



EDTA, iText and INBATEK Conference

# Extracting Data

iText Presentation 26/07/2017



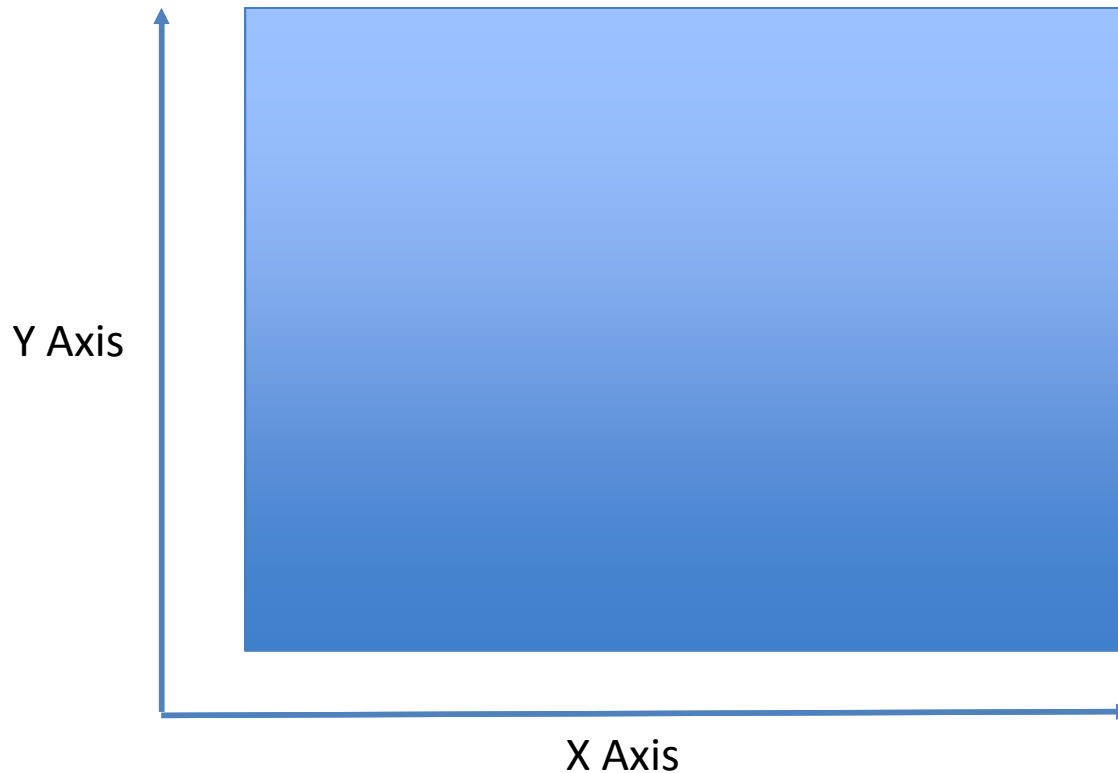
# ≡ Extracting Data

- 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](mailto:sales.isb@itextpdf.com)  
W: [www.itextpdf.com](http://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 EUR	Discount %	Total EUR
iText financing (luxury edition)	1.00	€ 2.000.00	0.00%	€ 2.000.00

# Step 2: Mark extraction area

The screenshot shows the pdf2Data online editor interface. The main document is an invoice from iText Software BVBA. A red rectangular box highlights the 'End user address' field, which contains the following text:

**End user address:**  
Hogwarts  
5 Diagon Alley  
WC2N London  
UK

The invoice also includes the following information:

**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](mailto:sales.isb@itextpdf.com)  
W: [www.itextpdf.com](http://www.itextpdf.com)  
VAT: BE 0838.649.627  
RPR Gent, afdeling Gent

**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 EUR	Discount %	Total EUR
iText figurine (luxury edition) Perpetual License - Unlimited use in Time - Unlimited Users 24 carat gold, luxury edition	1.00	€ 2 000,00	0.00%	€ 2 000,00
iText figurine (standard edition) Perpetual License - Unlimited use in Time - Unlimited Users plastic, standard edition	10.0	€ 44,00	0.00%	€ 440,00



# Step 3: Define Extraction Rules

The screenshot shows the pdf2Data online editor interface. The main document is an invoice from iText Software BVBA. The end user address is highlighted with a red box, indicating that an extraction rule has been defined for it. The invoice details are as follows:

**ITEXT**

**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](mailto:sales.isb@itextpdf.com)  
W: [www.itextpdf.com](http://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 EUR	Discount %	Total EUR
iText figurine (luxury edition) Perpetual License - Unlimited use in Time - Unlimited Users 24 carat gold, luxury edition	1.00	€ 2 000,00	0.00%	€ 2 000,00
iText figurine (standard edition) Perpetual License - Unlimited use in Time - Unlimited Users plastic, standard edition	10.0	€ 44,00	0.00%	€ 440,00

# Step 4: Select Page

The screenshot shows the pdf2Data online editor interface. The browser address bar displays 'pdf2data.online/editor/editTemplate'. The main content area shows an invoice template with the iText logo and contact information for iText Software BVBA. Two data fields are highlighted with red boxes: 'End user address' and 'Invoice address'. The 'End user address' field contains the text: 'Hogwarts', '5 Diagon Alley', 'WC2N London', 'UK'. The 'Invoice address' field contains the text: 'Hogwarts', 'To the att. of A. Dumbledore', 'WC2N London'. Below the addresses, the invoice number 'INVOICE I/ISB/12975' is displayed. A table of invoice details follows, including columns for Date, Y/Ref, VAT, and Customer N°. The table contains two rows of data for iText figurines (luxury and standard editions).

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

Description	Quantity	Unit price EUR	Discount %	Total EUR
iText figurine (luxury edition) Perpetual License - Unlimited use in Time - Unlimited Users 24 carat gold, luxury edition	1.00	€ 2 000,00	0.00%	€ 2 000,00
iText figurine (standard edition) Perpetual License - Unlimited use in Time - Unlimited Users plastic, standard edition	10.0	€ 44,00	0.00%	€ 440,00

# Step 5: Save the Template

The screenshot shows the pdf2Data online editor interface. The main content area displays an invoice template with the iText logo and contact information for iText Software BVBA. Two data fields are highlighted with red boxes: 'End user address' and 'Invoice address'. The 'End user address' field contains the text: 'Hogwarts', '5 Diagon Alley', 'WC2N London', and 'UK'. The 'Invoice address' field contains the text: 'Hogwarts', 'To the att. of A. Dumbledore', and 'WC2N London'. Below the invoice details, there is a table with columns for Date, Y/Ref, VAT, and Customer N°. The table contains two rows of data. Below the table, there is a table with columns for Description, Quantity, Unit price, Discount, and Total. The table contains two rows of data.

**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) Perpetual License - Unlimited use in Time - Unlimited Users 24 carat gold, luxury edition	1.00	€ 2 000,00	0.00%	€ 2 000,00
iText figurine (standard edition) Perpetual License - Unlimited use in Time - Unlimited Users plastic, standard edition	10.0	€ 44,00	0.00%	€ 440,00

**Result:**  
Extracted value 1: Hogwarts  
Extracted value 2: 5 Diagon Alley  
Extracted value 3: WC2N London  
Extracted value 4: UK

# Step 6: Return to Application

The screenshot shows the iText web application interface for parsing PDFs. The browser address bar shows 'pdf2data.online/loadTemplate'. The page title is 'Parse PDF' and the subtitle is 'Parse PDF using a template'. The progress bar indicates 'Step 1 Upload/Create your template' is active, and 'Step 2 Select PDF file to parse' is next. A PDF icon is shown with the filename 'demo\_invoice\_001.pdf' and a 'Download PDF template' link. Below this, it says 'Defined data fields: 4' with an 'Edit data fields' button. A 'Next' button is also present. The main content area shows 'Data field End user address, page 1' with a table of extracted values.

Selector	Property	Value
Page	Page: 1	Extracted value 1: Hogwarts
		Extracted value 2: 5 Diagon Alley
Boundary	left: 97.250, right: 606.500, width: 271.000, height: 45.500 Bounds: left right top bottom	Extracted value 3: WC2N London
		Extracted value 4: UK

Below this, the start of another table is visible: 'Data field Invoice address, page 1' with columns for Selector, Property, and Value.

# Step 7: Verify the Template

The screenshot shows a web browser window with the URL [pdf2data.online/loadTemplate](http://pdf2data.online/loadTemplate). The interface displays a configuration for a PDF template, showing several data fields with their selectors, properties, and extracted values.

**Data field Invoice address, page 1**

Selector	Property	Value
Page	Page: 1	Extracted value 1: Hogwarts
Boundary	left: 97.250, right: 606.500, width: 271.000, height: 45.500 Bounds: left right top bottom	Extracted value 2: 5 Diagon Alley Extracted value 3: WC2N London Extracted value 4: UK

**Data field Customer summary table, page 1**

Selector	Property	Value
Table	First column: 1, last column: 4 Row: 2 Headers: Data Y/Ref VAT Customer N°	25/01/2017 Your iText quote request Dec. 14th BE 838649627 62442
Page	Page: 1	

**Data field Invoice summary table, page 2**

Selector	Property	Value
Table	First column: 1, last column: 4	

# Step 8 : Extract the Data

The file has been uploaded and parsed successfully.

[Download recognized file](#) [Provide feedback](#)

Invoice summary table  
Page 1

Value	Screenshot	Correct?
VAT duty 21% € 900.45 € 1089.54		<input checked="" type="checkbox"/>

Invoice address  
Page 1

Value	Screenshot	Correct?
Levi Corp.		<input checked="" type="checkbox"/>
To the att. of Molly Weasley		<input checked="" type="checkbox"/>
Ottery St. Catchpole 2600		<input checked="" type="checkbox"/>
Devon		<input checked="" type="checkbox"/>

End user address  
Page 1

Value	Screenshot	Correct?
Levi Corp.		<input checked="" type="checkbox"/>
To the att. of Molly Weasley		<input checked="" type="checkbox"/>
Ottery St. Catchpole 2600		<input checked="" type="checkbox"/>
Devon		<input checked="" type="checkbox"/>

# XML Data

```
<?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/>