Reading water-quality reports at scale

How column geometry turns a messy PDF into clean text

Municipal water systems publish quality reports every year, but the numbers rarely travel far beyond the PDF. Analysts who want to compare lead levels across towns still copy values by hand, one table at a time. A small script can do better: open each report, find the sampling table, and write the results to a spreadsheet. The hard part is not the arithmetic. It is the layout. Reports mix narrative text in two columns with tables, footnotes and headers that repeat on every page. Text extraction that ignores the columns reads straight across the gutter and produces sentences that never existed. The fix starts with geometry. Every character on a PDF page has a position, and the gap between two columns is simply a vertical band where no characters fall. Find that band, split the page, and each half reads cleanly. Headers and footers are handled the same way: they sit in predictable strips at the top and bottom of the page, so cropping them away removes the noise before any text is joined together. None of this needs machine learning. It needs a library that exposes coordinates and lets you filter by them. That is exactly what a geometry-first extraction tool is for, and why it remains popular for born-digital documents even as newer model-based parsers appear. For scanned pages the story changes, because there is no text layer to measure; an OCR step has to create one first. Alder Creek shared its 2014 report as a 38-page PDF. It lists 23 sampling sites and 5 readings above the action level, set out in a table on page 8. The table has 5 columns and is ruled only between rows, so the column positions have to come from the header. Note 1 in our log records that the Alder Creek file mixes two fonts in the header and prints the header once. Bramford filed its 2015 report as a 41-page PDF. It lists 84 sampling sites and 9 readings above the action level, set out in a table on page 38. The table has 8 columns and is ruled only between rows, so the column positions have to come from the header. Note 2 in our log records that the Bramford file uses one font throughout and repeats the header on every page. Cedar Point posted its 2016 report as a 34-page PDF. It lists 70 sampling sites and 9 readings above the action level, set out in a table on page 2. The table has 8 columns and is ruled only between rows, so the column positions have to come from the header. Note 3 in our log records that the Cedar Point file uses one font throughout and prints the header once. Dunmore published its 2017 report as a 24-page PDF. It lists 18 sampling sites and 1 readings above the action level, set out in a table on page 16. The table has 4 columns and is ruled only between rows, so the column positions have to come from the header. Note 4 in our log records that the Dunmore file mixes two fonts in the header and repeats the header on every page. Elmhurst filed its 2018 report as a 19-page PDF. It lists 70 sampling sites and 5 readings above the action level, set out in a table on page 10. The table has 6 columns and is ruled only between rows, so the column positions have to come from the header. Note 5 in our log records that the Elmhurst file uses one font throughout and prints the header once. Fairhaven published its 2019 report as a 38-page PDF. It lists 53 sampling sites and 6 readings above the action level, set out in a table on page 23. The table has 4 columns and is ruled only between rows, so the column positions have to come from the header. Note 6 in our log records that the Fairhaven file uses one font throughout and repeats the header on every page. Glenrock released its 2020 report as a 24-page PDF. It lists 79 sampling sites and 8 readings above the action level, set out in a table on page 16. The table has 5 columns and is ruled only between rows, so the column positions have to come from the header. Note 7 in our log records that the Glenrock file mixes two fonts in the header and prints the header once. Harlow published its 2021 report as a 25-page PDF. It lists 90 sampling sites and 9 readings above the action level, set out in a table on page 24. The table has 6 columns and is ruled only between rows, so the column positions have to come from the header. Note 8 in our log records that the Harlow file uses one font throughout and repeats the header on every page. Ironbridge published its 2022 report as a 33-page PDF. It lists 54 sampling sites and 7 readings above the action level, set out in a table on page 27. The table has

8 columns and is ruled only between rows, so the column positions have to come from the header. Note 9 in our log records that the Ironbridge file uses one font throughout and prints the header once. Juniper Falls filed its 2023 report as a 13-page PDF. It lists 78 sampling sites and 7 readings above the action level, set out in a table on page 10. The table has 8 columns and is ruled only between rows, so the column positions have to come from the header. Note 10 in our log records that the Juniper Falls file mixes two fonts in the header and repeats the header on every page. Kingsbury released its 2024 report as a 39-page PDF. It lists 70 sampling sites and 6 readings above the action level, set out in a table on page 28. The table has 8 columns and is ruled only between rows, so the column positions have to come from the header. Note 11 in our log records that the Kingsbury file uses one font throughout and prints the header once. Lakeview published its 2025 report as a 39-page PDF. It lists 69 sampling sites and 8 readings above the action level, set out in a table on page 38. The table has 5 columns and is ruled only between rows, so the column positions have to come from the header. Note 12 in our log records that the Lakeview file uses one font throughout and repeats the header on every page. Millbrook published its 2014 report as a 39-page PDF. It lists 24 sampling sites and 0 readings above the action level, set out in a table on page 29. The table has 7 columns and is ruled only between rows, so the column positions have to come from the header. Note 13 in our log records that the Millbrook file mixes two fonts in the header and prints the header once. Northgate published its 2015 report as a 20-page PDF. It lists 26 sampling sites and 7 readings above the action level, set out in a table on page 16. The table has 9 columns and is ruled only between rows, so the column positions have to come from the header. Note 14 in our log records that the Northgate file uses one font throughout and repeats the header on every page. Oakridge shared its 2016 report as a 22-page PDF. It lists 68 sampling sites and 8 readings above the action level, set out in a table on page 14. The table has 9 columns and is ruled only between rows, so the column positions have to come from the header. Note 15 in our log records that the Oakridge file uses one font throughout and prints the header once. Pinecrest shared its 2017 report as a 47-page PDF. It lists 47 sampling sites and 8 readings above the action level, set out in a table on page 46. The table has 7 columns and is ruled only between rows, so the column positions have to come from the header. Note 16 in our log records that the Pinecrest file mixes two fonts in the header and repeats the header on every page. Queensford released its 2018 report as a 18-page PDF. It lists 59 sampling sites and 6 readings above the action level, set out in a table on page 10. The table has 4 columns and is ruled only between rows, so the column positions have to come from the header. Note 17 in our log records that the Queensford file uses one font throughout and prints the header once. Riverton shared its 2019 report as a 47-page PDF. It lists 12 sampling sites and 2 readings above the action level, set out in a table on page 31. The table has 8 columns and is ruled only between rows, so the column positions have to come from the header. Note 18 in our log records that the Riverton file uses one font throughout and repeats the header on every page. Stonehill released its 2020 report as a 18-page PDF. It lists 70 sampling sites and 5 readings above the action level, set out in a table on page 17. The table has 7 columns and is ruled only between rows, so the column positions have to come from the header. Note 19 in our log records that the Stonehill file mixes two fonts in the header and prints the header once. Thornbury shared its 2021 report as a 33-page PDF. It lists 73 sampling sites and 7 readings above the action level, set out in a table on page 4. The table has 6 columns and is ruled only between rows, so the column positions have to come from the header. Note 20 in our log records that the Thornbury file uses one font throughout and repeats the header on every page. Upton filed its 2022 report as a 12-page PDF. It lists 48 sampling sites and 7 readings above the action level, set out in a table on page 10. The table has 6 columns and is ruled only between rows, so the column positions have to come from the header. Note 21 in our log records that the Upton file uses one font throughout and prints the header once. Valemont filed its 2023 report as a 33-page PDF. It lists 65 sampling sites and 2 readings above the action level, set out in a table on page 6. The table has 9 columns and is ruled only between rows, so the column positions have to come from the header. Note 22 in our log records that the Valemont file mixes two fonts in the header and repeats the header on every page. Westfield published its 2024 report as a 29-page PDF. It lists 42 sampling sites and 7 readings above the action level, set out in a