Supplementary

A. Existing DC datasets

As the datasets from Table 2 did not satisfy large-scale benchmarking multi-page DC benchmarking requirements, we discuss them in supplementary for interested readers.
Tobacco-3482 [23] is another subset of IIT-CDIP with fewer samples and a smaller label set than RVL-CDIP.
Tobacco-800 [62] has been used for page stream segmentation ([56], similarly defined as in [36]) as it contains consecutively numbered multi-page business documents.
NIST The NIST Structured Forms Database [8] consists of 5,590 binary synthesized documents from 20 different classes of tax forms.
MARG The MARG (Medical Article Records Groundtruth) database [32] is a layout-based classification benchmark containing 1553 documents which are mainly the first pages of medical journals.
TAB [36] is a recently introduced page stream segmentation dataset targeting binary classification to detect document boundaries on multi-page streams. It consists of a sample of 44,769 PDF documents from the Truth Tobacco Industry Documents (TTID) archives.

B. Visualization of proposed DC datasets

As we have contributed two novel datasets consisting of multi-page documents in PDF format, adding visualizations is non-trivial. The datasets are hosted at the HuggingFace Hub (https://huggingface.co/datasets/bdpc), for which at the time of submission, the dataset viewer does not support PDF data. Rather than adding examples in the manuscript, which is tedious for PDF documents with multiple pages, we have built an interactive app (https://huggingface.co/spaces/jordyvl/viz_bdpc). This allows for the visualization of samples from the proposed datasets, with an additional filter on the labels, whereas both datasets follow the original RVL-CDIP label taxonomy.