EDTA, iText and INBATEK Conference

# Extracting Data

iText Presentation 26/07/2017

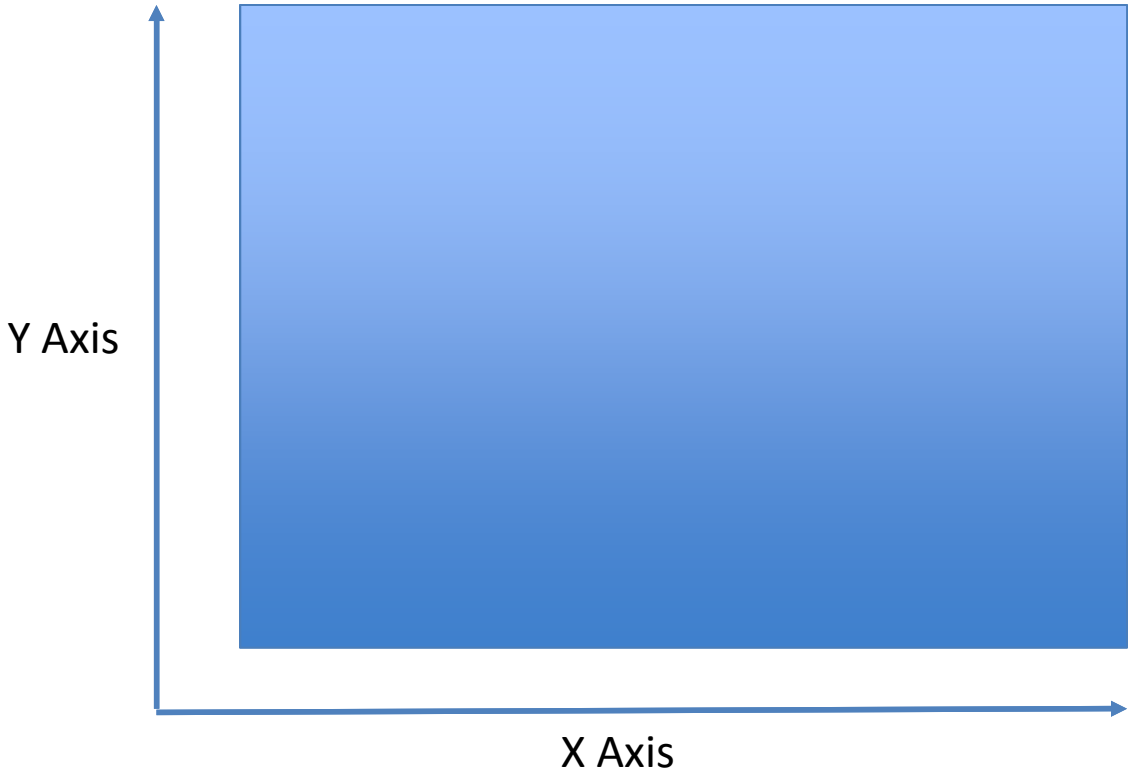
# Extracting Data

- Parsing PDF file to Text
- Parsing PDF file to Data records

# Parsing PDF file to Text

- Extraction of text from the PDF file
  - Uses **PdfTextExtractor** Class
- Extraction done using
  - Default Strategy
  - Location based Strategy
- Location Based strategy
  - Define rectangle by giving dimensions
    - X value, Y value, Width and Height
  - Text snippets that intersect the rectangle are extracted

# Parsing PDF file to Text



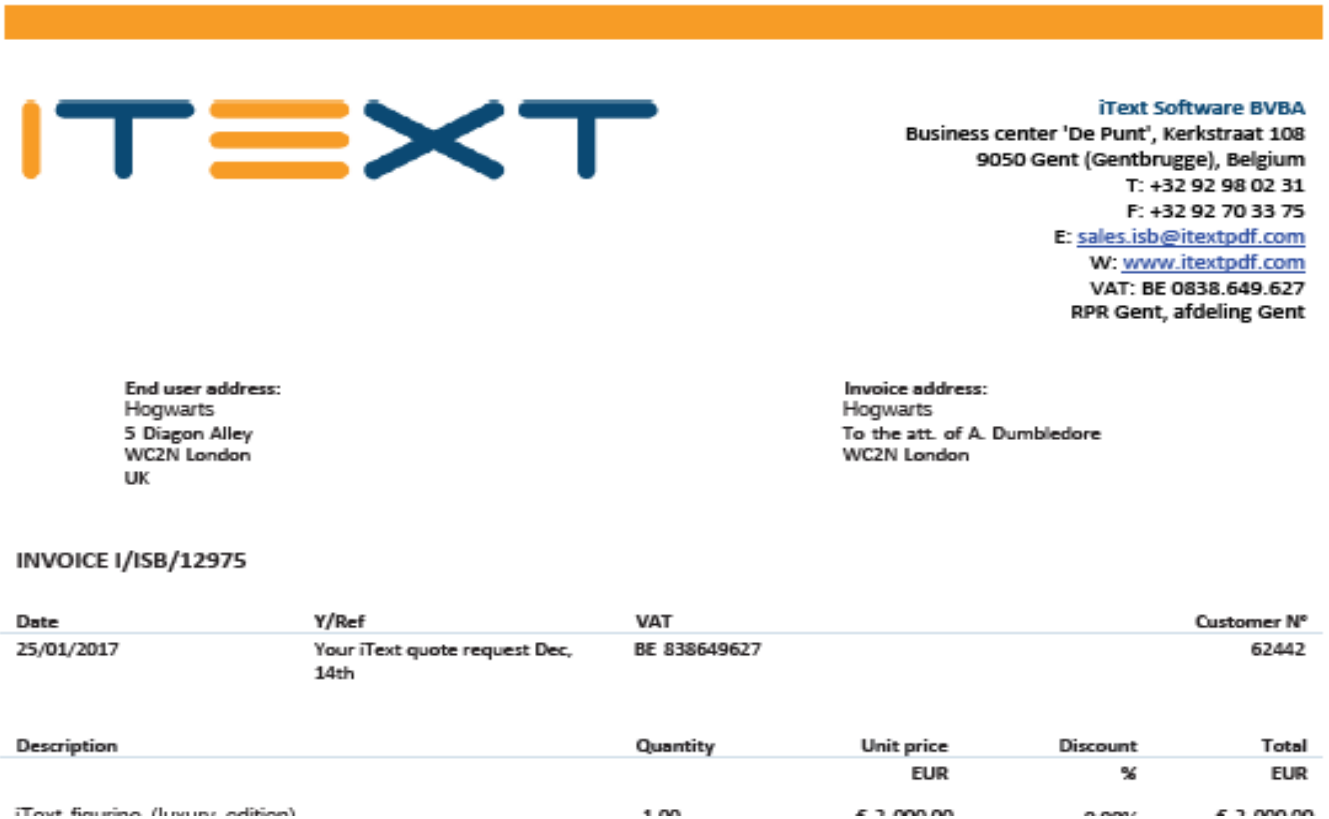

- Unit of the measurement in PDF is called "user unit".
- 72 user units = 1 inch

# Parsing PDF file to Records : pdf2Data

- What is pdf2Data
  - Tool to automate data extraction from similar PDFs
  - Extracted data in XML
- Use of pdf2Data
  - Good fit for PDFs that are in a **template format**
  - Documents that look the same, but have different information
    - Invoices
    - Internal reports
- Data extraction based on predefined fields/rules

# pdf2Data workflow

# Step 1: Upload Template PDF



iText Software BVBA
Business center 'De Punt', Kerkstraat 108
9050 Gent (Gentbrugge), Belgium
T: +32 92 98 02 31
F: +32 92 70 33 75
E: sales.isb@itextpdf.com
W: www.itextpdf.com
VAT: BE 0838.649.627
RPR Gent, afdeling Gent

End user address:
Hogwarts
5 Diagon Alley
WC2N London
UK

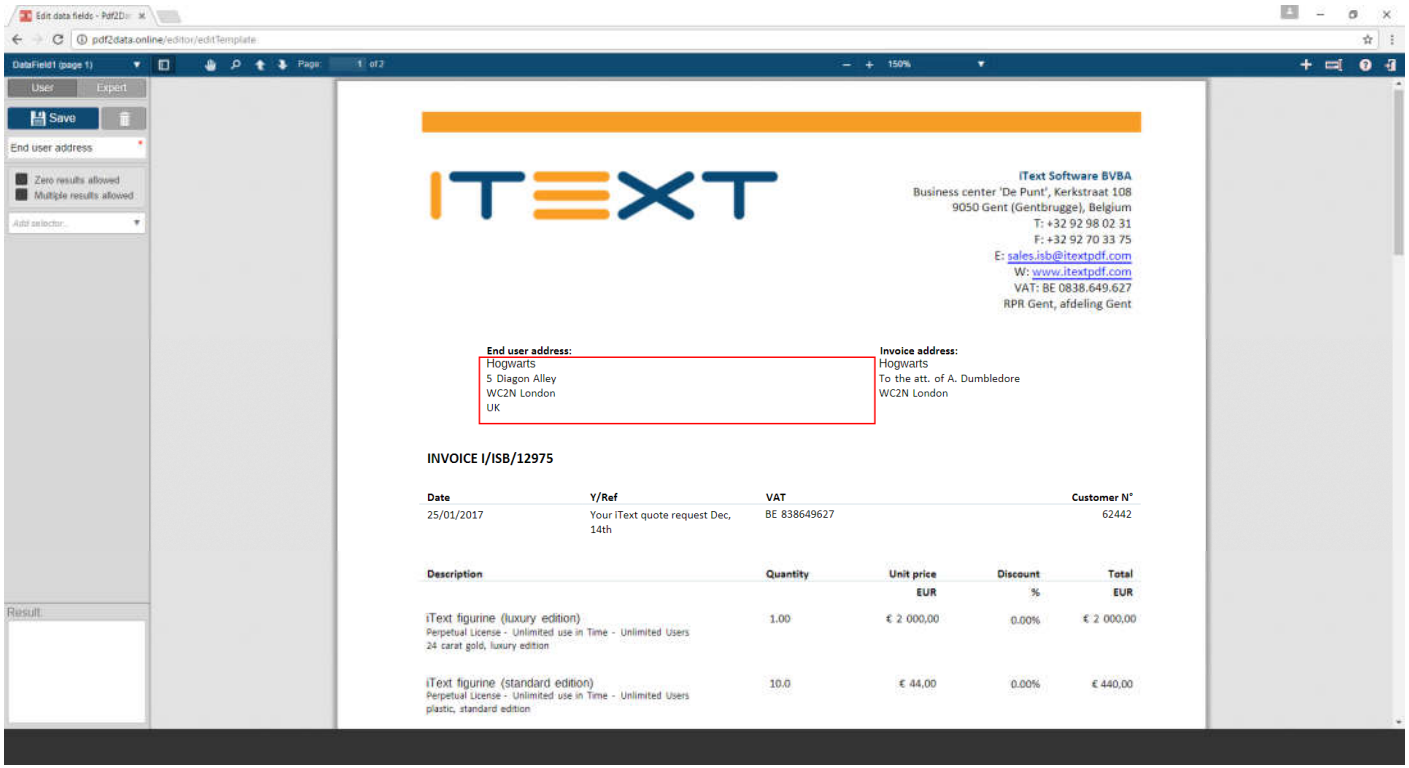
Invoice address:
Hogwarts
To the att. of A. Dumbledore
WC2N London

INVOICE I/ISB/12975

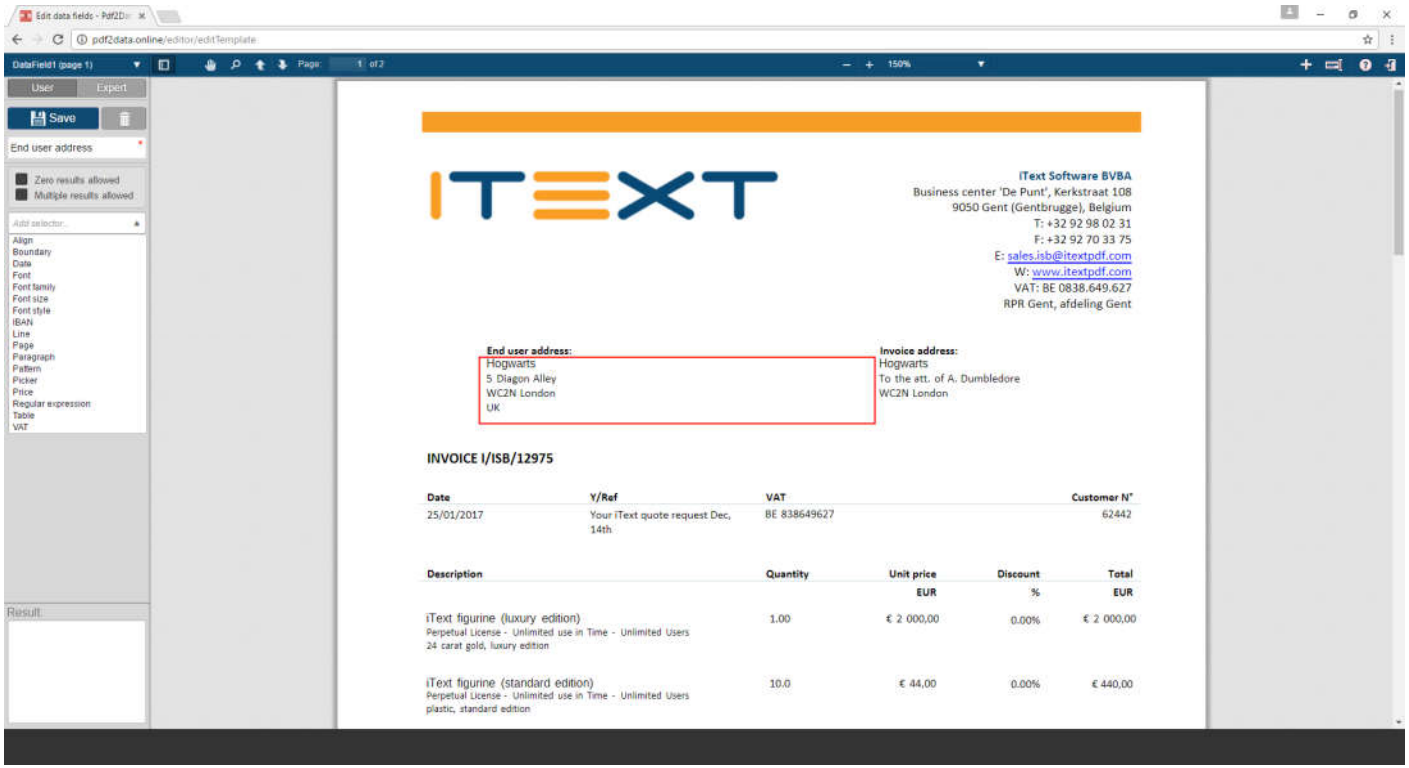
| Date | Y/Ref | VAT | Customer N° |
|---|---|---|---|
| 25/01/2017 | Your iText quote request Dec, 14th | BE 838649627 | 62442 |

| Description | Quantity | Unit price | Discount | Total |
|---|---|---|---|---|
| | | EUR | % | EUR |
| iText figurine (luxury edition) | 1.00 | € 2,000.00 | 0.00% | € 2,000.00 |

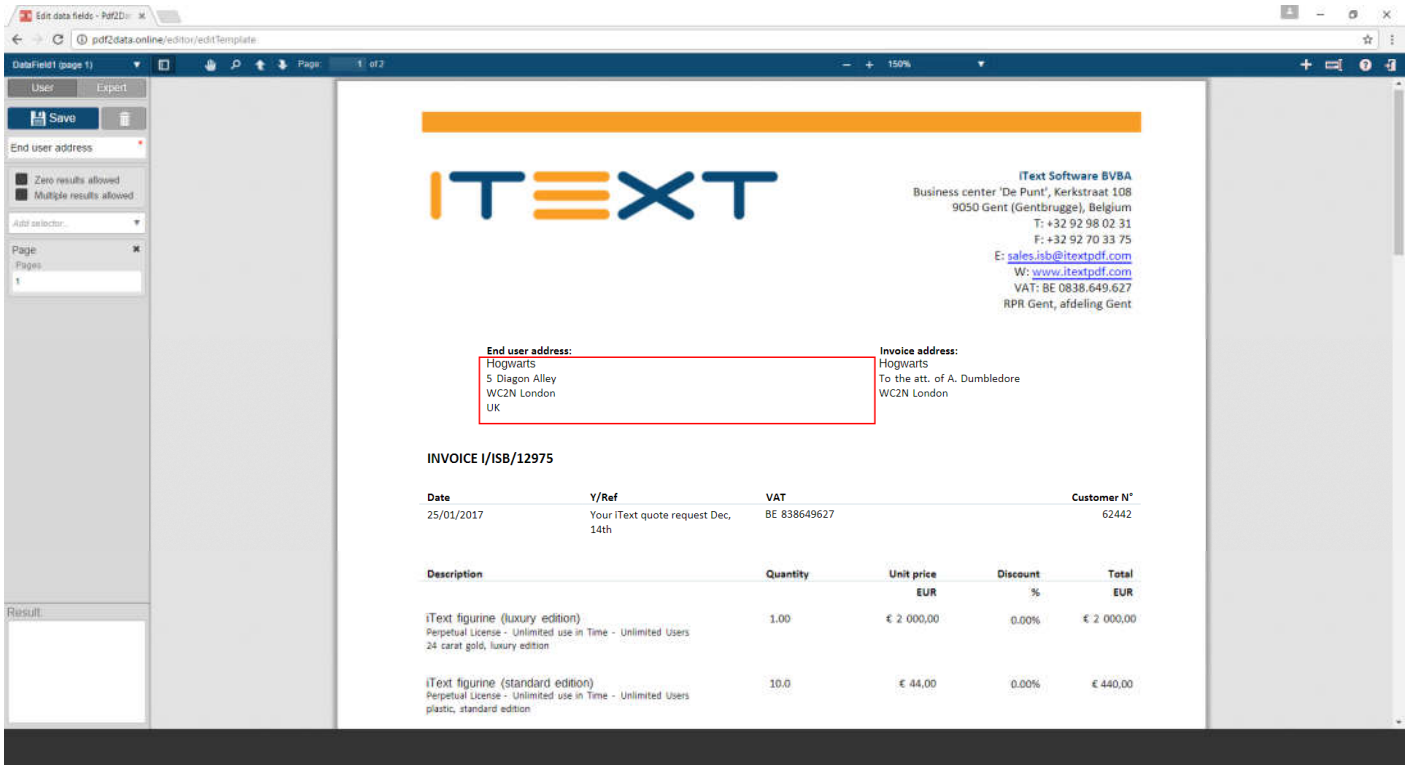
# Step 2: Mark extraction area



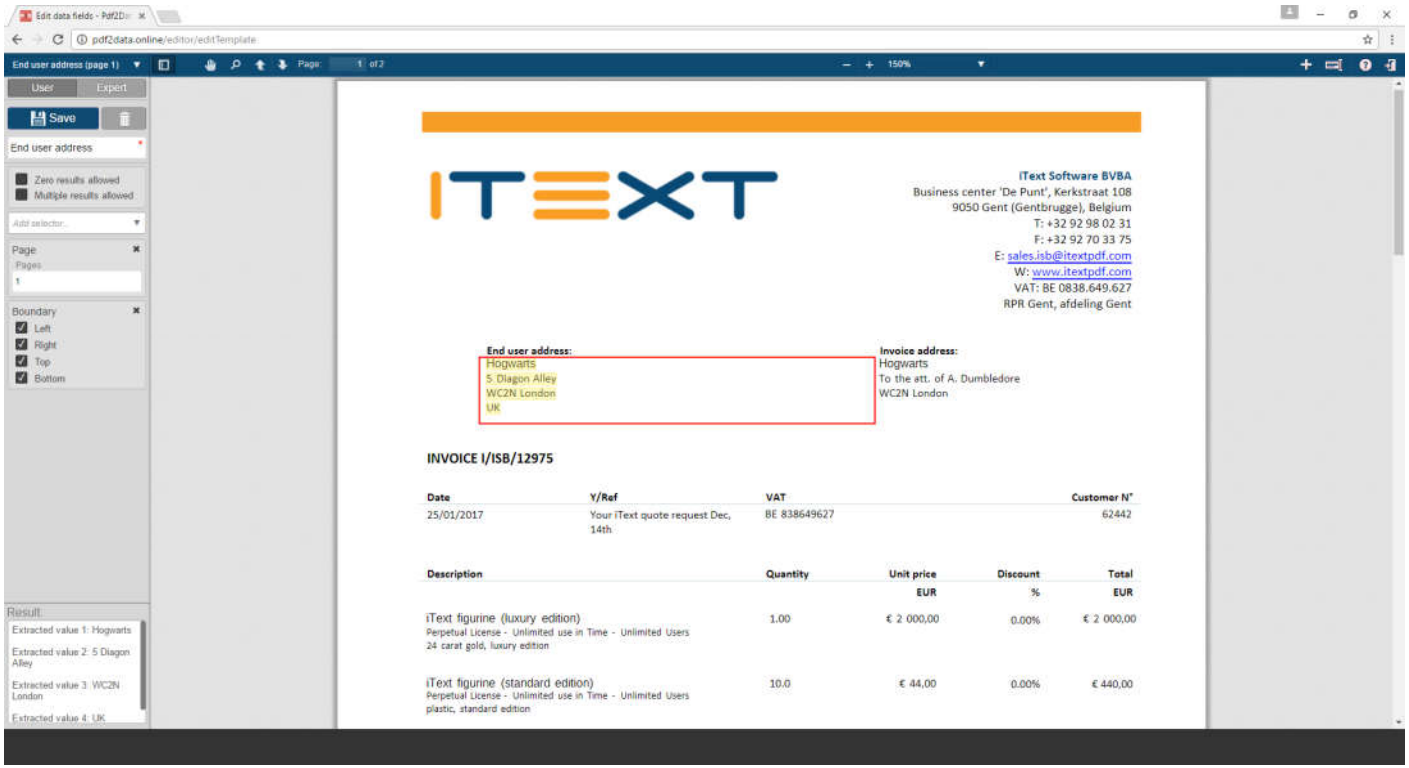
# Step 3: Define Extraction Rules



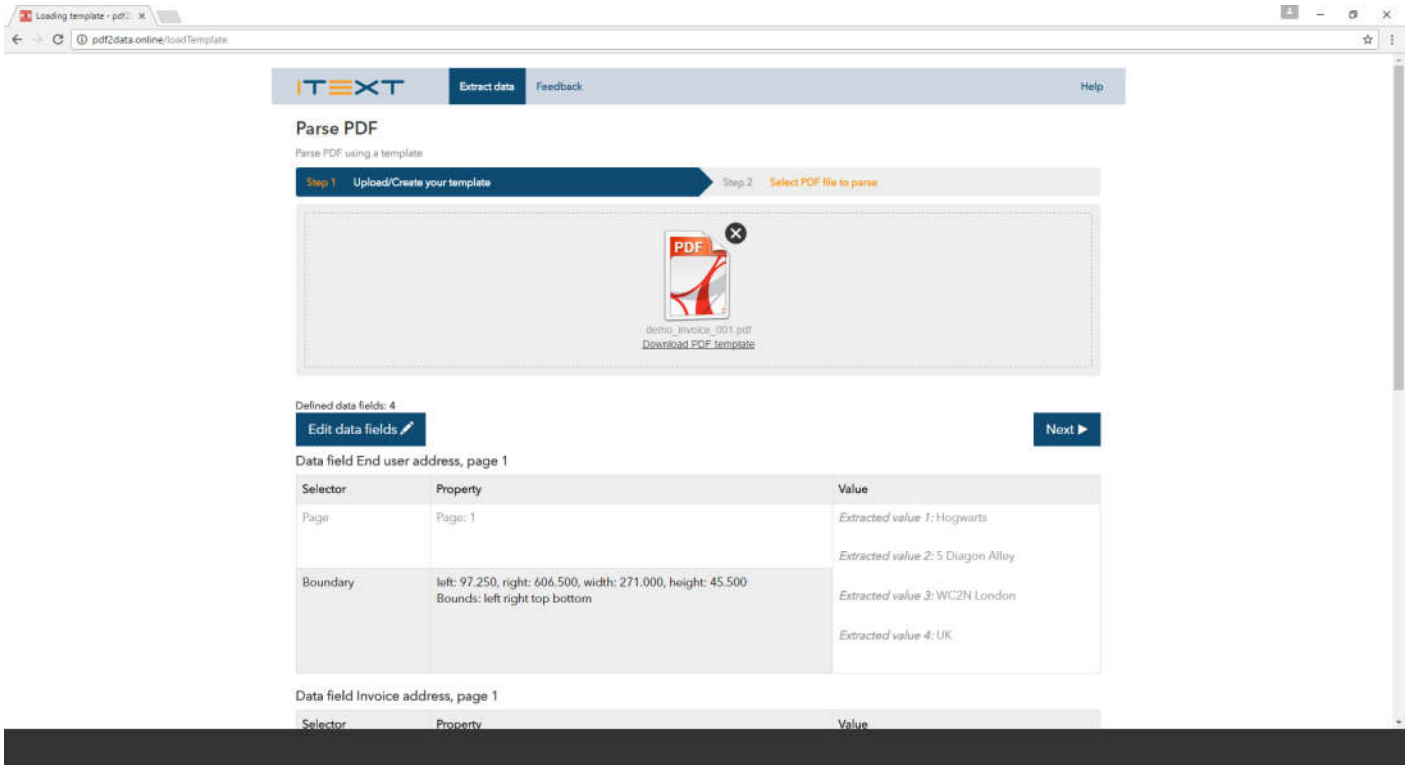
# Step 4: Select Page



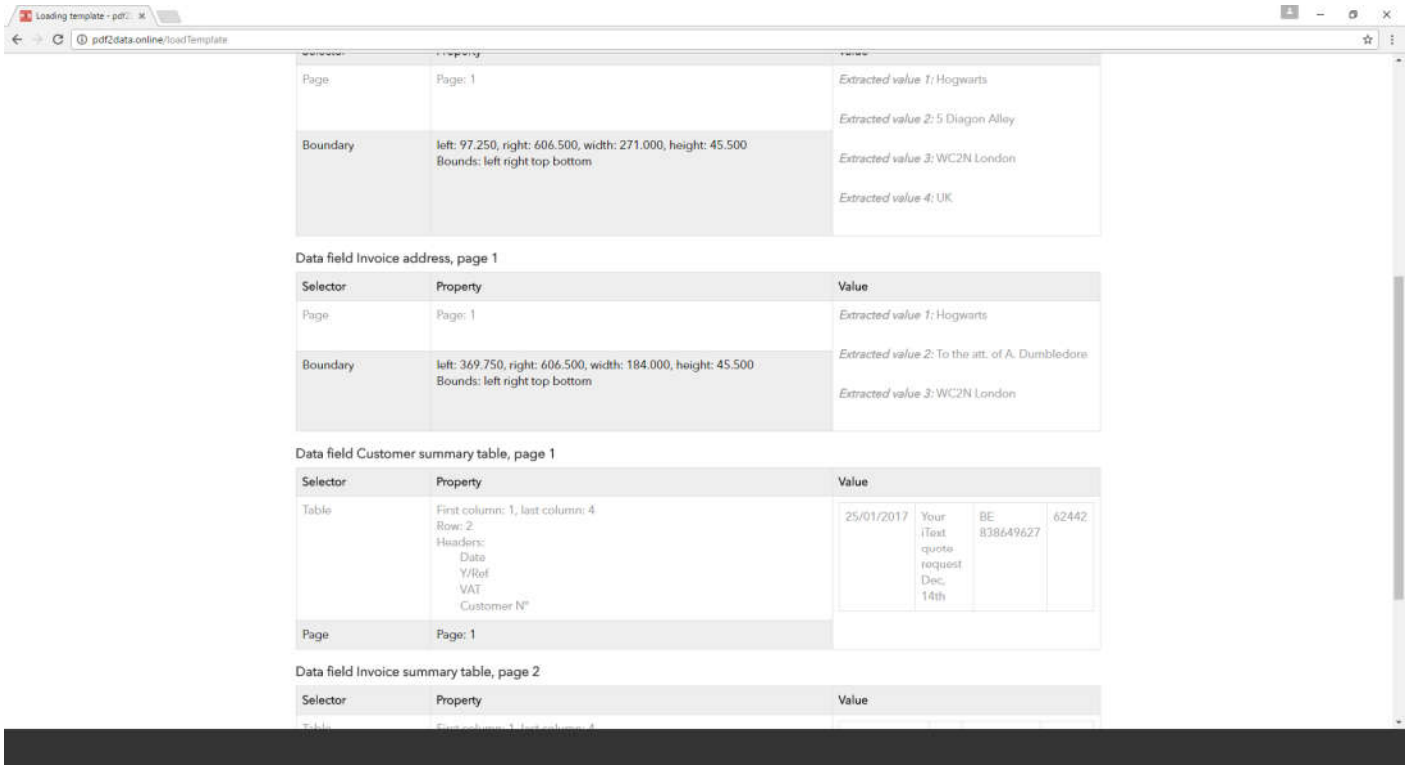
# Step 5: Save the Template



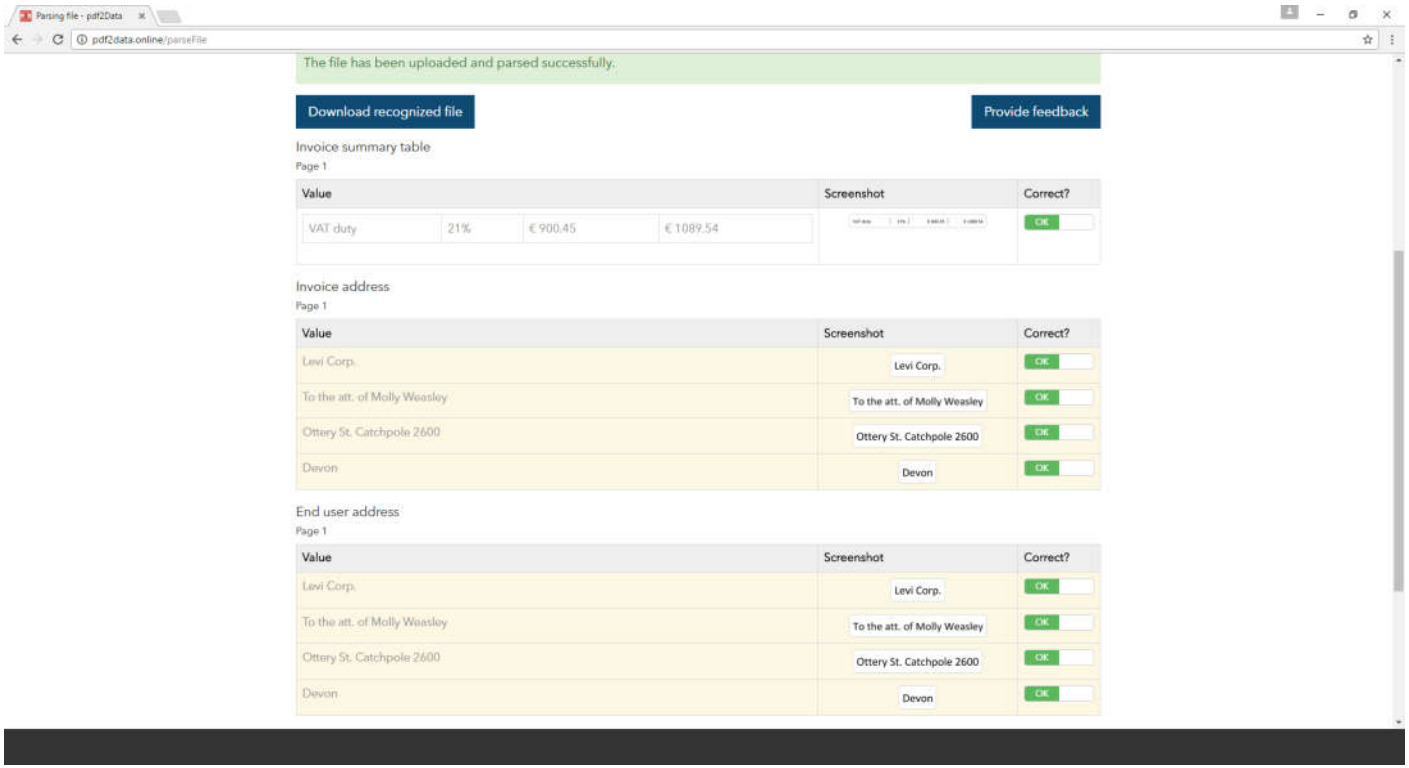
# Step 6: Return to Application



# Step 7: Verify the Template



# Step 8 : Extract the Data



# XML Data

```xml
<?xml version="1.0" encoding="UTF-8"?><elements>
  <data name="Customer summary table">
    <table>
      <row>
        <column>14/02/2017</column>
        <column>Your iText subscription 04/2017-
04/2018 PO-000288</column>
        <column>BE 7248162952</column>
        <column>54320</column>
      </row>
    </table>
  </data>
  <data name="DataField4">
    <table>
      <row>
        <column>VAT duty</column>
        <column>21%</column>
        <column>€ 900.45</column>
        <column>€ 1089.54</column>
      </row>
    </table>
  </data>
  <data name="End user address">
    <text>Levi Corp.</text>
    <text>To the att. of Molly Weasley</text>
    <text>Ottery St. Catchpole 2600</text>
    <text>Devon</text>
    <text>UK</text>
  </data>
  <data name="Invoice address">
    <text>Levi Corp.</text>
    <text>To the att. of Molly Weasley</text>
    <text>Ottery St. Catchpole 2600</text>
    <text>Devon</text>
    <text>UK</text>
  </data>
</elements>
```

# Programmatically extracting data

- Define selectors using Adobe reader comments
- Code to extract data to XML

```
Pdf2DataExtractor ext = new Pdf2DataExtractor(template);

ext.parsePdf(sampleFile, targetPDF, targetXML);
```

# Thank You!

- Log on to http://itextpdf.com/
- Log on to http://developers.itextpdf.com/