Introduction

Paged documents scatter their words across many small boxes, and a careful extractor has to put them back in reading order. This first paragraph is deliberately long so that it spans several lines of the page and so that the justification of the lines becomes visible to anybody who looks at the result with a critical eye. A few more sentences follow here only to fill up the first page, so that a later paragraph is forced to continue on the next page. The quick brown dog sleeps under the old oak tree while the children play in the garden and the bees buzz around the flowers.

The second paragraph is indented like every other paragraph that follows another one, and the extractor should notice that indentation and start a new paragraph here. Its reconstruction depends on the geometry alone, since the file itself does not record where one paragraph ends and the next one begins.

Short paragraph number three.

The fourth paragraph mentions a few typographic details: the office has fluffy waffles, an efficient staff and a “quoted phrase” with pages 10–20 and a pause—like this one. The café serves naïve guests.

The fifth paragraph runs across the page break, right past the running header and the page number, neither of which must appear in the extracted text. We continue writing some more text here so that this paragraph covers enough lines to overflow from the first page onto the second page, where it simply goes on as if nothing had happened, and the extractor must join both halves into one single paragraph without inserting any break at all. To make sure that this really happens, the paragraph goes on for a while longer with a few more sentences of plain filler text that nobody will ever read,

apart from the test suite, which checks every single word of it.

The sixth paragraph is the last one of the single column part and it ends the page with a sentence about zebras jumping over lazy foxes.

The seventh paragraph opens the two column part of the document. The left column is read completely before the right column, and the line numbers printed in the margins next to each column must be dropped by the extractor because they are not part of the text at all.

The eighth paragraph continues in the left column with another few lines of harmless filler text about apples, pears and plums that grow in an orchard near the river.

The ninth paragraph keeps going with filler about the weather, which was mild and pleasant in the morning but turned rather cold later on. The ninth paragraph keeps going with filler about the weather, which was mild and pleasant in the morning but turned rather cold later on.

The tenth paragraph is the first one of the right column and it talks about lighthouses that guide ships safely through the night along rocky coasts.

The eleventh paragraph concludes the document with a final remark.