}{k} \infty \text{a constant).}
6 \text{Lemma 1 is obtained as a special case of the more general lemma 4:}
7 \text{Let there exist an infinite sequence of integers $\geq 0$ such that $a_k \leq k^s$ and $\varphi(k)$ irrational.}
8 \text{When $i_1$ and $i_2$ are consecutive suffixes of $\Gamma(n)$, then there is a $k$ with \( i_1 < k < i_2 \)}.
9 \text{Under these conditions all series \( \sum_{k=1}^{m_i} \frac{a_k + b_k}{t^k} \) are irrational.}
10 \text{The author deduces Theorem 2:}
11 \text{Let \( \{a_{\alpha}\} \text{ and } \{b_{\lambda}\} \text{ be two infinite sequences of integers } \geq 0 \text{ such that } a_k \leq k^s }.
12 \text{The author shows that the series \( \sum_{n=1}^{\infty} t^{-\varphi(n) -\varphi (n)} \) is Euler’s function, and the sum of divisors.}\n\end{verbatim}
5 The road ahead

We are planning to apply for grant money to eliminate the dark spot of scanned but not fully digitised reviews in zbMATH. As a supplement to zbMATH Open, we are planning to investigate the capabilities of MathOCR tools further and combine the strength of them. With well-defined evaluation metrics, we will be able to continuously improve the conversion quality and involve the community in the final human judgment of the quality. We are committed to the FAIR and open principles, and therefore we will document share and open-source our developments. Thus not only the more than 250,000 zbMATH pages will become available as \LaTeX code, but also the effort required to convert from scans to \LaTeX code will decrease for follow-up projects. We will pave the road for more than 100 million pages of mathematical literature that is not available as \LaTeX code to eventually become digital.

References

[1] P. Erdős. “On the Irrationality of Certain Series”. In: Nederlandse Akademie van Wetenschappen. Proceedings. Series A. Indagationes Mathematicae 60 (1957), pp. 212–219.

[2] A. Greiner-Petter et al. “MathTools: An Open API for Convenient MathML Handling”. In: Intelligent Computer Mathematics. Ed. by F. Rabe et al. Vol. 11006. Cham: Springer International Publishing, 2018, pp. 104–110.

[3] P. Ion and O. Teschke. “Continuing toward a Global Digital Mathematics Library”: AMS Special Session on Mathematical Information in the Digital Age of Science at the Joint Mathematics Meetings (San Diego, CA). 2018.

[4] S. A. Oliveira, B. Seguin, and F. Kaplan. “dhSegment: A Generic Deep-Learning Approach for Document Segmentation”. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR) (2018), pp. 7–12. arXiv: 1804.10371.

[5] K. Sain, A. Dasgupta, and U. Garain. “EMERS: A Tree Matching–Based Performance Evaluation of Mathematical Expression Recognition Systems”. In: International Journal on Document Analysis and Recognition (IJDAR) 14.1 (2011), pp. 75–85.

[6] M. Schubotz and G. Wicke. “Mathoid: Robust, Scalable, Fast and Accessible Math Rendering for Wikipedia”. In: Intelligent Computer Mathematics. Ed. by S. M. Watt et al. Vol. 8543. Cham: Springer International Publishing, 2014, pp. 224–235.

[7] R. Simanowski. “Autorschaften in digitalen Medien. Einleitung”. In: Text + Kritik. 152. 2001, pp. 3–21.

[8] R. Simanowski. “Reading Digital Literature A Subject Between Media and Methods”. In: Reading Moving Letters. Ed. by R. Simanowski, J. Schäfer, and P. Gendolla. Bielefeld: transcript Verlag, 2010.

[9] R. Zanibbi et al. “Stroke-Based Performance Metrics for Handwritten Mathematical Expressions”. In: 2011 International Conference on Document Analysis and Recognition, ICDAR 2011, Beijing, China, September 18–21, 2011. IEEE Computer Society, 2011, pp. 334–338.