Advancing PDF in Scholarly Publications

Peter Wyatt

Introduction

In the late 1990s, PDF became the digital file format of choice for scientific and technical publishers. Thanks to precise and exact typesetting of paginated content, device-independent color and vector graphics, and guaranteed results both on-screen and in print, even the first generation of PDF documents from 1993 are still fully functional today.

Unbeknownst to many, the PDF format has undergone enormous change in its 30+ years. In 2007, Adobe surrendered control over the format to an ISO committee, which has since defined PDF as ISO 32000, an open international standard developed under consensus-based processes. As a result, support for PDF is ubiquitous, with creators, viewers, and other PDF software an integral part of all browsers, platforms, and devices.

Common perceptions of PDF, however, have not significantly changed since the early 2000s, when both the file format and the most popular viewer were controlled by a single organization. Today, 17 years after PDF became an international standard, various misconceptions remain commonplace:

- **"Adobe PDF"** Though a common reference, this is a misnomer. PDF became an open international standard in 2008 and is supported today by thousands of vendors providing users with many alternatives.
- **"The notion that PDF content is never searchable or extractable."** In the early days of PDF, file size, font licensing, the complexity of digital font technologies, and limited PDF software impacted access to PDF's text content. Modern PDF applications that directly export PDF will always embed necessary font data as is required by the latest PDF 2.0 standard and all ISO-standardized subsets of PDF.

Peter Wyatt (https://orcid.org/0009-0007-1282-9675) is CTO, PDF Association.

Opinions expressed are those of the authors and do not necessarily reflect the opinions or policies of their employers, the Council of Science Editors, or the Editorial Board of Science Editor.

https://doi.org/10.36591/SE-4801-11

- **"PDF is inaccessible to those who need assistive technologies (AT) in order to read and navigate documents."** Tagged PDF,[1] the feature that enables accessible PDF,[2] was added to (then) Adobe PDF in 2001. The first ISO standard defining the correct use of PDF for universal accessibility—ISO 14289, PDF/UA—was first published in 2012. Unfortunately, some authoring applications still do not fully support Tagged PDF when exporting to PDF; hindering the creation of truly accessible PDFs.
- **"PDF has remained unchanged since the 1990s."** This fallacy was reinforced by various authoring applications that, until relatively recently, only saved PDF files using legacy PDF versions with severely reduced feature sets. A global focus on accessibility, prompted by laws and regulations (e.g., the European Accessibility Act [EAA][3]) requiring accessible content, triggered developers to update their office applications so authors could export richer and more accessible documents across HTML, EPUB, and PDF.
- **"Offline paginated content is outmoded."** Standalone, single-file, paginated content remains relevant to scholarly publishers and their end users, including professors, students, librarians, researchers, and other academics, as continued demand for both PDF and EPUB demonstrate.

PDF 2.0

ISO 32000-1, published in 2008, represented an ISO-standardized version of Adobe's PDF 1.7[4] specification. Nine years later, in 2017, the first consensus-based, vendor-neutral open standard for PDF was published as PDF 2.0[5] (ISO 32000-2). While maintaining backward compatibility with past versions of PDF, PDF 2.0 introduced several new file format features and requirements relevant to STEM publishers.

- All font data is now required for every PDF file. Legacy versions of PDF allowed a dependency on external fonts, which led to varying appearances and difficulties in extracting text.
- Support for the latest Unicode standard, ensuring that content in any language can be reliably represented for extraction and reuse.

CONTINUED

- The addition of MathML 3.0 as a "first-class citizen" in PDF's Logical Structure feature enables full accessibility for complex mathematical typography (Figure 1).
- An updated set of semantic "tags" to improve accessibility and reuse of a wide range of content.
- A new Associated File feature, wherein embedded files of any format can be associated with any PDF object along with a semantic relationship to that content, such as the original source data (e.g., a CSV for a chart), an alternate representation, a data schema, etc.
- A vendor-neutral portable collection feature, enabling single-download distribution of multiple files (of any format) in a single PDF package. If PDF documents are contained in the collection, they may also reference (hyperlink) each other.
- Interactive 3D content can be supported via multiple 3D formats (U3D, PRC, glTF, and STEP AP242) for use in medical, engineering, and other disciplines (Figure 2).
- The addition of geospatial coordinate measurement features used in cartographic and related applications.
- Updated digital signature technology, capable of providing authenticity guarantees.
- Support for the latest, modern encryption algorithms for secured content.
- An "unencrypted wrapper" feature enabling proprietary digital rights management (DRM) with a controllable publisher-defined experience.
- XMP-based metadata is now the preferred metadata format, enabling a far richer metadata vocabulary and easier discovery.

The PDF Association[7] continues to develop new PDF specifications, extensions, guidelines, and test suites to maximize interoperability and ensure a consistent understanding of PDF standards. For example, as of this writing, and at the request of stakeholders in the publishing industry, the organization is developing a specification for the inclusion of ONIX[8] payloads in PDF files. The PDF Association is also working with office application suite developers to enable the export of semantically rich documents to modern PDF.

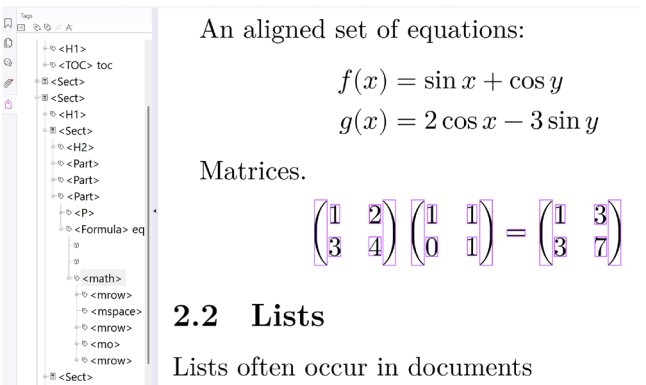

Figure 1. This example PDF file (https://pdfa.org/download-area/examples/MathML-in-PDF.pdf) was generated by the LaTeX Project's WTPDF generator at https://latex3.github.io/tagging-project/documentation/wtpdf-from-latex, using default (as of February 12, 2025) settings.

Leveraging ISO Standards for PDF in Publishing

Unlike the transient nature of the web with content and URLs that come and go, PDFs are fully self-contained documents that can persist indefinitely. Once a PDF document is added to a library, whether it be an institution or a personal library, that *exact PDF* remains available and usable forever under the librarian's sole control.

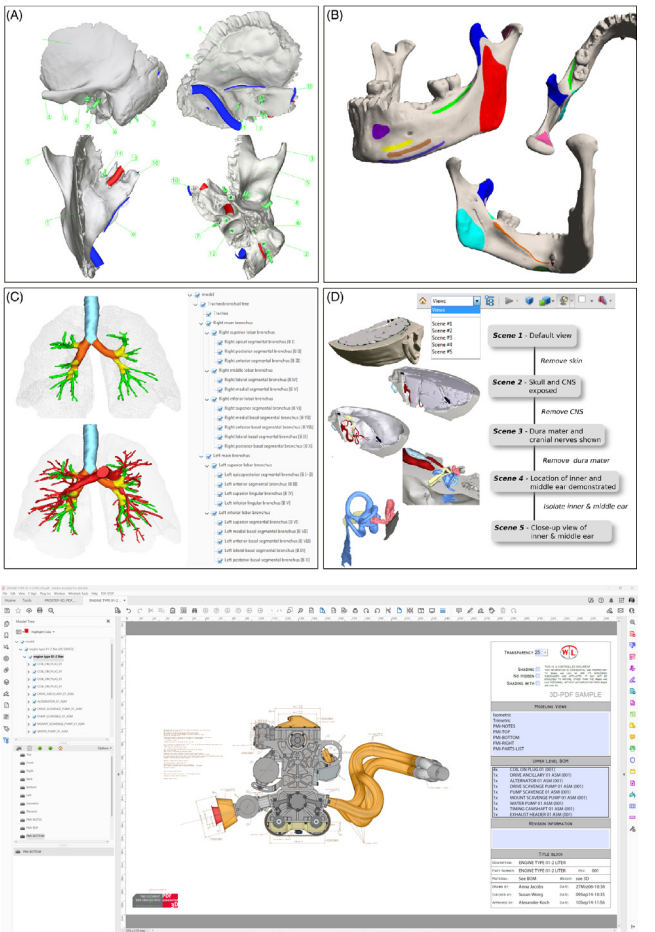

Figure 2. (top) Image reprinted from Azkue.[6] (bottom) Screenshot from Adobe Acrobat that is supposed to represent "engineering content" (https://pdfa.org/3d-pdf-showcase/#technical; https://pdfa.org/wp-content/uploads/2021/12/PROSTEP-3D_PDF_TDP.pdf).

CONTINUED

For formal long-term preservation needs, such as those required by national archives, museums, and some libraries, a specialized profile known as PDF/A[9] ("A" means "archival") was first formalized as ISO 19005 in 2005. This profile is designed so that PDF files that declare conformance to PDF/A can be easily machine-validated (i.e., checked by software) upon submission or ingestion.

To support production of physical products (i.e., books, journals, and other printed materials), ISO 15930 (better known as PDF/X[10] ["X" means "eXchange"]) was first standardized in 2001 for the graphics arts and commercial printing industries. PDF/X supports the blind exchange of PDF content (meaning only the PDF file is needed) with all data necessary to ensure exact and reliable printed output across disparate print providers. This permits geographically dispersed printing, reducing distribution and mailing costs while ensuring that all printed copies are identical.

The principles of web accessibility as defined in the Web Content Accessibility Guidelines (WCAG)[11] are applicable to web content, but WCAG's guidance, being oriented towards web content, does not fully address PDF. To ensure that PDF documents are readable and navigable by users who must use AT, the PDF technology community developed PDF/UA[12,13] ("UA" means "Universal Accessibility"), first formalized as ISO 14289 in 2012. Authoring applications now include accessibility checking to support WCAG, and as a result, can produce Tagged PDF as well as accessible HTML and EPUB.

PDF's evolution from a free (but proprietary) specification to an open international standard continues to add new features and vendor-neutral capabilities, with more to come. Each significant generation of PDF is matched with updated PDF/A, PDF/X, and PDF/UA standards to ensure ongoing success for the industries that depend on these documents. Today, many other industries have built on top of these widely adopted standards to create specialized applications that leverage the general availability software that creates, validates, views, or otherwise uses PDF.

Author Guidance

As capable as modern PDF is, any given PDF document can only be as good as the way in which it was created. For example, PDF can be entirely accessible with fully extractable and reusable rich content, but only when both the document's author and their creation software prioritize the steps necessary to produce such content. This rich content (for the eye and the printed page) can also be of archival quality as preferred by libraries. Thus, publishers are responsible for ensuring that capable authoring applications are chosen and correctly used to ensure that the author's intended semantics are appropriately captured and can be exported to PDF, HTML, or EPUB. Simply formatting text to look like a heading will not make it a heading, using a dash or asterisk does not make something a list, etc. Thankfully, all modern office application suites now use style sheets with accessibility suggestions supported by artificial intelligence. These applications can generate a Tagged PDF and even PDF/UA-compliant documents, while the latest updates to LaTeX[14] enable PDF/A and PDF/UA generation from STEM content.

In light of the ongoing evolution of PDF, publishers should update their workflows and author guidance in several ways:

- Ensure all authors' application templates are updated and include the necessary accessibility features and clear instructions on how to export to PDF. These application features are critical for ensuring that exported content, whether HTML or PDF, can be accessible. Instructions are important because PDFs created via print pipelines, although identical in appearance, will not contain rich features or semantics.
- Avoid legacy PDF versions by requiring PDF 1.7 and PDF 2.0 as this helps to ensure the use of up-to-date software and provides the best chance of receiving high quality semantically rich content at the smallest file size.
- Accept and publish all PDF publications as Tagged PDFs, ideally as PDF/UA (ISO 14289) compliant, to meet EAA, Section 508,[15] and other regulations that support users who need assistive technology.
- For publications that include mathematics, ensure PDF 2.0 and MathML are used.
- Only accept PDF publications that include all fonts and related Unicode data. Complying with either PDF/A, PDF/UA, or PDF/X guarantees this is always achieved. Out-of-date authoring software with legacy PDF versions or creating PDFs via printing pipelines cause such issues.
- Ensure authors understand and use predefined styles wherever possible, and limit the use of inline styling, as manually applied inline styling cannot convey the same necessary semantics.
- Provide authors with PDF validation tools and training so they can check their documents prior to submission.
- Encourage the use of PDF 2.0 or PDF/A files with associated embedded files for publications supporting open data with reasonably-sized data sets. These data files can be semantically associated with specific PDF content, such as a chart or image. PDF also supports efficient data compression.
- Accept and publish PDF documents that include interactive 3D and geospatial content, as these are standardized PDF features.
- Accept and publish PDF documents with accurate document XMP metadata.

CONTINUED

Table. PDF as defined by ISO standards.

Technology	Nomenclature	PDF 1.7	PDF 2.0
Core PDF specification	PDF	ISO 32000-1:2008	ISO 32000-2:2020
PDF for archiving	PDF/A	ISO 19005-3:2012	ISO 19005-4:2020
PDF for universal accessibility	PDF/UA	ISO 14289-1:2014	ISO 14289-2:2024
PDF for graphic arts/printing	PDF/X	ISO 15930-8:2010	ISO 15930-9:2020

- Always refer to the PDF file format in a vendor-neutral manner. PDF is best referred to as "Portable Document Format" or simply "PDF." If technical precision is important, reference ISO 32000. For specialized applications, other PDF nomenclature and related ISO standards might also be used, such as PDF/UA or ISO 14289 for accessibility, PDF/A or ISO 19005 for long-term preservation, or PDF/X or ISO 15930 for print publications.

These recommendations assume that publishers are themselves using modern, up-to-date PDF software in their workflows. This entails—at a minimum—full support for PDF 1.7 based on ISO 32000-1 and preferably, PDF 2.0 because the occurrence of PDF 2.0 is increasing with more and more technical authoring applications recognizing the clear benefits of new features such as those listed above. Publishers that fail to support PDF 2.0 for technical and scholarly publications in the near future face reputational risk, increased costs to their business, and potential regulatory risk.

Conclusion

PDF is a living, thriving file format developed and actively maintained in the PDF Association, a consensus-based, vendor-neutral standards organization, and formally standardized via ISO. The principal ISO standards defining PDF are listed in the Table.

By supporting the rich feature set defined in modern PDF specifications such as PDF 2.0, publishers can ensure that all readers have an optimal experience with rich content. By further leveraging existing ISO standards such as PDF/UA, PDF/A, and PDF/X, publishers can reduce their costs while meeting regulatory and policy requirements.

Perhaps the most difficult challenges lie in convincing (and helping!) authors to competently use up-to-date, capable application software that will then export best-in-class PDF documents with modern features.

References and Links

1. https://pdfa.org/wtpdf/
2. https://pdfa.org/accessibility
3. https://en.wikipedia.org/wiki/European_Accessibility_Act
4. https://pdfa.org/resource/iso-32000-1/
5. https://www.pdfa.org/resource/iso-32000-2-pdf-2-0/
6. Azkue JJ. Embedding interactive, three-dimensional content in portable document format to deliver gross anatomy information and knowledge. Clin Anat. 2021;34:919–933. https://doi.org/10.1002/ca.23755.
7. https://pdfa.org
8. https://www.editeur.org/8/ONIX/
9. https://pdfa.org/archival-pdf/
10. https://pdfa.org/resource/iso-15930-pdfx/
11. https://www.w3.org/WAI/standards-guidelines/wcag/
12. https://pdfa.org/iso-14289-2-pdfua-2/
13. https://pdfa.org/resource/iso-14289-pdfua/
14. https://latex3.github.io/tagging-project/
15. https://www.section508.gov
16. https://www.iso.org/committee/53674.htmlmmittee/53674.html

CONTINUED

knowledge sharing that is adapted to different linguistic and cultural situations, to ensure that there is no regionalisation and fragmentation of the knowledge-sharing infrastructure. For Crossref, a large part of this is listening to the needs of our current, highly diverse membership, as well as reaching out to those who are not yet fully part of our community.

Challenges in Getting There

Cultural change is hard, and in order for the vision of the research nexus to come to fruition, we have to work on open data and open research becoming the default and change incentive structures for how research is assessed. Publishing an article in a high Impact Factor journal is not sufficient. Another challenge is financial—research and scholarly publishing require significant resources, as does the creation and maintenance of high-quality metadata. This all needs support from, and collaboration between, government, funding bodies, research institutions, researchers, open infrastructure providers, scholarly societies, and commercial companies.

A key underpinning for our vision of the future are the Principles of Open Scholarly Infrastructure (POSI).[9] These are 16 principles covering open data, sustainability, and inclusive governance that are essential for metadata and will continue to be as relevant in 2050 as they are now.

Supporting and embracing technological innovation in a measured way and being globally inclusive are also very important. More work is needed to expand the scholarly record to more fully include the Global South and expand the scholarly record to cover areas such as grey literature and Indigenous Knowledge.

All the elements are in place for ensuring that in 2050, we will have overcome the current challenges so that metadata supports a fully open and dynamic global research ecosystem.

References and Links

1. https://orcid.org/
2. https://ror.org/
3. Hendricks G, Tkaczyk D, Lin J, Feeney P. Crossref: the sustainable source of community-owned scholarly metadata. Quant Sci Stud. 2020;1:414–427. https://doi.org/10.1162/qss_a_00022.
4. Rittman M. Retraction Watch retractions now in the Crossref API. Crossref Blog. January 29, 2025. https://doi.org/10.13003/692016.
5. Tkaczyk D, Buttrick A. Metadata matching 101: what is it and why do we need it? Crossref Blog. May 16, 2024. https://doi.org/10.13003/aewi1cai.
6. https://www.dublincore.org/
7. Atkins H, Lyons C, Ratner H, Risher C, Shillum C, Sidman D, Stevens A. Reference linking with DOI: a case study. D-Lib Magazine 2000;6. https://doi.org/10.1045/february2000-risher.
8. Mugaanyi J, Cai L, Cheng S, Lu C, Huang J. Evaluation of large language model performance and reliability for citations and references in scholarly writing: cross-disciplinary study. J Med Internet Res 2024;26:e52935. https://doi.org/10.2196/52935.
9. Bilder G, Lin J, Neylon C. The principles of open scholarly infrastructure. 2020. https://doi.org/10.24343/C34W2H.

(Continued from p. 11)

on how to adopt them responsibly. A risk management framework tailored to the diverse applications of AI tools is the first step in this process.

By developing clear risk profiles, approving vetted tools, differentiating between substantive and nonsubstantive uses, and implementing reliability scoring systems, publishers can navigate the complexities of AI adoption with confidence. Equally important is the commitment to education and training, ensuring that every stakeholder in the publishing ecosystem understands both the opportunities and the risks of AI.

The future of scientific publishing lies not in avoiding AI but in embracing it thoughtfully, with robust safeguards in place. The responsibility now falls on publishers, editors, and researchers to collaborate in building a publishing environment where AI serves as a tool for progress, integrity, and innovation.

Disclosure

I uploaded lectures and slides I created on my own to ChatGPT for a first draft of this article. I then reviewed, revised, and edited it before sharing with ChatGPT for feedback. After implementing some changes I agreed with, I then uploaded once more to ChatGPT for an edit/proofread. The responsibility for the content in this article is mine entirely.

References and Links

1. https://artificialintelligenceact.eu/high-level-summary/
2. https://european-research-area.ec.europa.eu/news/living-guidelines-responsible-use-generative-ai-research-published
3. https://www.digital-science.com/tldr/article/dark-matter-whats-missing-from-publishers-policies-on-ai-generative-writing/
4. https://www.chronicle.com/article/scholars-are-supposed-to-say-when-they-use-ai-do-they
5. https://sr.ithaka.org/our-work/generative-ai-product-tracker/
6. https://pubs.acs.org/doi/10.1021/acsnano.3c01544