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

Marco Beck (University of Wuppertal, Germany), Isabel Beckenbach (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

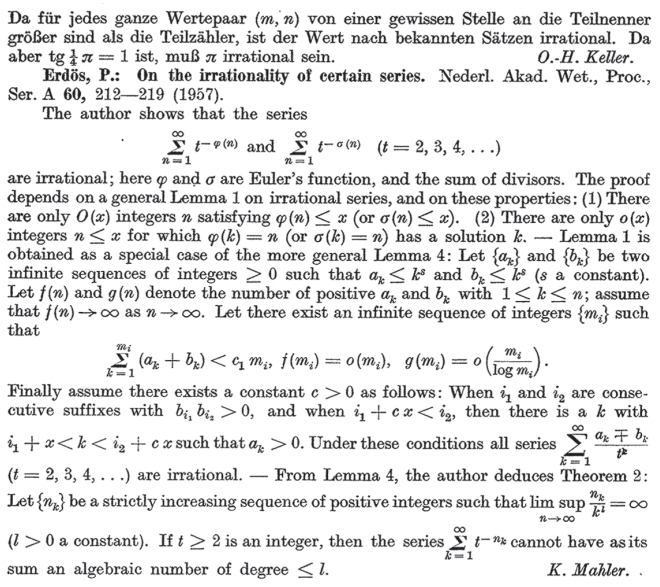
Currently, zbMATH, and the aforementioned volumes in particular, includes more than 800 000 reviews and abstracts that exist as scanned images alone. Those items are distributed over 250 000 pages. For today's zbMATH users, the usage experience for those items is not satisfying. For example, the fonts 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 retrodigitise 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 digitaliation 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-



Da für jedes ganze Wertepaar (m, n) von einer gewissen Stelle an die Teilnenner größer sind als die Teilzähler, ist der Wert nach bekannten Sätzen irrational. Da aber $\operatorname{tg} \frac{1}{4}\pi = 1$ ist, muß π irrational sein. *O.-H. Keller.*

Erdös, P.: On the irrationality of certain series. Nederl. Akad. Wet., Proc., Ser. A **60,** 212—219 (1957).

The author shows that the series

$$\sum_{n=1}^{\infty} t^{-\varphi(n)} \text{ and } \sum_{n=1}^{\infty} t^{-\sigma(n)} \quad (t = 2, 3, 4, \ldots)$$

are irrational; here φ and σ are Euler's function, and the sum of divisors. The proof depends on a general Lemma 1 on irrational series, and on these properties: (1) There are only $O(x)$ integers n satisfying $\varphi(n) \leq x$ (or $\sigma(n) \leq x$). (2) There are only $o(x)$ integers $n \leq x$ for which $\varphi(k) = n$ (or $\sigma(k) = n$) has a solution k. — Lemma 1 is obtained as a special case of the more general Lemma 4: Let $\{a_k\}$ and $\{b_k\}$ be two infinite sequences of integers ≥ 0 such that $a_k \leq k^s$ and $b_k \leq k^s$ (s a constant). Let $f(n)$ and $g(n)$ denote the number of positive a_k and b_k with $1 \leq k \leq n$; assume that $f(n) \to \infty$ as $n \to \infty$. Let there exist an infinite sequence of integers $\{m_i\}$ such that

$$\sum_{k=1}^{m_i} (a_k + b_k) < c_1 m_i, \; f(m_i) = o(m_i), \; g(m_i) = o\left(\frac{m_i}{\log m_i}\right).$$

Finally assume there exists a constant $c > 0$ as follows: When i_1 and i_2 are consecutive suffixes with $b_{i_1} b_{i_2} > 0$, and when $i_1 + c\,x < i_2$, then there is a k with $i_1 + x < k < i_2 + c\,x$ such that $a_k > 0$. Under these conditions all series $\sum_{k=1}^{\infty} \frac{a_k \mp b_k}{t^k}$ ($t = 2, 3, 4, \ldots$) are irrational. — From Lemma 4, the author deduces Theorem 2: Let $\{n_k\}$ be a strictly increasing sequence of positive integers such that $\limsup_{n\to\infty} \frac{n_k}{k^l} = \infty$ ($l > 0$ a constant). If $t \geq 2$ is an integer, then the series $\sum_{k=1}^{\infty} t^{-n_k}$ cannot have as its sum an algebraic number of degree $\leq l$. *K. Mahler.*

Figure 1. Excerpt from Zentralblatt für Mathematik, Series 79, page 74. Zbl 0079.07401 [1]

ature which, in contrast to digitised literature, is existentially dependent on the digital medium [8, p. 15–28]. It is characterised by at least one of the specific features of digital media: interactivity, intermediality and staging [7, p. 3–21]. Accordingly, digital literature is described as hypertextual and multimedia rather than as a self-contained work. If we transfer this to zbMATH reviews and abstracts, this means that formulæ from the publication can be copied into Word/LaTeX for further processing. An additional advantage of digital research literature is the adaptation of page and font size as well as fonts to the respective end device so that the reader is provided with better usability. A manual full-text entry, i.e., typing the complete text and formulæ, is hardly conceivable due to the high personnel costs and the high error rate.

Associate image and metadata

A first problem that arises when digitising the scanned images of zbMATH reviews and abstracts is that they are not yet separated. Namely, there is usually more than one review on a scanned page and a review can start on one page and end on another. Thus, we first have to solve a special segmentation problem where we have to split the scanned pages into individual reviews and associate them to their corresponding zbMATH entry. We know that the review of an article starts with the author names and the title of the given article and ends with the name of the reviewer. The author names, title and name of the reviewer of an article are metadata in our database. Thus, all this information is already given in a digital form and known in advance. This is very different from other segmentation problems in the field of OCR, as usually the next section is unknown.

In recent years, machine learning approaches, in particular deep recurrent and convolutional neural networks were successfully applied to document segmentation problems, see for example [4]. We investigate how these methods can be applied to our setting and how we can make use of the special metadata information we have.

Convert images to LaTeX code

According to our research, there are two major commercial services that specialised in the conversion of images of mathematical documents to LaTeX code. The Infty project from Japan is a research-driven project that has been developing technology to convert scanned images to LaTeX code since 1998. A new alternative is the MathPix project, which provides an online API as well as a desktop app to convert images of mathematical formulæ to LaTeX source code. While the desktop app was originally designed to convert individual formulæ, a recent API addon was published to allow conversion of whole pages at once. In contrast to the Infty Reader, Mathpix does not link an extracted fragment to the position in its image. See Section 3 for an exemplary comparison of the two selected commercial services.

Assess conversion quality

After converting the scanned images to 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 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 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 have been integrated into this development. In this respect, we want to conduct a legal investigation into the concrete meaning of new legal provisions for Text and Data Mining (TDM), deep learning techniques and digital research for the digital expansion of the directory of mathematical literature. With the investigation, we intend to make clear how the current approach to copyright law will impact open access transformation projects. In addition, editors, journal managers and

Listing 1: Infty

```
The author shows that the series
$\displaystyle \sum_{n\Rightarrow1}^{\infty}
   t^{-\varphi(n)}$ and $\displaystyle
   \sum_{n=1}^{\infty} t^{-\sigma(n)} (t=2,3,4,\ .\ .\
   .)$
are irrational; here $\varphi$ and $\sigma$ are
   Euler's function, and the sum of divisors.
The proof depends on a general Lemma 1 on irrational
   series, and on these properties:
(1) There are only $0(x)$ integers $n$ satisfying $
   \varphi(n)\leq x$ (or $\sigma(n)\leq x$).
(2) There are only $o(x)$ integers $n\leq x$ for which
   $\varphi(k)=n$ (or $\sigma(k)=n$) has a solution $
   k$. --
Lemma 1 is obtained as a special case of the more
   general Lemma 4:
Let $\{a_{\alpha}\}$ and $\{b_{k}\}$ be two infinite
   sequences of integers $\geq 0$ such that $a_{k}\leq
   k^{s}$ and $b_{k}\leq k^{s}$ ($s$ a constant).

Let $f(n)$ and $g(n)$ denote the number of positive $
   a_{k}$ and $b_{k}$ with $1\leq k\leq n$; assume
   that $ f(n)\rightarrow\infty$ as $
   n\rightarrow\infty$.
Let there exist an infimte sequence of integers $
   \{m_{i}\}$ such that
$$
\sum_{k=1}^{m_{i}}(a_{k}+b_{k})<\mathrm{q}m_{i},\
   f(m_{i})=o(m_{i}),\ g(m_{i})=o(\frac{m_{i}}{\log
   m_{i}})\ .
$$
Finally assume there exists a constant $c>0$ as
   follows:
When $i_{1}$ and $i_{2}$ are conse-cutive suffixes
   with $b_{i_{1}}b_{i_{-}}. >0$, and when $
   i_{1}+cx<i_{2}$, then there is a $k$ with $
   i_{1}+x<k<i_{2}+cx$ such that $a_{\gamma}>0$.
Under these conditions all series $\displaystyle
   \sum_{k=1}^{\infty}\frac{a_{k}\mp
   b_{\lambda}}{t^{1}}$ $(t=2,3,4,\ .\ .\ .)$ are
   irrational.
$-\Gamma \mathrm{r}\mathrm{o}\mathrm{m}$ Lemma 4, the
   author deduoes Theorem 2:
Let $\{n_{k}\}$ be a strictly increasing sequence of
   positive integers such that $\displaystyle
   \lim_{n\rightarrow}\sup_{\infty}\frac{n_{k}}{k}=\infty
   ($l>0$ a constant).
If $t\geq 2$ is an integer, then the series $
   \displaystyle \sum_{k=1}^{\infty}t^{-n_{k}}$ cannot
   have asits sum an algebraic number of degree $\leq
   1$.
```

Listing 2: Manual transcript

```
The author shows that the series
\[
\sum_{u=1}^\infty t^{-\varphi (n)} \text{ and }
   \sum_{n=1}^\infty t^{-\sigma(n)}
\]
\((t=2,3,4,...)\)
are irrational; here \(\varphi\) and \(\sigma\) are
   Euler's function, and the sum of divisors.
The proof depends on a general Lemma 1 on irrational
   series, and on these properties:
(1) There are only \(0(x)\) integers \(n\) satisfying \(
   \varphi(n) \leq x\) (or \(\sigma(n) \leq x\)).
(2) There are only \(o(x)\) integers \(n \leq x\) for
   which \(\varphi(k) = n\) (or \(\sigma(k) = n\)) has a
   solution \(k\).
Lemma 1 is obtained as a special case of the more
   general Lemma 4:
Let \(\{a_k\}\) and \(\{b_k\}\) be two infinite sequences
   of integers \(\geq 0\) such that \(a_k \leq k^s\) and \(
   b_k \leq k^s\) (\(s\) a constant).

Let \(f(n)\) and \(g(n)\) denote the number of p0ositive \(
   a_k\) and \(b_k\) with \(1\leq k \leq n\); assume that \(
   f(n) \to \infty\) as \(n \to \infty\).

Let there exist an infinite sequence of integers \(
   \{m_i\}\) such that
\[
\sum_{k=1}^{m_i} (a_k+b_k) < c_1 m_i,\qquad f(m_i) =
   o(m_i),\qquad g(m_i) = o\left({m_i \over \log
   m_i}\right).
\]
Finally assume there exists a constant \(c>0\) as
   follows:
When \(i_1\) and \(i_2\) are consecutive suffixes with \(
   b_{i_1}b_{i_2} >0\), and when \(i_1+cx < i_2\), then
   there is a \(k\) with \(i_1+x < k < i_2+cx\) such that \(
   a_k > 0\).
Under these conditions all series \[
\sum_{k=1}^\infty {a_k \mp b_k \over t^k}\qquad
   (t=2,3,4,...)
\] are irrational.
From Lemma 4, the author deduces Theorem 2:

Let \(\{n_k\}\) be a strictly increasing sequence of
   positive integers such that \(\limsup_{n \to \infty}
   {n_k \over k^l} = \infty\) \((l>0\) a constant).
If \(t \geq 2\) is an integer, then the series \(
   \sum_{k=1}^\infty t^{-n_k}\) cannot have as its sum
   an algebraic number of degree \(\leq 1\).
```

Listing 3: Mathpix

```
The author shows that the seriess
\[
\sum_{n=1}^{\infty} t^{-\varphi(n)} \text { and }
   \sum_{n=1}^{\infty} t^{-\sigma(n)} \quad(t=2,3,4,
   \ldots)
\]
are irrational; here \( \varphi \) and \( \sigma \) are
   Euler's function, and the sum of divisors.
The proof depends on a general Lemma 1 on irrational
   series, and on these properties:
(1) There are only \( 0(x) \) integers \( n \) satisfying \(
    \varphi(n) \leq x(\text { or } \sigma(n) \leq x) \)
(2) There are only \( o(x) \) integers \( n \leq x \) for
   which \( \varphi(k)=n(\text { or } \sigma(k)=n) \)
   has a solution \( k .- \)
Lemma 1 is obtained as a special case of the more
   general Lemma 4:
Let \( \left\{a_{k}\right\} \) and \( \left\{b_{k}\right\} \)
   be two infinite sequences of integers \( \geq 0 \)
   such that \( a_{k} \leq k^{s} \) and \( b_{k} \leq
   k^{s} \) ( \( s \) a constant).
Let \( f(n) \) and \( g(n) \) denote the number of
   positive \( a_{k} \) and \( b_{k} \) with \( 1 \leq k
   \leq n \); assume that \( f(n) \rightarrow \infty \)
   as \( n \rightarrow \infty . \)
Let there exist an infinite sequence of integers \( \left
   \{m_{i}\right\} \) such that
\[
\sum_{k=1}^{m_{i}}\left(a_{k}+b_{k}\right)<c_{1} m_{i},
   \quad f\left(m_{i}\right)=o\left(m_{i}\right), \quad
   g\left(m_{i}\right)=o\left(\frac{m_{i}}{\log
   m_{i}}\right)\]
Finally assume there exists a constant \( c>0 \) as
   follows:
When \( i_{1} \) and \( i_{2} \) are consecutive suffixes
   with \( b_{i_{1}} b_{i_{2}}>0, \) and when \( i_{1}+c
   x<i_{2}, \) then there is a \( k \) with \(
   i_{1}+x<k<i_{2}+c x \) such that \( a_{k}>0 . \)
Under these conditions all series \(
   \sum_{k=1}^{\infty} \frac{a_{k} \mp b_{k}}{t^{k}} \)
   \( (t=2,3,4, \ldots) \) are irrational. \( - \)

From Lemma \( 4, \) the author deduces Theorem 2:

Let \( \left\{n_{k}\right\} \) be a strictly increasing
   sequence of positive integers such that \( \lim _{n
   \rightarrow \infty} \sup \frac{n_{k}}{k^{l}}=\infty
   \) \( \left(l>0 \text { a constant).
If } t \geq 2 \text { is an integer, then the series }
   \underset{k=1}{\stackrel{\infty}{\Sigma}} t^{-n_{k}}
\) cannot have asits \right. sum an algebraic number of
   degree \( \leq 1 \).
```

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 Infty 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 `$,$$,\(,\),\[,\]` as well as irrelevant grouping brakes `{}`, and `\left(\right]` 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 Infty. Infty has spelling issues for the word 'infinite sequence' in line 12 and 'deduces' line 19. Moreover, Infty recognised the word 'From' as formula Γrom (line 19).

Regarding the mathematical formulæ, the human reviewer made a typo in the summation in line 3 where u was used instead of the correct identifier n. While both MathOCR solutions recognised this correctly, Infty recognised an $\Rightarrow$ instead of an = sign in line 3. Both Mathpix and Infty did not correctly recognise $\limsup_{n\to\infty}$ in line 20. They did split limit and sup as separate operators. Additionally, the very last inline formula $\leq l$ was correctly recognised by both systems, but was falsely manually transcribed as ≤ 1. In this example, Infty had several issues with sub and superscripts

$$a_k \to a_\alpha \tag{l.9}$$
$$c_1 \to q \tag{l.14}$$
$$b_{i_2} \to b_{i\_} \tag{l.17}$$
$$a_k \to a_\gamma \tag{l.17}$$
$$b_k \to b_\lambda \tag{l.18}$$
$$t^k \to t^1. \tag{l.18}$$

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

[1] http://www.inftyproject.org
[2] https://mathpix.com

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.

Marco Beck [m.beck@beck-notz.de] is a doctoral researcher at the Data & Knowledge Engineering Group the University of Wuppertal. His main research interests are systematic use of computer-based methods and digital resources in the humanities and cultural sciences.

Isabel Beckenbach [isabel.beckenbach@fiz-karlsruhe.de] studied mathematics at FernUniversität Hagen, Technical University of Berlin, and Free University of Berlin. From 2013 to 2019 she worked at Zuse Institute Berlin in the optimization department where she completed her PhD in combinatorial optimization and graph theory. In 2020 she moved to FIZ Karlsruhe and now she works on the transition of zbMATH towards an open information platform for mathematics.

Thomas Hartmann [tho.hartmann@fiz-karlsruhe.de] is a researcher for intellectual property rights. He is specialised on copyright, licence management, data rights and legal support for open access.

Photos and CVs of the other authors can be found in previous Newsletter *issues.*