# GemBox Support Center



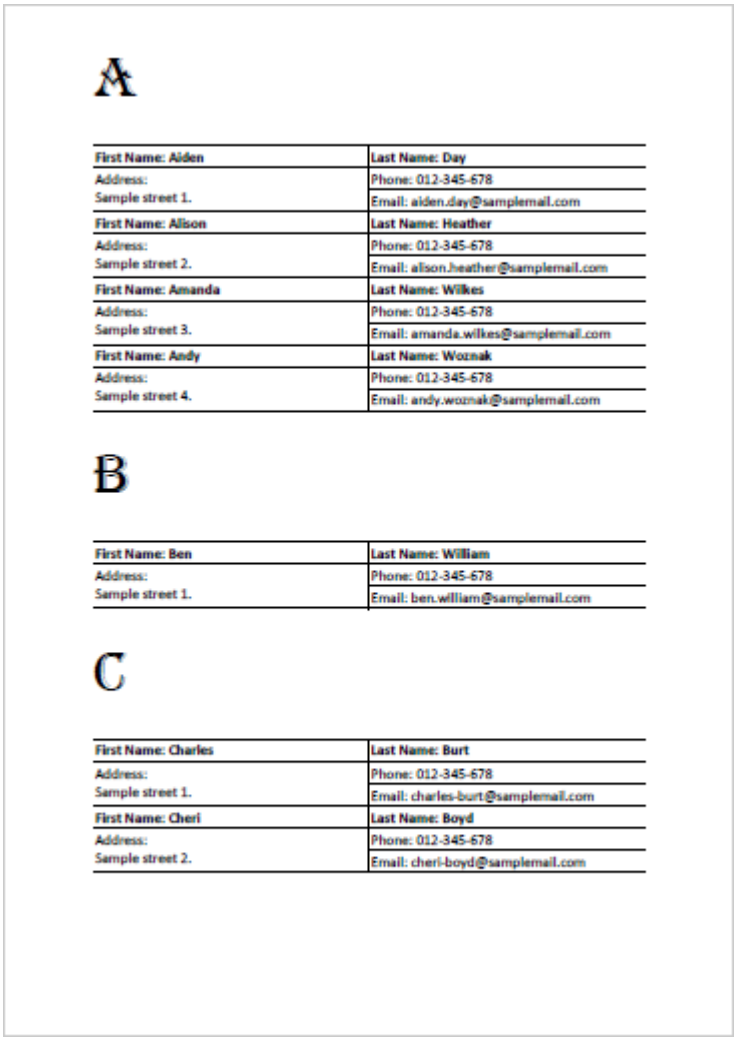
## Reading PDF files and extracting table elements

Mario - GemBox - 2017-02-10 - 0 Comments - in GemBox.Document

GemBox.Document can now read PDF files, which will be progressively improved over time based on user feedback. The current support level for reading PDF is described on this help page.

The following sample will demonstrate how we can read a PDF file that contains multiple table elements, extract the content of those table elements, and import the content into a single HTML table.

The following sample input file ("Address Book.pdf") is used:



A

| First Name: Alden | Last Name: Day |
|---|---|
| Address:<br>Sample street 1. | Phone: 012-345-678<br>Email: alden.day@samplemail.com |
| **First Name: Alison** | **Last Name: Heather** |
| Address:<br>Sample street 2. | Phone: 012-345-678<br>Email: alison.heather@samplemail.com |
| **First Name: Amanda** | **Last Name: Wilkes** |
| Address:<br>Sample street 3. | Phone: 012-345-678<br>Email: amanda.wilkes@samplemail.com |
| **First Name: Andy** | **Last Name: Woznak** |
| Address:<br>Sample street 4. | Phone: 012-345-678<br>Email: andy.woznak@samplemail.com |

B

| First Name: Ben | Last Name: William |
|---|---|
| Address:<br>Sample street 1. | Phone: 012-345-678<br>Email: ben.william@samplemail.com |

C

| First Name: Charles | Last Name: Burt |
|---|---|
| Address:<br>Sample street 1. | Phone: 012-345-678<br>Email: charles-burt@samplemail.com |
| **First Name: Cheri** | **Last Name: Boyd** |
| Address:<br>Sample street 2. | Phone: 012-345-678<br>Email: cheri-boyd@samplemail.com |

This PDF file contains multiple table elements, which store some example contacts. We'll read through all of those contact entries and add them to the table element in "Template.html" file.

### C# code

```
// Load PDF document.
DocumentModel pdfDocument = DocumentModel.Load("Address Book.pdf");
// Load HTML template document.
DocumentModel htmlDocument = DocumentModel.Load("Template.html");

// Get HTML document's table element.
Table htmlTable = (Table)htmlDocument.Sections[0].Blocks[0];

// Iterate through PDF document's table elements.
foreach (Table table in pdfDocument.GetChildElements(true, ElementType.Table).Cast<Tabl
e>())
{
   for (int i = 0; i < table.Rows.Count; i += 3)
   {
      // Extract PDF document's table content.
      TableCell cell = table.Rows[i].Cells[0];
      string name = cell.Content.ToString().Replace("First Name:", "").Trim();

      cell = table.Rows[i].Cells[1];
      string surname = cell.Content.ToString().Replace("Last Name:", "").Trim();

      cell = table.Rows[i + 1].Cells[0];
      string address = cell.Content.ToString().Replace("Address:", "").Trim();

      cell = table.Rows[i + 1].Cells[1];
      string phone = cell.Content.ToString().Replace("Phone:", "").Trim();

      cell = table.Rows[i + 2].Cells[0];
      string email = cell.Content.ToString().Replace("Email:", "").Trim();

      // Add new table row with table cells.
      htmlTable.Rows.Add(
         new TableRow(htmlDocument,
            new TableCell(htmlDocument,
               new Paragraph(htmlDocument, name))
            { CellFormat = { BackgroundColor = new Color(251, 228, 213) } },
            new TableCell(htmlDocument,
               new Paragraph(htmlDocument, surname))
            { CellFormat = { BackgroundColor = new Color(251, 228, 213) } },
            new TableCell(htmlDocument,
               new Paragraph(htmlDocument, address))
            { CellFormat = { BackgroundColor = new Color(251, 228, 213) } },
            new TableCell(htmlDocument,
               new Paragraph(htmlDocument, phone))
            { CellFormat = { BackgroundColor = new Color(251, 228, 213) } },
            new TableCell(htmlDocument,
               new Paragraph(htmlDocument, email))
            { CellFormat = { BackgroundColor = new Color(251, 228, 213) } }));
   }
}

// Save HTML document.
htmlDocument.Save("Address Book.html");
```

### VB.NET code

```
' Load PDF document.
Dim pdfDocument As DocumentModel = DocumentModel.Load("Address Book.pdf")
```

```
' Load HTML template document.
Dim htmlDocument As DocumentModel = DocumentModel.Load("Template.html")

' Get HTML document's table element.
Dim htmlTable As Table = DirectCast(htmlDocument.Sections(0).Blocks(0), Table)

' Iterate through PDF document's table elements.
For Each table As Table In pdfDocument.GetChildElements(True, ElementType.Table).Cast(
Of Table)()
    For i As Integer = 0 To table.Rows.Count - 1 Step 3
        ' Extract PDF document's table content.
        Dim cell As TableCell = table.Rows(i).Cells(0)
        Dim name As String = cell.Content.ToString().Replace("First Name:", "").Trim()

        cell = table.Rows(i).Cells(1)
        Dim surname As String = cell.Content.ToString().Replace("Last Name:", "").Trim()

        cell = table.Rows(i + 1).Cells(0)
        Dim address As String = cell.Content.ToString().Replace("Address:", "").Trim()

        cell = table.Rows(i + 1).Cells(1)
        Dim phone As String = cell.Content.ToString().Replace("Phone:", "").Trim()

        cell = table.Rows(i + 2).Cells(0)
        Dim email As String = cell.Content.ToString().Replace("Email:", "").Trim()

        ' Add new table row with table cells.
        htmlTable.Rows.Add(
            New TableRow(htmlDocument,
                New TableCell(htmlDocument,
                    New Paragraph(htmlDocument, name)) With { _
                .CellFormat = New TableCellFormat() With {.BackgroundColor = New Color(251, 2
28, 213)}},
                New TableCell(htmlDocument,
                    New Paragraph(htmlDocument, surname)) With { _
               .CellFormat = New TableCellFormat() With {.BackgroundColor = New Color(251, 2
28, 213)}},
                New TableCell(htmlDocument,
                    New Paragraph(htmlDocument, address)) With { _
               .CellFormat = New TableCellFormat() With {.BackgroundColor = New Color(251, 2
28, 213)}},
                New TableCell(htmlDocument,
                    New Paragraph(htmlDocument, phone)) With { _
                .CellFormat = New TableCellFormat() With {.BackgroundColor = New Color(251, 2
28, 213)}},
                New TableCell(htmlDocument,
                    New Paragraph(htmlDocument, email)) With { _
                .CellFormat = New TableCellFormat() With {.BackgroundColor = New Color(251, 2
28, 213)}}))
    Next
Next

' Save HTML document.
htmlDocument.Save("Address Book.html")
```

The following is the resulting "Address Book.html" file:

| Name | Surname | Address | Phone | Email |
|---|---|---|---|---|
| Aiden | Day | Sample street1. | 012-345-678 | aiden.day@samplemail.com |
| Alison | Heather | Sample street2. | 012-345-678 | alison.heather@samplemail.com |
| Amanda | Wilkes | Sample street3. | 012-345-678 | amanda.wilkes@samplemail.com |
| Andy | Woznak | Sample street4. | 012-345-678 | andy.woznak@samplemail.com |
| Ben | William | Sample street1. | 012-345-678 | ben.william@samplemail.com |
| Charles | Burt | Sample street1. | 012-345-678 | charles-burt@samplemail.com |
| Cheri | Boyd | Sample street2. | 012-345-678 | cheri-boyd@samplemail.com |