# 21.0.3 – SUPPORTED TABLE STRUCTURES

## Table of Contents

## Factors influencing Table Extraction

The extraction of tables depends on the following:

1. **The OCR output of the page** i.e., how well the words and lines on the page are detected by the OCR Engine.
2. **The table fields annotated or defined in Automation Document Processing (ADP) Designer which become part of the document type definition**. For any header of the table to be extracted, it must be annotated as the table header OR have the text of the header present in the list of possible names for the table header.
3. For the **summary** to be extracted, it **should be explicitly defined** in the table field.

## Known Limitations in Table Extraction

Here are some factors that influence table detection.

1. **Inappropriate Vertical lines between table headers** might impact the header extraction. For example, in a header row, if only a few headers have a vertical line separation, but others do not, then it's possible that the text of headers not separated by lines might be grouped together thus impacting the effectiveness of the column extraction.
2. **Multi-line table headers with non-aligned lines:** If headers have multiple lines and the alignment of these lines are different across different headers, then the header for the table might not be extracted properly.
3. **Tables with rows not separated by lines** might have some problems in accuracy of extraction. For example, each line in a multi-line row might be treated as separate rows OR multiple rows might be treated as a single row with multiple lines.
4. **For grid-less table(s), the horizontal and vertical alignment of data matters:** If the table in the document does not have lines that separate rows and columns, then the extraction of rows and columns depends on the horizontal and vertical alignment of text respectively.
5. **Single row table without grid lines:** In a grid-less table, any data in the document that is below the table and aligns the table headers might be treated as part of the table.
6. **Any watermarks on the page** that interferes with the OCR extraction of words in tables might impact the extraction results.
7. **Additional fields** annotated as part of the table definition in ADP Designer are not yet supported for extraction.
8. **Inverse Header Table** where the table header text is in a white font against a solid dark color might cause problems in detecting drawn lines, which might reduce table extraction accuracy. Also, **rows having alternate colors** might be detected inaccurately.
9. Tables with **Hierarchical Headers** are not supported.
10. Tables with **paragraphs of text between Header and Data** might have gaps in detection.

# Supported tables

## 1. Standard Grid lines table

Here are the characteristics of the table:

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- The header is the first line of the table.
- The header can span multiple lines. Example: **ITEM DESCRIPTION**.
- The cell can span multiple lines. Example: HARD BOUND NOTEBOOK.
- Some of the cells can be empty. Example: For the WRITING PAD row, the 2nd and 4th cells are empty.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| HARD BOUND NOTEBOOK | 800 | 1 | 800 |
| WRITING PAD | | 2 | |

## 2. Tables with summary data defined as cells in table with blank cells to the left

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- The summary data is present as cells at the end of the table.
- The summary data is an L-shaped extension with the cells on the left of the summary seen as white space without any borders.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |
| | | | |
| | | | |
| | | SUBTOTAL | 1400 |
| | | TAX | 10% |
| | | GRAND TOTAL | 1540 |

## 3. Tables with blank rows in-between

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- There can be blank rows in-between.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| | | | |
| NOTEBOOK | 800 | 1 | 800 |
| WRITING PAD | 400 | 3 | 1200 |
| | | | |
| HARD BOUND NOTEBOOK | 800 | 1 | 800 |

## 4. Tables as part of a grid

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- Table spans the width of the page.
- Table is part of a bigger grid (especially for forms and similar documents) that might have other text outside the table, but in the same grid with different alignment.

| **Contact Address**<br>John Smith<br>New York | **Destination Address**<br>Jane Doe<br>Chicago | **Code**<br>AB123 | **Origin**<br>Australia | **Date**<br>01/01/2021 |
|---|---|---|---|---|

| ITEM ID | ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|---|
| 1 | PEN | 300 | 2 | 600 |
| 2 | NOTEBOOK | 800 | 1 | 800 |
| 3 | WRITING PAD | 400 | 3 | 1200 |
| | | | | |

## 5. Table with rows containing merged cells

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- Cells are merged and span a whole row in the table.
  Example: "Customer No. 123456789" that spans the whole row.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| Customer No: 123456789 | | | |
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |
| | | | |

## 6. Tables with summary data defined as cells in table

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- The summary is present as cells to the end of the table with the cells to the left of summary seen as empty cells with borders.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |
| | | | |
| | | | |
| | | SUBTOTAL | 1400 |
| | | TAX | 10% |
| | | GRAND TOTAL | 1540 |

# Partially Supported and Unsupported Tables

## 1. Header with different background color

- The header row has a different background color than the color of the cells.
- Here, the headers might not be extracted as expected.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| HARD BOUND NOTEBOOK | 800 | 1 | 800 |
| WRITING PAD | | 2 | |
| | | | |

## 2. Tables with summary data defined partially with cell borders

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- The summary data is present as cells to the end of the table.
- The summary data does not have the borders defined, but its values are defined inside borders.
- Here, the summary header might be part of the table and hence may not be extracted properly.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |
| | | | |
| | | SUBTOTAL | 1400 |
| | | TAX | 10% |
| | | GRAND TOTAL | 1540 |

## 3. Tables with summary data defined outside the table

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- The summary data present outside the table bounding box might not be extracted.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |
| | | | |

| | |
|---|---|
| SUBTOTAL | 1400 |
| TAX | 10% |
| GRAND TOTAL | 1540 |

## 4. Tables with additional properties defined beside summary data

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- Summary can be present at the bottom of the table, and there can be additional data present beside the summary.
- In such cases, summary might not be properly extracted.
- For Example: Account No, Payment Terms are present beside the summary.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |
| | | | |
| | | | |

| | | |
|---|---|---|
| Account No: 12345678 | SUBTOTAL | 1400 |
| Payment Terms: Payment within 30 days | TAX | 10% |
| | GRAND TOTAL | 1540 |

## 5. Tables with watermark text or stamps

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- Table shall have a watermark or stamp that overlaps the headers or row data. In such cases, there might be problems with the accuracy of the extraction.
- In the following examples:
    - the watermark overlaps with the column header thus impacting the detection of the column.
    - the stamp overlaps the table, which impacts the table data extraction. Note that stamps not only add random text, but they can also have lines which further impacts recognition and line detection.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |
| | | | |
| | | | |



| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |
| | | | |
| | | | |



## 6. Tables with Hierarchical Headers

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- The Headers are hierarchical in nature, in that multiple (sub) headers are grouped logically under one header.
- Here, the table might not get extracted at all OR might not be extracted as expected.
- Example: Instead of having unique headers - ACCESSORIES QUANTITY and ACCESSORIES PRICE, they are consolidated under a single header - ACCESSORIES.

| ITEM DESCRIPTION | ACCESSORIES | | SPORTS | | TOTAL |
|---|---|---|---|---|---|
| | QUANTITY | PRICE | QUANTITY | PRICE | |
| KIT BAG | 30 | 200 | 60 | 100 | 12000 |
| GLOVES | 80 | 100 | 80 | 50 | 12000 |
| | | | | | |
| | | | | | |

## 7. Tables with paragraph of text between Header and Data

Here are the characteristics of the table.

- The table may be surrounded by a border.
- There could be a paragraph with text that does not align with the header columns.
- Table might not be extracted.

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| There could be some random text that might mis-align with the headers (as this paragraph) thus making it difficult to extract the table. | | | |
| NOTEBOOK | 800 | 1 | 800 |
| WRITING PAD | 400 | 3 | 1200 |
| | | | |
| HARD BOUND NOTEBOOK | 800 | 1 | 800 |

## 8. Tables attached to each other

Here are the characteristics of the table.

- The table is surrounded by a border.
- Headers, rows, and cells in the row are separated by lines.
- Tables have their own alignment.
- Tables might be attached, for example there is no space that separates the tables.

| Agent | Code | Location | Terms |
|---|---|---|---|
| John Smith | 111 | New York | Net 300 |
| Jane Doe | 222 | Chicago | Net 600 |

| ITEM ID | ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|---|
| 1 | PEN | 300 | 2 | 600 |
| 2 | NOTEBOOK | 800 | 1 | 800 |
| 3 | WRITING PAD | 400 | 3 | 1200 |
| | | | | |

## 9. Tables without borders

Here are the characteristics of the table.

- The table is not surrounded by any kind of border.
- Headers, rows, and cells in the row are not separated by lines.
- The header can be distinguished from the values of the table.
- Note: The extraction depends on the alignment of text.

Header has a bold font

| ITEM DESCRIPTION | UNIT PRICE | QUANTITY | TOTAL PRICE |
|---|---|---|---|
| PEN | 300 | 2 | 600 |
| NOTEBOOK | 800 | 1 | 800 |