Subject: Getting tables out of DHS Final Report PDFs: converting to Markdown first
Posted by Pdf2mdTools on Tue, 08 Sep 2026 04:50:16 GMT
View Forum Message <> Reply to Message

Most of what I work with from The DHS Program arrives as a Final Report PDF, and the tables in those reports are where the numbers I actually need live. Retyping a table of contraceptive prevalence by region into a spreadsheet is slow and it is where I have introduced the most mistakes.

What has worked well for me is converting the report to Markdown first, then pulling the tables out of the text. Markdown keeps the heading structure of the report, so chapter and table numbering survive, and a pipe table can be pasted straight into R or a spreadsheet.

The converter I have been using runs entirely in the browser, so the report never gets uploaded to a server, which matters when a colleague is working with a restricted dataset: https://pdf2md.me

A couple of practical notes for anyone trying the same thing:

- Convert the whole report in one pass rather than page by page; the heading levels come out consistent that way.
- Scanned reports (older survey rounds) still need OCR first, this only helps with text based PDFs.
- Always check the converted table against the PDF for the footnote markers, those are the easiest thing to lose.

Curious whether anyone else here has a workflow for getting the Final Report tables into an analysis tool, or whether most people go straight to the STATcompiler / microdata instead.