Reading 29,000 COVID-19 papers

Jonathan Fine

On 17 March 2020, *The Register* wrote [5]:

> A dataset of more than 29,000 scientific papers focused on COVID-19, and the coronavirus family as a whole, has been publicly shared to ultimately help the medical world thwart the bio-nasties.
>
> Specifically, it is hoped AI-based tools can be developed to comb through this COVID-19 Open Research Dataset (CORD-19) and dig up vital clues and insights on how to treat and contain the virus.

This article considers how TEX, or more exactly LATEX, fits into this massive and important research effort. We read in [1] that each paper is represented as a single JSON object, which conforms to a schema [2].

And now we see the LATEX problem. Even though the source document for the PDF is an informally structured document, it does not conform to a standard. It does not have a schema. It is sure to be machine readable by only one software system, namely LATEX, along with the specified preamble.

Similarly, the resulting PDF is machine readable only by a PDF reader, and in general the only machine readable semantic information it contains are the links. In particular, screen readers often fail to work well with LATEX-produced PDF files.

The US National Library of Medicine provides a Journal Article Tag Suite (JATS) [4] which "is a continuation of the NLM Archiving and Interchange DTD work begun in 2002" by the (US) National Center for Biotechnology Information.

My preliminary web search shows little activity in the area of JATS and LATEX, even though JATS is an important schema for scholarly articles. The Scholastica journal management platform [6] provides a typesetting service (probably based on LATEX) that will "generate HTML, PDF, and full-text JATS XML versions of articles".

Scholastica "was founded in 2012 in response to a growing need in academia for an easier, more modern way to peer review research articles and publish high-quality open access journals online" by three people who met when they were graduate students at the University of Chicago [7].

Fields medallist Tim Gowers is a key Editor [3] of the journals *Discrete Analysis* and *Advances in Combinatorics*, which are arXiv overlay journals, published on the Scholastica platform.

We now return to the dataset of 29,000 scientific papers focused on COVID-19. Clearly, the primary value of these papers is the experience, training, skill and dedication of the authors of these papers. Any system that allows for discovery, analysis, extracting and other use of these papers adds further value.

This is the theme of the article in *The Register*. Can Artificial Intelligence help humanity make sense of and use these 29,000 articles? I hope the answer is Yes, because any contribution helps. We can also ask: Has LATEX similarly helped humanity?

For us as TEX and LATEX users (and developers) more important than a Yes or No is this question: What can we do now, and in the future, to make TEX more useful for scientific and human challenges such as COVID-19?

References

[1] `https://pages.semanticscholar.org/coronavirus-research`

[2] `https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/2020-03-13/json_schema.txt`

[3] `https://gowers.wordpress.com/2019/10/30/advances-in-combinatorics-fully-launched/`

[4] `https://jats.nlm.nih.gov/`

[5] `https://www.theregister.co.uk/2020/03/17/ai_covid_19/`

[6] `https://scholasticahq.com/`

[7] `https://en.wikipedia.org/wiki/Scholastica_(company)`

◇ Jonathan Fine
jfine2358@gmail.com
https://jfine2358.github.io