Appendix

1 PubTables1M Canonicalization Algorithm

1. Correct the column header annotation: In other words, determine if there are additional rows that should be added to the column header annotation.
    - (a) If the first row starts with a blank cell, label it as part of the column header.
    - (b) Split every blank spanning cell into blank grid cells (pre-processing step used by several future steps).
    - (c) If the first row is labeled as part of the column header, do the following to ensure that the column header continues to as many rows as necessary so that the column header tree has a unique leaf node in every column:
        - i. For every column, determine the first row in that column that has a cell that does not span multiple columns.
        - ii. Determine the last row of the rows identified in the previous step.
        - iii. Set every row in the table up to the row determined in the previous step to be part of the column header.
2. Canonicalize the column header: In other words, merge cells that are unnecessarily split into multiple cells.
    - (a) Merge any neighboring cells that occupy the exact same columns.
    - (b) Merge cells with blank cells below them: For every column header cell, occupying columns [M,M+1,...,M+C] and rows [N-R,N-R+1,...,N], merge it with any blank cells in rows [N+1,...,N+X] such that every cell occupying columns [M,M+1,...,M+C] and rows [N+1,...,N+X] is blank.
    - (c) Merge cells with blank cells above them: For every column header cell, occupying columns [M,M+1,...,M+C] and rows [N,N+1,...,N+R], merge it with any blank cells in rows [N-1,...,N-X] such that every cell occupying columns [M,M+1,...,M+C] and rows [N-1,...,N-X] is blank.
3. Infer (portions of) the row header annotation: Detect some portions of the row header, which is not annotated at all in this dataset.
    - (a) Detect projected row headers in the table: These are rows below the column header where only one cell in the row is not blank.
    - (b) Detect if the first column is part of the row header: If any cells below the column header in the first column are spanning cells or are blank, the first column is part of the row header.
4. Canonicalize the row header: In other words, merge cells that are unnecessarily split into multiple cells.
    - (a) For every projected row header, merge all of the cells in the row into a single cell.
    - (b) Merge any blank cells in the row header with the first non-blank cell above it and any blank cells in between, if all of these cells only occupy the same columns and these columns are all in the row header.