Aesthetics of XML documents with TeX and tagged PDF enabled*

Rishikesan Nair T, Rajagopal CV, Apu V

Abstract

Typesetting and printing of research content in the legacy era was done mainly as a means of disseminating information, with pages playing a central role. Visual clarity and structural precision were essential in presenting content effectively. With the introduction of the Internet and other societal changes, this objective evolved into a broader set of purposes, reshaping how information is shared, accessed, and experienced. This has resulted in the shift in the legacy nature of typesetting of the research content, paying less importance to the visual clarity or the beauty of the pages.

In this article, we will explore the evolution of typesetting practices in the Internet age, focusing on the challenges introduced by the changing dissemination models. We will also examine the XML-first workflow and highlight Primo — a purpose-built tool designed to create visually appealing pages while adapting seamlessly to both the current XML-first workflow and emerging technologies.

1 Introduction

Even before the digital age, maintaining the aesthetic appearance of pages has been always important, as it adds value to the content or ideas being conveyed by making reading easier [1]. The aesthetic appearance is maintained by arranging text and elements in a systematic and consistent order with respect to the content, mainly within the boundaries of the page. The perfect assembling of the font, font size, leading, skips between different elements, alignment of the equations, tables, placement of the floats like figures, tables, footnotes, etc. make a perfect page. Typographers and type designers place great value on the aesthetic quality of typeset content, often investing significant time and care to perfect it. If you've ever admired a beautiful article/book layout, a page layout and thought, "Wow, that just looks right", you've likely been appreciating someone's typesetting mastery.

2 Art of typesetting

Perfect typography is an art — an art through which the ideas and information are conveyed. As we all know, Donald Knuth created TeX in response to a personal and professional disappointment that turned into a legendary act of innovation. TeX wasn't just a software project — it was Knuth's way of preserving typographic excellence in the digital age. And it ended up revolutionizing how technical documents are written and shared. The name TeX is derived from the Greek word technē, which means art, skill, or craft. Donald Knuth chose this name to emphasize that typesetting is not just a technical task — it's a form of artistic expression.

> Good typography, first, makes words readable. At its best, it does something more: it helps express the animating spirit of the ideas behind the words. (Michael Bierut[1])

Although TeX's popularity continues to grow in academia — where it is fast becoming a de facto standard — one cannot fault a devoted TeX user for questioning whether it is still being used in the true spirit of its design: namely, as a medium of artistic expression. This concern extends beyond TeX and typesetting, reflecting a broader trend in academic publishing where the typographical quality of research content appears to be diminishing — often justified by appeals to convenience or efficiency.

In contrast, the aesthetic expectations for other forms of digital media have skyrocketed. Today, we demand polished visual presentation in everything from websites and applications to software GUIs. The discrepancy is hard to ignore: while typesetting leans toward utilitarian minimalism, digital interfaces embrace meticulous design.

Given the reality as discussed here, it becomes essential to examine the challenges, bottlenecks, and underlying reasons for the diminished expectations surrounding the typographical quality of academic research content.

2.1 Print to digital

Research content as disseminated through journals, books, and conference proceedings has long relied on well-established publishing houses that have played a vital role for centuries. The transition from print to digital formats has opened up a broad array of opportunities. Accessibility, plagiarism detection, efficient full-text search, and faster publishing workflows are just a few examples. While embracing the advantages of new technologies, these shifts in workflow — often driven by necessity — have introduced numerous challenges. One notable consequence has been a compromise in typographical quality. Let us explore some of the reasons behind this decline — factors that, in many cases, are neither inevitable nor justifiable.

* This article is dedicated to the typographers and type designers.

[1] `en.wikipedia.org/wiki/Michael_Bierut`

doi.org/10.47397/tb/46-2/tb143rishi-xml-aesthetics

2.1.1 Decentralised publishing models

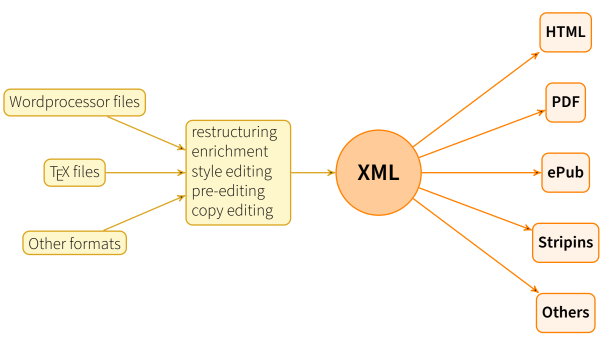
In earlier stages of research publishing, a centralized typesetting model was common. A single typesetter handled the entire production process — from author communication and galley proof preparation to generating print-ready files.

Today, this model has largely been replaced by a decentralized workflow, where responsibilities are split among different individuals or teams. One typesetter may prepare the galley proof, while another handles proof corrections and final production of web- or print-ready PDFs.

This segmentation often leads to a disconnect in formatting continuity. The author's preferences — particularly regarding layout and typographic details — may not be carried through seamlessly across stages. As a result, the final output can diverge from the author's intended presentation, creating friction between contributors and compromising typographical quality.

2.1.2 XML-first workflow

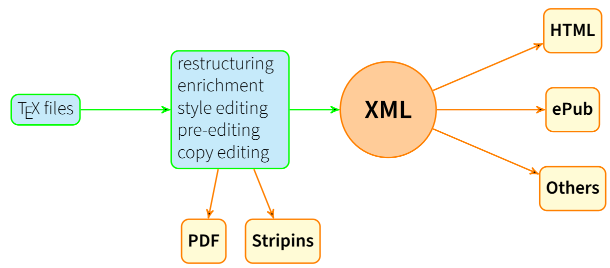
An XML-first workflow demands the production of the PDF files from XML, which is considered as the valid and definitive source.



In an XML-first workflow — where PDFs are automatically generated from XML — we operate primarily through global optimization settings rather than local, document-specific configurations [2]. This means pagination cannot be fine-tuned on a case-by-case basis, unlike traditional typesetting systems where granular control is possible. (See [3] for Frank Mittelbach's experiment in global optimization.)

For instance, actions such as eliminating a loose line, tying strings to prevent poor breaks, or adjusting font sizes for visual balance become difficult or even impossible within this framework. This limitation stems from XML's structure-centric foundation, governed by a DTD or schema that defines the logical organization of content — akin to a database. However, it offers no inherent support for the formatting nuances essential to typesetting systems, which are responsible for producing visually elegant output.

Some journals that accept only LaTeX submissions from authors maintain a strict commitment to visual balance and page elegance. These journals adopt a dedicated LaTeX-based workflow, which ensures the production of beautifully typeset pages while simultaneously leveraging all the advantages of digitisation. In this workflow, the final PDF is generated directly from the LaTeX source rather than from an intermediary XML format, allowing for precise control over layout, typography, and page design.



2.1.3 Unpredictable content

The high volume and vast, ever-changing nature of today's information landscape contribute to the unpredictability of published content. Unlike newspapers and other news-based materials, research content cannot be edited solely for pagination purposes. While typesetters are permitted a limited scope of language editing and the implementation of a publisher's house style, this freedom is minimal. As a result, the typesetter's role becomes particularly demanding and time-intensive. Regardless of the automation being introduced, manual intervention remains essential in most documents to achieve a visually appealing and professionally polished presentation.

2.1.4 Reduced turnaround time (TAT)

In pursuit of faster turnaround times (TAT) and high-level automation to cut production costs, publishers have relaxed traditional pagination rules. These relaxations include more leniency in column balancing, equation alignment, and uniformity of line spacing — just to name a few. Ironically, these very elements are the ones that contribute significantly to the visual elegance of research content. When compromised, the aesthetics and readability of scholarly material often suffer, highlighting the tension between automation-driven efficiency and the craftsmanship inherent in quality typesetting.

2.1.5 XML-based proofing tools

Unlike the legacy period when galley proofs were shared as printed material or PDF documents, the current trend is to use XML-based proofing tools. These tools render the XML version of the content directly in a browser and offer additional features such as track changes, query listings, multiple navigation panels, restricted editing, collaboration, and more. They are also designed to generate PDFs on the fly from the XML which is called a preview PDF, but the typographical quality of these machine-generated preview PDFs is often suboptimal. As previously discussed, the lack of fine control over formatting and pagination plays a major role in this degradation. Nevertheless, authors are typically asked to review and approve these automated galley proofs — even when the pagination and overall appearance are far from visually pleasing.

A major limitation of many XML-based proofing tools, except Primo, is their inability to allow authors to interactively paginate their articles. This frequently leads to tension between authors and typesetters, especially when poor page rendering affects readability and aesthetics. As a result, despite their technological advantages, XML-based proofing platforms are often unpopular among authors — particularly those who use TeX. To address this gap and satisfy the LaTeX authors, publishers adopted LaTeX-based proofing systems, such as Neptune [4], which offer greater visual fidelity, typesetting control and accessible PDFs [6], providing a far more satisfying experience for LaTeX authors. The major advantage of the LaTeX-based proofing tool is that authors can make changes directly in their LaTeX file, compile it, see the output, paginate it and sign off on the author proofing stage with confidence and satisfaction.

3 Primo sets a standard

Primo, a promising XML-based proofing solution, is now available to redefine the author proofing experience [5]. With its combined functionality as both an authoring and a typesetting system, it empowers authors to actively participate in the pagination and aesthetic refinement of their articles. Unlike traditional XML proofing solutions that limit layout control, Primo enables visually balanced output without disarming the author's formatting intentions.

Primo harnesses the power of TeX for PDF generation — a declaration that speaks volumes about its potential in high-quality page design. From creating visually striking pages to generating tagged and accessible PDFs that comply with standards, Primo positions itself as a comprehensive publishing solution. Its blend of typographic finesse and semantic fidelity will make it a game-changer in the author proofing and delivery workflow.

Primo may well be the first XML-based proofing tool to unify authoring, typesetting, and accessible PDF generation within a single, seamless framework. In the absence of comparable tools in the public domain, it's reasonable to consider Primo a trailblazer — setting a benchmark others may one day follow. Whether similar platforms are quietly emerging or yet to be imagined, Primo's journey from concept to pilot has been long, deliberate, and worthy of its ambition. Its continued evolution promises to shape the future of author engagement and publishing precision.

3.1 Features of Primo

Here are some of the features of Primo:

1. A tool from a house of TeX people, introducing the beauty of TeX typesetting to the non-TeX community.
2. Accessible PDF on the fly.
3. Eliminates the need for intervention by other tools to make the underlying XML valid.
4. Supports collaborative editing.
5. Keeps the advantages of XML-based workflows that facilitate controlled authoring and/or editing in accordance with the specific DTDs.
6. Journal template-based PDF generation using TeX.
7. A well-designed proofing environment with all relevant features to help easy proofing.

4 Conclusion

Maintaining the aesthetic integrity of research documents enhances readability and preserves the art of typesetting. Among XML-based tools, Primo stands out for redefining design standards in modern publishing.

Acknowledgement

We gratefully acknowledge the valuable suggestions and comments provided by Radhakrishnan CV and Rahul Krishnan S.

References

[1] archive.org/details/artpracticeoftyp00gres
[2] journal-bcs.springeropen.com/articles/10.1007/s13173-012-0066-6
[3] tug.org/tug2016/abstracts/mittelbach.pdf
[4] neptune.texfolio.org
[5] tug.org/TUGboat/tb44-2/tb137rishi-primo.pdf
[6] tug.org/TUGboat/tb46-2/tb143rahul-neptune-tagged.pdf

◇ Rishikesan Nair T, Rajagopal CV, Apu V
STM Document Engineering Pvt Ltd.,
Trivandrum 695571, Kerala
India
rishi (at) stmdocs dot in,
cvr3 (at) stmdocs dot in,
apu.v (at) stmdocs dot in
https://stmdocs.in