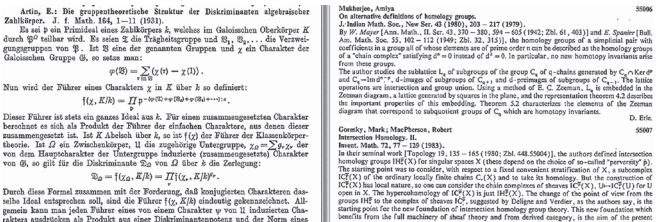
Four Decades of TEX at zbMATH

Moritz Schubotz (FIZ Karlsruhe, Berlin, Germany) and Olaf Teschke (FIZ Karlsruhe, Berlin, Germany)

In April 2019, zbMATH was completely transformed from TEX to LATEX sources. On this occasion, we give a brief history of typesetting Zentralblatt volumes, and describe the challenges, methods and benefits of the transition.

A very short overview of typesetting Zentralblatt

TEX and LaTeX have been the standard tools for creating documents for at least two generations of mathematicians. Today it is almost inconceivable that mathematical content was typeset before. Indeed, mathematical typesetting has a much longer history at zbMATH. zbMATH has existed for about 150 years if one includes its printed predecessors, *Jahrbuch über die Fortschritte der Mathematik* and *Zentralblatt für Mathematik*. Although it would be somewhat tempting to retell the history of mathematical typesetting in these periods based on the appearance of the *Jahrbuch* and *Zentralblatt* volumes in a similar manner as, e.g., [5], this would go beyond the scope of the present article. Instead, the figure below provides impressions of volumes based on various lead types, linotypes, or phototypesetting using IBM golfball typewriters.



Artin, E.: Die gruppentheoretische Struktur der Diskriminanten algebraischer Zahlkörper. J. f. Math. 164, 1–11 (1931).

Es sei $\mathfrak{p}$ ein Primideal eines Zahlkörpers k, welches im Galoischen Oberkörper K durch $\mathfrak{P}^e$ teilbar wird. Es seien $\mathfrak{T}$ die Trägheitsgruppe und $\mathfrak{V}_1, \mathfrak{V}_2, \ldots$ die Verzweigungsgruppen von $\mathfrak{P}$. Ist $\mathfrak{H}$ eine der genannten Gruppen und χ ein Charakter der Galoischen Gruppe $\mathfrak{G}$, so setze man:

$$\varphi(\mathfrak{H}) = \sum_{\tau \in \mathfrak{H}} (\chi(\tau) - \chi(1)).$$

Nun wird der Führer eines Charakters χ in K über k so definiert:

$$\mathfrak{f}(\chi, K/k) = \prod_{\mathfrak{p}} \mathfrak{p}^{\varphi(\mathfrak{T}) + \varphi(\mathfrak{V}_2) + \varphi(\mathfrak{V}_3) + \cdots}.$$

Dieser Führer ist stets ein ganzes Ideal aus k. Für einen zusammengesetzten Charakter berechnet es sich als Produkt der Führer der einfachen Charaktere, aus denen dieser zusammengesetzt ist. Ist K Abelsch über k, so ist $\mathfrak{f}(\chi)$ der Führer der Klassenkörpertheorie. Ist Ω ein Zwischenkörper, $\mathfrak{U}$ die zugehörige Untergruppe, $\chi_0 = \sum_{\tau \in \mathfrak{G}} \chi_{\tau}$ der von dem Hauptcharakter der Untergruppe induzierte (zusammengesetzte) Charakter von $\mathfrak{G}$, so gilt für die Diskriminante $\mathfrak{D}_\Omega$ von Ω über k die Zerlegung:

$$\mathfrak{D}_\Omega = \mathfrak{f}(\chi_0, K/k) = \prod \mathfrak{f}(\chi, K/k)^{\varphi}.$$

Durch diese Formel zusammen mit der Forderung, daß konjugierten Charakteren dasselbe Ideal entsprechen soll, sind die Führer $\mathfrak{f}(\chi, K/k)$ eindeutig gekennzeichnet. Allgemein kann man jeden Führer eines von einem Charakter ψ von $\mathfrak{U}$ induzierten Charakters ausdrücken als Produkt aus einer Diskriminantenpotenz und der Norm eines Führers von $\mathfrak{f}(\psi)$. Alle diese Formeln sind „elementar" beweisbar; nur zum Beweis, daß der Führer ein ganzes Ideal von k ist, braucht man die Klassenkörpertheorie.

Herbrand, J.: Sur les unités d'un corps algébrique. C. r. Acad. Sci. Paris 192, 24–27 (1931).

Mukherjee, Amiya 55006
On alternative definitions of homology groups.
J. Indian Math. Soc., New Ser. 43 (1980), 203 – 217 (1979).
By *W. Meyer* [Ann. Math., II. Ser. 43, 570 – 589, 594 – 605 (1942; Zbl. 61, 403)] and *E. Spanier* [Bull. Am. Math. Soc. 55, 102 – 112 (1949; Zbl. 32, 315)], the homology groups of a simplicial pair with coefficients in a group all of whose elements are of prime order n can be described as the homology groups of a "chain complex" satisfying $d^n = 0$ instead of $d^2 = 0$. In particular, no new homotopy invariants arise from these groups.
The author studies the subalition L_q of subgroups of the group C_q of q-chains generated by $C_q \cap \operatorname{Ker} d^n$ and $C_q \cap \operatorname{Im} d^{n-1}$, d-images of subgroups of C_{q+1}, and d-preimages of subgroups of C_{q-1}. The lattice operations are intersection and group union. Using a method of E. C. Zeeman, L_q is embedded in the Zeeman diagram, a lattice generated by squares in the plane, and the representation theorem 4.2 describes the important properties of this embedding. Theorem 5.2 characterizes the elements of the Zeeman diagram that correspond to subquotient groups of C_q which are homotopy invariants.
D. Erle.

Gorenky, Mark; MacPherson, Robert 55007
Intersection Homology. II.
Invent. Math. 72, 77 – 129 (1983).
In their seminal work [Topology 19, 135 – 165 (1980; Zbl. 448.55004)], the authors defined intersection homology groups $IH^{\bar{p}}(X)$ for singular spaces X (them depend on the choice of so-called "perversity" $\bar{p}$). The starting point was to consider, with respect to a fixed convenient stratification of X, a subcomplex $IC^{\bar{p}}_*(X)$ of the ordinary locally finite chains $C_*(X)$ and to take its homology. But the construction of $IC^{\bar{p}}_*(X)$ has local nature, so one can consider the chain complexes of sheaves $IC^{\bar{p}}_*(X)$, $U \mapsto IC^{\bar{p}}_*(U)$ for U open in X. The hypercohomology of $IC^{\bar{p}}(X)$ is just $IH^{\bar{p}}_*(X)$. The change of the point of view from the groups $IH^{\bar{p}}_*$ to the complex of sheaves $IC^{\bar{p}}_*$, suggested by Deligne and Verdier, as the authors say, is the starting point for the new foundation of intersection homology group theory. This new foundation which benefits from the full machinery of sheaf theory and from derived category, is the aim of the present paper.
The paper is divided into the following eight paragraphs: §1. Sheaf Theory, §2. IH, as Hypercohomology for P.L. Pseudo-manifolds), §3. Sheaf theoretic construction of $IC^{\bar{p}}_*$ §4. Topological invariance of $IC^{\bar{p}}_*$ §5. Basic Properties of IH, §6. The middle group, §7. Lefschetz Theorem on hyperplane sections, §9(!) Generalized Deligne construction and duality. An extensive bibliography is also given.
M. Stein.

Formulæ from Zentralblatt volumes 1 and 529.

In general, mathematics was always very expensive to typeset.[1] The various developments until the 1970s aimed to make this process more efficient. Famously, the quality of the more cost-efficient phototypesetting technique decreased in comparison to classical lead techniques [5]. Indeed, in the *Zentralblatt* volumes from the years of phototypesetting one comes across several formulae which needed to be added by hand, before one could start to create the phototypesetting master by cutting and glueing.

However, at *Zentralblatt*, the shift to phototypesetting was inevitable due to the growing amount of content, which resulted in the production of immense register volumes. This created the most obvious and urgent need for further digitisation.

From the beginnings of TEX to its adaptation

Donald Knuth developed the TEX system to overcome the limitations of phototypesetting. As a welcome side effect, the fact that TEX is an open system established the autonomy of mathematical writing. With hindsight, this appears to be an inevitable development, though in the transition period it was not. For instance, a subject of detailed debate was whether it would be more efficient to employ scientists as their typesetters, instead of having specialised staff [9].

A major milestone in this process was the TEX82 version with both its improvements and stability. The 1982 meeting of the TEX Users Group at Stanford was the first TUG meeting lasting for about a week. Incidentally, a *Zentralblatt* editor – who was involved at that time in the editors' exchange program with Mathematical Reviews[2] – took part, and advocated later in Berlin the adaptation of TEX for the production of the *Zentralblatt* volumes. At Mathematical Reviews TEX became fully productive as of 1985, after several years of preparation [7]. In contrast, it took until 1992 before the advent of TEX at *Zentralblatt*. In the meantime, the phototypesetting technique became no longer sustainable. In 1984, a proprietary, internal Springer system was employed that resembled many of the typesetting functions of TEX. The commands of the Springer system were designed for specialised technical staff instead of self-use. A key argument for the Springer system in contrast to TEX was the resulting lower number keystrokes for volume production.

From today's viewpoint, the greatest advantage of moving away from phototypesetting was that at least the texts and formulae were available in a digital form for the volumes 531–734. However, the disadvantage was that after the Springer system became outdated in the early 1990s, the introduction of TEX required conversion: simultaneously to the Springer-based production plan. Since the transition to TEX also included a migration from Springer servers to FIZ Karlsruhe, the expected delay occurred.[3] Fortunately, the TEX expertise acquired in the meantime allowed for a successful transition. It turned out that even most formulae could be translated automatically, though some constructions

1 Typically, there were only a few highly specialised typesetters involved. One anecdote that has passed through generations was that one single typesetter was responsible for producing Zentralblatt volumes for many years. Eventually, he was able to spot errors in mathematical formulae without any semantical knowledge.

2 That was also the time when a merger of both services was actively pursued.

3 The 3-month hiatus between Vol. 734 and 735 is by far the largest post-war gap.

had inherent problems; e.g., matrices needed conversion from column-based encoding to row-based encoding.[4]

From TEX to LATEX

For some years after the switch to TEX, the production of printed volumes had been the main objective. The main objective was the appropriate presentation. A standardised encoding, which is desirable from an information retrieval viewpoint, was less relevant. This changed in the second half of the 1990s after the online database became the primary objective. In 2004, adapting to these needs, the database production switched to a PostgreSQL system. This stored all available information in ASCII-coded TEX texts. This framework hadn't changed significantly until recently.

Of course, despite the impressive robustness and stability of TEX, technical development didn't stop in the 80s. Probably already in the mid-90s LATEX was the preferred dialect for many users. Today LATEX accounts for more than 90% of zbMATH review submissions. Since the database routines previously interacted with TEX, reviews in LATEX required a re-standardisation. Furthermore, an increasing amount of data is provided in the UTF-8 format. Its conversion to TEX encoding for internal storage and later reconversion for online presentation may cause errors and a loss of information. This pertains especially to bibliographic reference data which may include the need to encode native script such as Arabic, Chinese, Farsi, Hebrew, Japanese, Korean, or Russian. This makes it desirable to have the XETEX expansion available.

How does one convert ~ 20 million formulae?

Therefore, a switch of the production system underlying zbMATH to XETEX/LATEX was in preparation for several years. The initial work-intensive step was made when MathML was introduced in 2010 [1]. Standard tools for MathML conversion employed by zbMATH, such as Tralics [8], require LATEX source. Thus, it was necessary to convert different TEX commands to LATEX – at least, those commands which could be processed by MathML converters. This part of the conversion could be amply addressed by regular expressions. However, it turned out to be necessary to build the full expression tree for TEX formulae. This step, which was finally taken in April of this year, was preceded by an upgrade of PostgreSQL, which allows for UTF-8 handling and a 14-hour routine that converted approximately 18 million standard inline mathematical expressions and 600,000 displaystyle formulae. These included more sophisticated environments, like \alignat or \gathered. Fortunately, only a few environments, e.g., commutative diagrams, require additional manual transition. Overall, the introduction of LATEX resulted in a pause of zbMATH updates of about a month. In contrast to the introduction of TEX, which took three months, this is a significant improvement. The most visible difference for zbMATH users looking at the review sources is the replacement of $…$ by \(…\). Standardisation will allow for much easier integrity checks, and for a much more seamless integration of the submitted reviews. Additionally, considerably better presentation is possible by new functions available via LATEX packages. MathJax (needed for maths presentation in browsers not capable of MathML) works better on widely used LATEX commands instead of less frequent TEX commands. Another advantage pertains to the MathML generation: while Tralics works well for presentation purposes, it hasn't been developed further for some years, so it is reasonable to look for alternative options like LATEXML [6]. The availability of LATEX code makes such alternative implementations much more feasible.

Paving ground for future developments

Even more importantly, further developments in mathematical information retrieval and processing will most likely be based on LATEX. LATEX became the de facto standard in mathematics, and most working mathematicians use LATEX in their publication workflows. Moreover, most websites use LATEX as an input language for mathematical formulae. To make mathematical content better discoverable, multiple approaches exist for enhancing semantics in mathematical formulae. For example, the NIST Digital Library of Mathematical Functions developed a set of semantic LaTeX macros for mathematics. These macros are easy to use for mathematicians fluent with LATEX. By using these macros, with minimal overhead, information retrieval systems would be able to better disambiguate the syntax and semantics for mathematical expressions. Eventually, this will provide better search and recommendation functionality for users of mathematical digital libraries [2]. To some extent, such approaches have already been applied to realise the zbMATH formula search [4, 3], but having standardised LATEX sources at hand will certainly make further developments in this direction much more feasible.

References

[1] P. Baier, O. Teschke, Zentralblatt MATHMLized, *Eur. Math. Soc. Newsl.* 76, 55–57 (2010; Zbl 1278.68088)

[2] H.S. Cohl et al., Growing the digital repository of mathematical formulae with generic LaTeX sources, *Lect. Notes Comput. Sci.* 9150, 280–287 (2015; Zbl 06512424)

[3] F. Müller, O. Teschke, Full text formula search in zbMATH, *Eur. Math. Soc. Newsl.* 102, 51 (2016; Zbl 1366.68355)

[4] M. Kohlhase, H. Mihaljević-Brandt, W. Sperber, O. Teschke, Mathematical formula search, *Eur. Math. Soc. Newsl.* 89, 56–58 (2013; Zbl 1310.68217)

[5] D.E. Knuth, Mathematical typography, *Bull. Am. Math. Soc., New Ser.* 1, 337–372 (1979; Zbl 0404.92025)

[6] LATEXML, a LATEX to XML/HTML/MathML converter, https://dlmf.nist.gov/LaTeXML/

[7] D.C. Latterner, W.B. Wolf, TEX at *Mathematical Reviews*, *Tugboat* 10, No. 4, 639–654 (1989)

[8] Tralics: a LATEX to XML translator, https://www-sop.inria.fr/marelle/tralics/

[9] Letters et alia, *Tugboat* 4, No. 2, 81–102 (1983), https://www.tug.org/TUGboat/tb04-2/tb08letters.pdf

[4] By looking at the sources of these, a reader can easily indicate that they were not genuine TEX-coded; e.g., single variables were not set in formula italics in the old system, so they lack conversion until now.

Moritz Schubotz [moritz.schubotz@fiz-karlsruhe.de] is a senior researcher for mathematical information retrieval and open science. He maintains the support for mathematical formulæ in Wikipedia and is off-site collaborator at NIST.

Olaf Teschke's photo and CV of the author can be found in previous Newsletter issues.