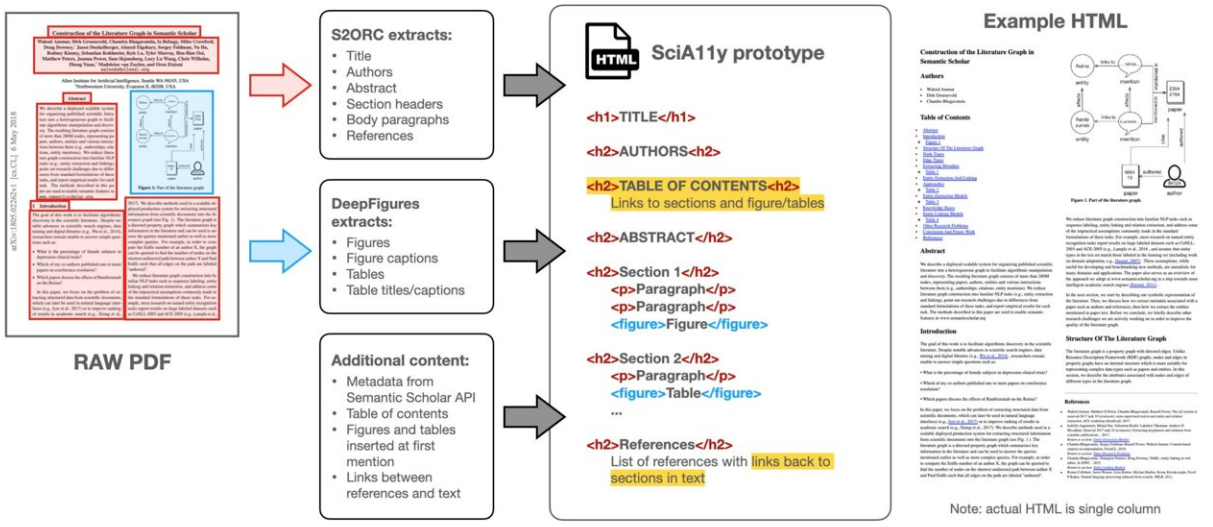
New tool will make math-heavy research papers easier to view online

January 3 2024



A schematic for creating the SciA11y HTML render from a paper PDF. Starting with the raw two-column PDF on the left, S2ORC [24] is used to extract the title, authors, abstract, section headers, body text, and references. S2ORC also identifies links between inline citations and references to figures and table objects. DeepFigures [43] is used to extract figures and tables, along with their captions. The output of these two models is merged with metadata from the Semantic Scholar API. Heuristics are used to construct a table of contents, insert figures and tables in the appropriate places in the text, and repair broken URLs. We add HTML headers as illustrated (header tags for sections, paragraph tags for body text, and figure tags for figures and tables); highlighted components (table of contents and links in references) are not in the PDF and novel navigational features that we introduce to the HTML render. An example HTML render of parts of a paper document is shown to the right (the actual render is a single column, which is split here for presentation). Credit:

The complex formulas in physics, math and engineering papers might be intimidatingly difficult reading matter for some, but there are many people who have trouble merely seeing them in the first place. The National Institute of Standards and Technology (NIST) has created a tool that makes these papers easier on the eyes for those with visual disabilities, and it's about to be adopted in a major way.

The tool, which converts one commonly used format for displaying math formulas into another, could help make the latest and greatest research papers accessible to all. Most new research papers are distributed as PDF files, which many people in the research community have difficulty reading.

According to the World Health Organization, more than a quarter of the world's population has a diagnosed vision impairment, and Yale's Center for Dyslexia and Creativity reports that in the United States 20% of people have dyslexia. In a recent study of scientific papers distributed as PDFs, researchers found that only 2.4% of the documents they sampled satisfied their accessibility criteria.

"If you're not someone who has been struggling to publish math papers all your life, you might wonder why this is a problem," said NIST's Bruce Miller, a physicist by training who specializes in math software. "PDFs look great on the printed page. But if you want math formulas to be read out loud, or be legible on a different-sized screen, like a tablet or a phone, the mismatch can be painful. You can't easily repurpose PDFs for other media."

How are PDFs typically generated? A scientist creating a paper

manuscript that uses many formulas will generally use the language LaTeX (pronounced "lay-tech") or one of its close relatives to render the formulas. LaTeX has been in use since the 1980s and is widely respected for the high-quality typesetting that it creates, but it is designed to produce printed pages in static form.

Since the 1990s, webpage creators have used HTML, which makes it possible to adjust the look, behavior and layout of the displayed text depending on its context. If you've ever dragged a webpage into a different size and watched its text smoothly reposition itself to fit within the new rectangle's boundaries, you are seeing a feature that readers with vision disabilities want.

Modern HTML includes extensions that not only permit this ability to "re-flow" type, but also allow the math formulas to be read aloud by machine for those who can't read the text themselves. These features make HTML ideal for creating accessible text, but for years there was no effective way to convert LaTeX into HTML. This presented a problem to Miller when he needed a way to bring the more than 1,000 pages of NIST's venerable Handbook of Mathematical Functions into the digital realm.

"At the time, some programs purported to convert LaTeX to webpages, but none worked well enough," he said. "I figured, let's try to make our own."

The resulting NIST tool was LaTeXML, which reads a LaTeX source file and builds a representation of the document that it can turn into HTML. LaTeXML was the key to creating the online Digital Library of Mathematical Functions, and several years later the managers of a major online resource realized it could help them too.

This resource is *arXiv* (pronounced "archive"), a repository of scholarly

articles that have yet to be published in scientific journals. Maintained by Cornell University, *arXiv* currently hosts more than 2 million articles that are free to view and download as PDFs. The server has become a prominent way station, where authors can post findings and discuss them with their peers before formally announcing them.

"Per a survey *arXiv* conducted in 2022, only 30% of users who rely on assistive technology can access all of the research they need without help. The same survey found that PDF formatting is the biggest barrier," said Shamsi Brinn, lead researcher on *arXiv*'s accessibility report and manager of the HTML papers project.

That will change with *arXiv*'s use of the LaTeXML converter, Brinn said. The server will generate HTML versions of papers and include the HTML version next to the link to download a PDF.

The *arXiv* repository will convert papers on a rolling basis, offering the first in December 2023. The move follows a broader trend of requiring accessible web and electronic information, according to Joe Zesski, assistant director of the Northeast ADA Center. Not only will the change help the scientific community adhere to the White House's updated policy on making federally-funded research freely available, but it will also make information accessible to young scientists, who have grown up using electronic resources.

"There is a growing reliance on the web and electronic information in education alongside a growing expectation of equal access by and for young people with disabilities," Zesski said. "Taking steps to make the information those students will need to access accessible and usable to them is important."

Provided by National Institute of Standards and Technology