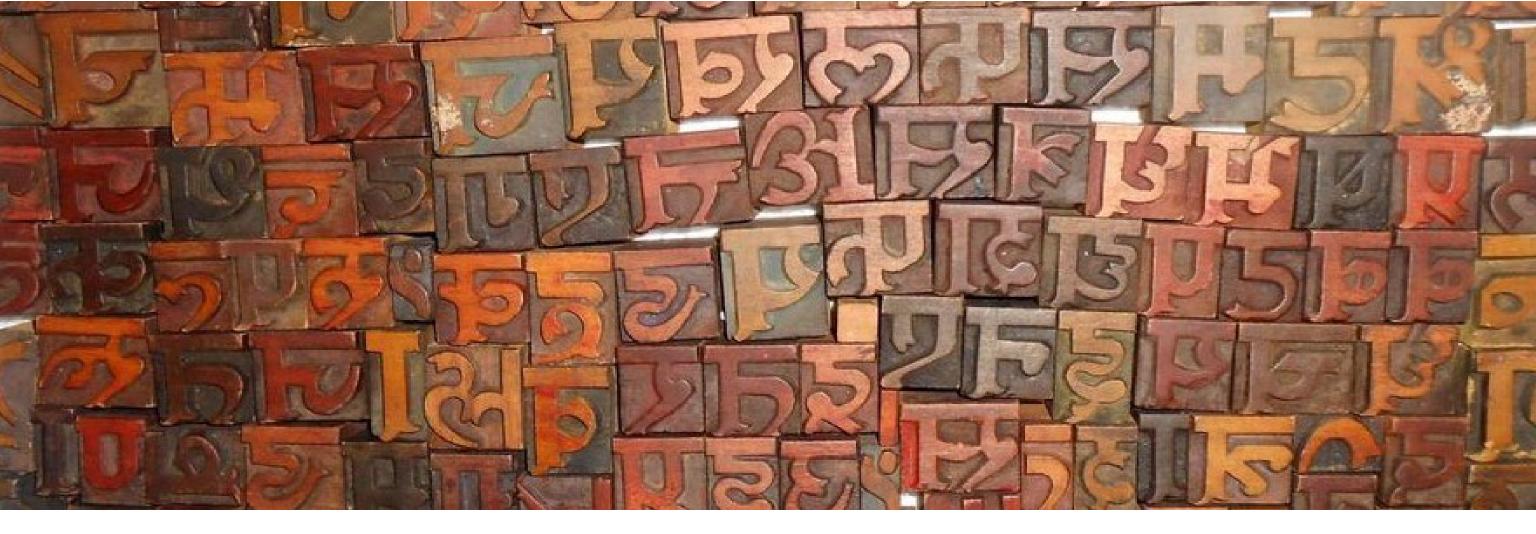


# Support of complex scripts in PDF

Lego game of text composition and text extraction algorithms

Alexey Subach iText Group NV



မင်္ဂလာညနေခင်းပါ

مساء الخير

Добрый вечер

सुसंध्या

ערב טוב

Good evening

អរុណស្ចស្ដី

สวัสดีตอนเย็น

Labas vakaras

God kväll



Buonasera



## Complex scripts are complex!

To get this:

تَعليمُ اللغة العَربية لغيرِ النَّاطقِين بهَا

Instead of that:

الميل الميال الم

...you have got a long way to go



# Steps in processing complex scripts

- Cluster characters into syllables or words
- Apply Unicode character reordering
- Apply Open Type features (shaping)
- = Apply *layout-related* processing, e.g. hyphenation
- Apply Unicode Bidi algorithm
- Write content using correct syntax (to view and extract data)
- Create tag structure
- = Extract content back from PDF to Unicode



# Steps in processing complex scripts

- Cluster characters into syllables or words
- Apply Unicode character reordering
- = Apply Open Type features (shaping)
- = Apply *layout-related* processing, e.g. hyphenation
- **=** Apply Unicode Bidi algorithm
- Write content using correct syntax (to view and extract data)
- Create tag structure
- **Extract content back from PDF to Unicode**



# Example: Arabic text

أهب ن يقط أن اريغ ل قيب رَعل الفغل الميل عَت تعليمُ اللغة العربية لغير النّاطقين بها تعليم اللغة العربية لغير التاطقين بها تعليم اللغة العربية لغير التاطقين بها



## Example: Khmer text

បាដ្ឋេឋាវេចនាសម្ព័ព័ន្ធធុរូវការ ហេដ្ឋារចនាសម្ព័ន្ធត្រាវារារ ហេដ្ឋារចនាសម្ព័ន្ធត្រាវការ



Logical representation

Visual representation



**Unicode Bidirectional Algorithm** 



■ Mixed directions in text: numbers and English text are still shown left-to-right

# October PDFest 2020 שלום



■ Mixed directions in text: numbers and English text are still shown left-to-right

# October PDFest 2020 שלום

**■ /ReversedChars:** informs the PDF processor that *show strings* within a marked-content sequence *contain characters in reverse order*. In order for such text to be extracted or read out aloud, the sequence of the characters as found in the show string operator shall be reversed before using them



= Character mirroring

שלום עולם (כל היבשות) היבשות)

= Unicode:

\u05E9...\u05DD (\u05DB\u05DC ...\u05EA)

= /ToUnicode – does not work if there are both mirrored and regular brackets

≡ Using multiple fonts – bugs when editing the PDF (e.g. form fields)



= Character mirroring

שלום עולם (כל היבשות) היבשות)

= Unicode:

\u05E9...\u05DD (\u05DB\u05DC ...\u05EA)

- = /ToUnicode does not work if there are both mirrored and regular brackets
- ≡ Using multiple fonts bugs when editing the PDF (e.g. form fields)
- = /ActualText



#### Is /ActualText our savior?

- Can be specified for content that does translate into text but that is represented in a nonstandard way (ISO 32000-2)
- = Replacement text can be specified for the following items:
  - = A structure element, by means of the optional **ActualText** entry (PDF 1.4) of the structure element dictionary.
  - = (PDF 1.5) A marked-content sequence, through an **ActualText** entry in a property list attached to the marked-content sequence with a **Span** tag.



#### Example

```
BT
/F2 30 Tf
91.75 700.03 Td
/Span <</ActualText <feff0028> >> BDC # Mirroring the bracket for extraction
<000c> Tj
                                         # ")" is shown, "(" is extracted
EMC
ET
BT
/F1 30 Tf
108.75 700.03 Td
/ReversedChars BMC
<001b001a00130020> Tj
EMC
ET
```



#### RTL Text Extraction: In practice

- ≡ If you don't use /ReversedChars: many major PDF viewers/consumers will still understand that this is RTL word
- ≡ If you use /ReversedChars: behavior of **many** PDF viewers/consumers is exactly the same as if you did not
- If you don't use /ActualText to mirror the paired characters: **some** PDF consumers will still apply the inverse Bidi algorithm and extract the correct logical sequence
- If you use /ReversedChars with /ActualText for mirroring, very few PDF consumers will understand the intention when performing text extraction and typically brackets will be mirrored even though they were not intended to



#### RTL text: correct tagging

Econtent shall be marked in the structure tree with semantically appropriate tags in a logical reading order (ISO 14289)

- Document
  - o P
    - Span
    - Span





## Inverse Bidi algorithm

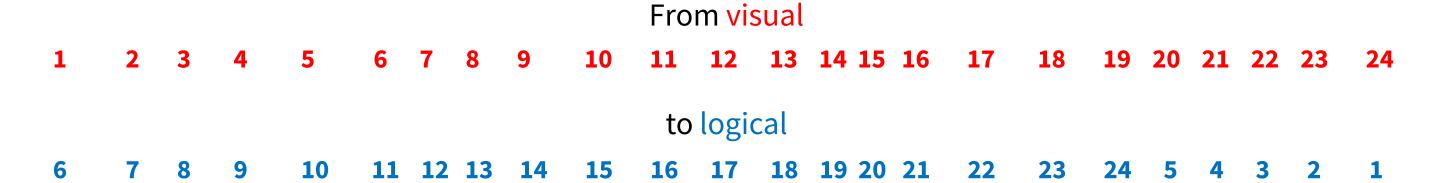
- = Restores logical sequence from visual one
- Needed in case /ReversedChars is missing
- Not necessarily whole text chunk in Tj, glyphs can be in random order which requires sorting in the first place
- Even for non-random order you need to detect text orientation for independent glyphs if they are in separate text showing operators
- = Many-to-one function, results may be unexpected



### Inverse Bidi algorithm

Sort the glyphs by their position and feed them to the algorithm!

# October PDFest 2020 שלום



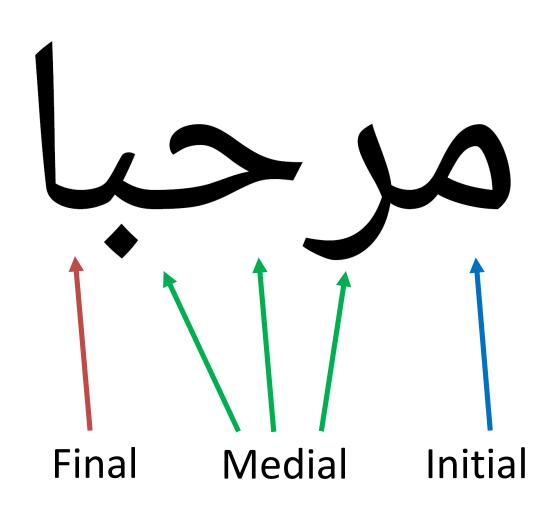


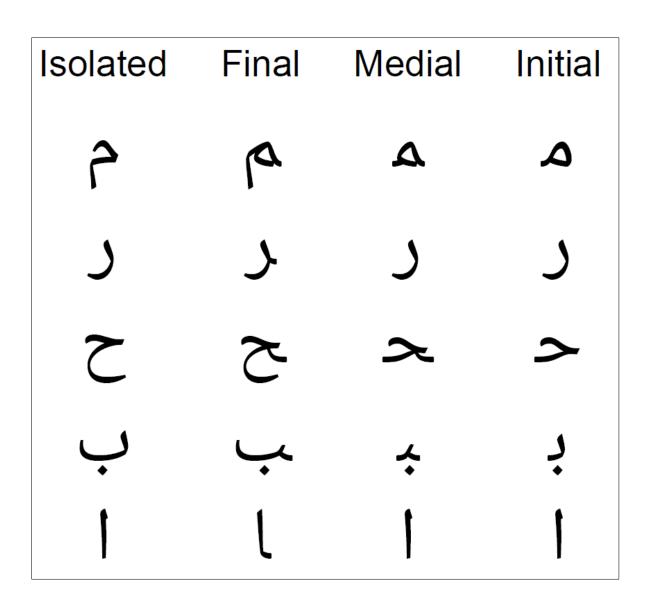
## On the way to success

اهب نيقطانالاريغل قيبرَعلا قغللا ميلغت الميم اللغة العربية لغير الناطقين بها تعليم اللغة العربية لغير التاطقين بها تعليم اللغة العربية لغير التاطقين بها



# Arabic shaping: 4 glyph forms

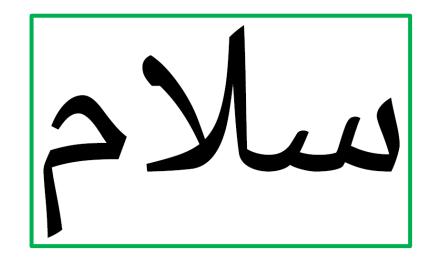


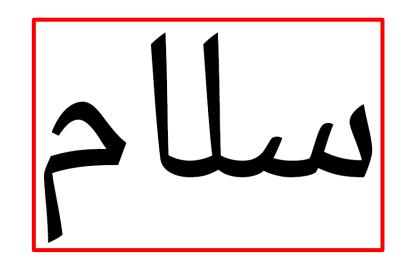




## OpenType Implementation

- OpenType fonts contain tables with the relevant information: init, medi, fina features in GSUB table
- = Fonts must be of good quality as they provide vital information for shaping
- = Do not forget about **rlig** (Required Ligatures) feature







- = Some operations can be done purely on Unicode level, e.g. Arabic shaping
- = Extensive Unicode information is a requirement in that case (1442 code points)
  - = Arabic Presentation Forms-A (FB50–FDFF, 611 characters)
  - = Arabic Presentation Forms-B (FE70–FEFF, 141 characters)
- Default (isolated), initial, medial, final, forms of glyphs require 4x characters in Unicode



- = Some operations can be done purely on Unicode level, e.g. Arabic shaping
- = Extensive Unicode information is a requirement in that case (1442 code points)
  - = Arabic Presentation Forms-A (FB50–FDFF, 611 characters)
  - = Arabic Presentation Forms-B (FE70–FEFF, 141 characters)
- Default (isolated), initial, medial, final, forms of glyphs require 4x characters in Unicode
- Arabic Letter Khah Medial Form (U+FEA8)



- = Some operations can be done purely on Unicode level, e.g. Arabic shaping
- = Extensive Unicode information is a requirement in that case (1442 code points)
  - = Arabic Presentation Forms-A (FB50–FDFF, 611 characters)
  - = Arabic Presentation Forms-B (FE70–FEFF, 141 characters)
- Default (isolated), initial, medial, final, forms of glyphs require 4x characters in Unicode
- = Arabic Letter Khah Medial Form (U+FEA8) 🗻
- = Arabic Ligature Salam Isolated Form (U+FDF5)



- = Some operations can be done purely on Unicode level, e.g. Arabic shaping
- = Extensive Unicode information is a requirement in that case (1442 code points)
  - = Arabic Presentation Forms-A (FB50–FDFF, 611 characters)
  - = Arabic Presentation Forms-B (FE70–FEFF, 141 characters)
- Default (isolated), initial, medial, final, forms of glyphs require 4x characters in Unicode
- = Arabic Letter Khah Medial Form (U+FEA8) 🗻
- = Arabic Ligature Salam Isolated Form (U+FDF5)
- = Arabic Ligature Bismillah Ar-Rahman Ar-Rageem (U+FDFD) الرحيمان



#### Not for all complex scripts

- **Substitution types** 
  - One to one (Arabic shaping)
  - Many to one (ligatures)
  - Many to many (incl. context-aware)
  - One to many (decomposition)
- ■691 substitutions for Arabic in a font compared to 1270 substitutions for a Devanagari font while Devanagari only has 208 Unicode characters reserved
- Impractical to store all the combinations at the Unicode level, contextual substitutions make it even harder, font lookups are more flexible

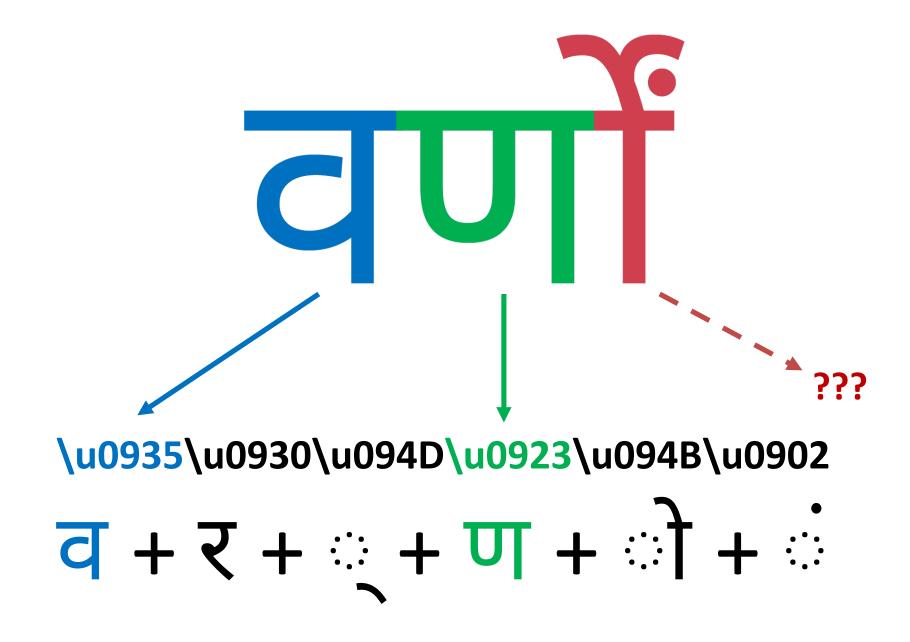


#### PDF Implementation

- ≡ Easy for initial, medial, final forms just use /ToUnicode to map to the default form
- No Unicode mapping for some of the glyphs from more complex substitutions (e.g. Devanagari)
- Reordering at Unicode level for clusters before applying font features makes ligature-like /ToUnicode impossible
- ≡ Syllable (cluster) is a minimal unit that can guarantee correspondence between visual representation and Unicode sequence that you get on copy-paste



#### Text extraction: getting trickier





#### /ActualText is our savior!

- = Turns out in many cases it has to be specified
- You should use the minimal possible sequence in the /ActualText block to allow copying small chunks of text

```
/Span <</ActualText <FEFF 0930 094D 0923 094B 0902> >> BDC <002701C4>Tj
EMC
```



## /ActualText: In practice

- Not supported in many PDF viewers
- Problems with determining spaces when extracting text in some viewers that support /ActualText
- Problems with extracting diacritical marks when extracting text in some viewers that support /ActualText
- If each of two (or more) consecutive structure or marked-content sequences has an **ActualText** entry, they shall be treated as if no word break is present between them (ISO 32000-2)

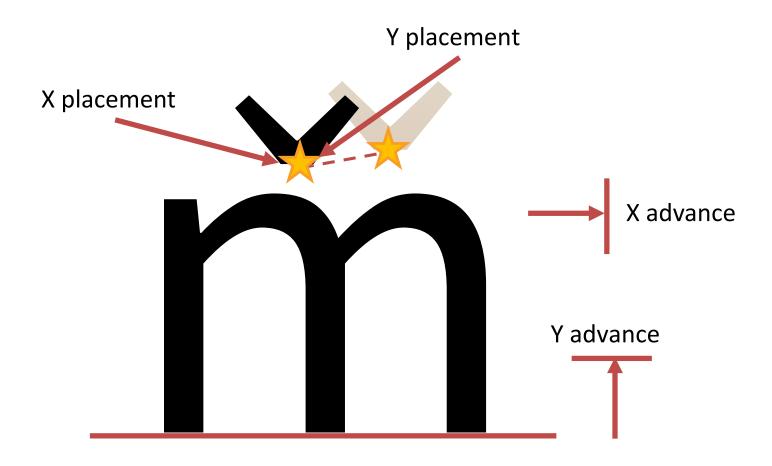


## On the way to success

اهب نيقطانالاريغل قيبرَعلا قغللا ميلغت تعليمُ اللغة العَربية لغير النّاطقين بها تعليم اللغة العَربية لغير التاطقين بها تعليم اللغة العربية لغير التاط



# Attaching glyph to one another

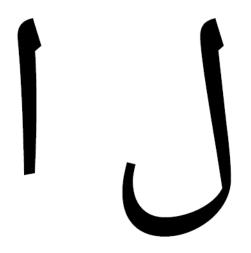




- Several possible attachment components force us into tracking which ligature component the mark is associated with
- = /ActualText plays vital role in extracting the text correctly

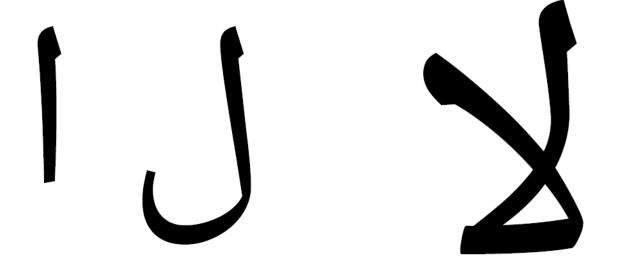


- Several possible attachment components force us into tracking which ligature component the mark is associated with
- = /ActualText plays vital role in extracting the text correctly



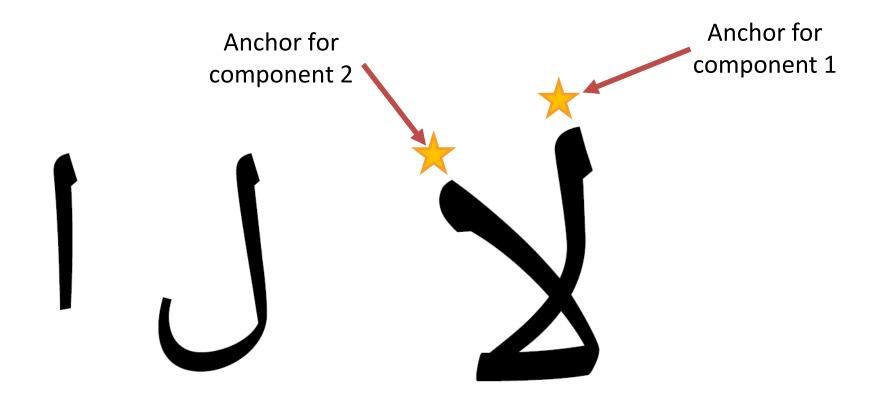


- Several possible attachment components force us into tracking which ligature component the mark is associated with
- = /ActualText plays vital role in extracting the text correctly





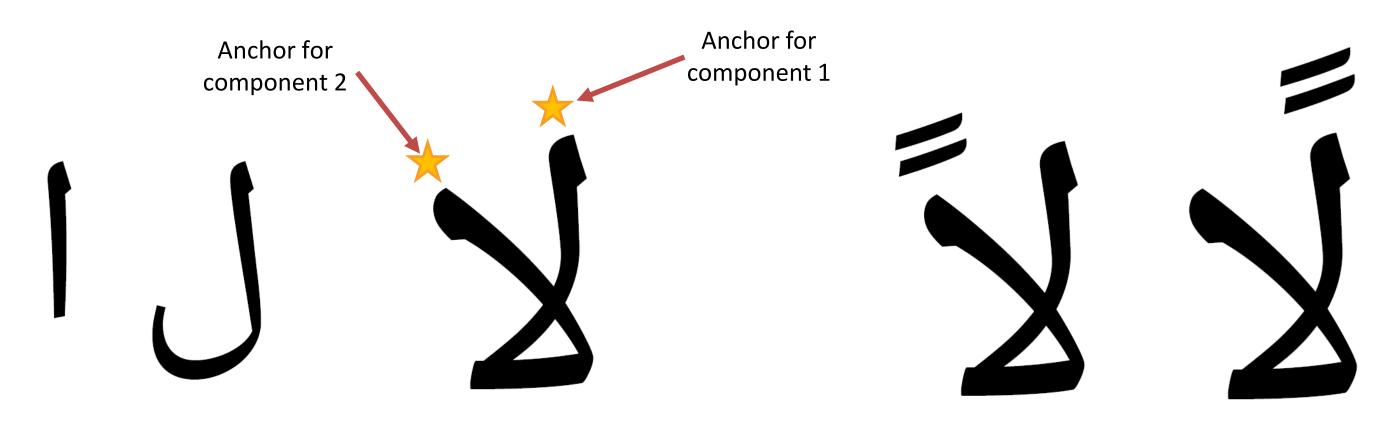
- Several possible attachment components force us into tracking which ligature component the mark is associated with
- = /ActualText plays vital role in extracting the text correctly





## Mark to ligature attachment

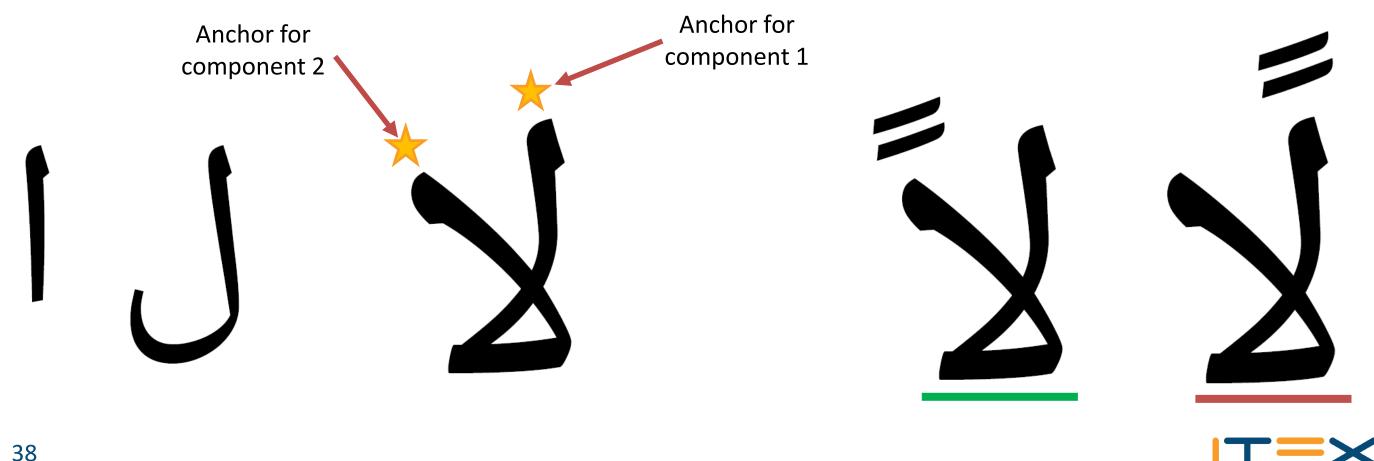
- Several possible attachment components force us into tracking which ligature component the mark is associated with
- = /ActualText plays vital role in extracting the text correctly





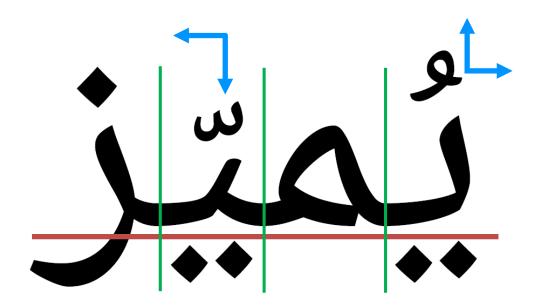
## Mark to ligature attachment

- = Several possible attachment components force us into tracking which ligature component the mark is associated with
- = /ActualText plays vital role in extracting the text correctly



## Extracting base glyphs + marks

- ≡ It is a common practice to perform some glyph sorting as a part of text extraction pipeline
- ≡ Glyph positioning: offsets and attachments affect the sorted order of the glyphs, both along X and Y axes, both mark to base and mark to mark
- = Sorting needs to be adjusted to take that into account





#### Conclusions

- All the pieces of the puzzle must fall together for the complex scripts
  to be supported (assuming the font is of good quality already ②)
- Poorly-generated documents provide a very limited possibility for extracting data from them in automatic way
- Not supporting complex scripts in products we make creates interoperability problems and might slow down adoption of PDF as the *universal* document interchange format in countries natively using complex scripts



### References

- PDF and OpenType, PDF Days 2016 talk https://www.youtube.com/watch?v=qZnXZppH2KI
- OpenType specification <a href="https://www.microsoft.com/en-us/Typography/OpenTypeSpecification.aspx">https://www.microsoft.com/en-us/Typography/OpenTypeSpecification.aspx</a>
- Microsoft Typography <a href="https://www.microsoft.com/en-us/Typography/default.aspx">https://www.microsoft.com/en-us/Typography/default.aspx</a>
- = FontForge Open Source tool <a href="https://fontforge.org">https://fontforge.org</a>
- OpenType CookBook <a href="http://opentypecookbook.com">http://opentypecookbook.com</a>





# Solutions offered by iText

- Leading PDF technology company
- Board member of PDF Association
- Open Source SDK is available for 20 years
- = 125M users worldwide
- = 7000+ clients worldwide





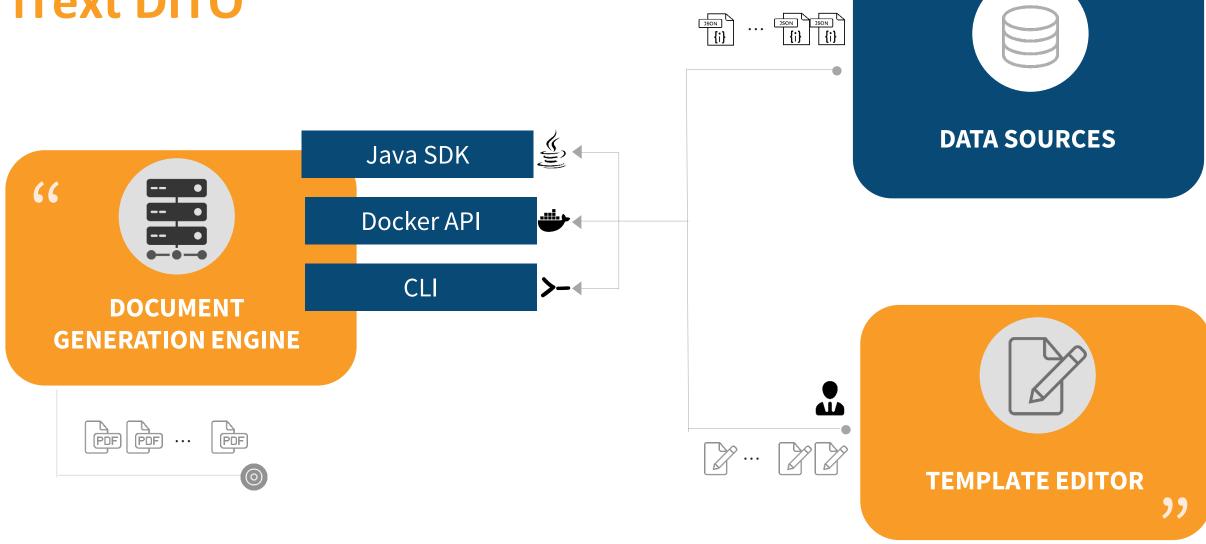
- Process PDF documents in Java or .NET
- iText 7 add-on: pdfHTML
  Convert HTML and CSS into
  PDF documents
- iText 7 add-on: pdfSweep
  Securely redact data from PDF
  documents

- j iT
  - iText 7 add-on: pdfXFA
    Flatten dynamic XFA forms to
    static PDF
- iText 7 add-on: pdfCalligraph Unlock advanced features in PDF typography
- iText 7 add-on: pdfOCR
  Convert images into searchable
  PDF documents

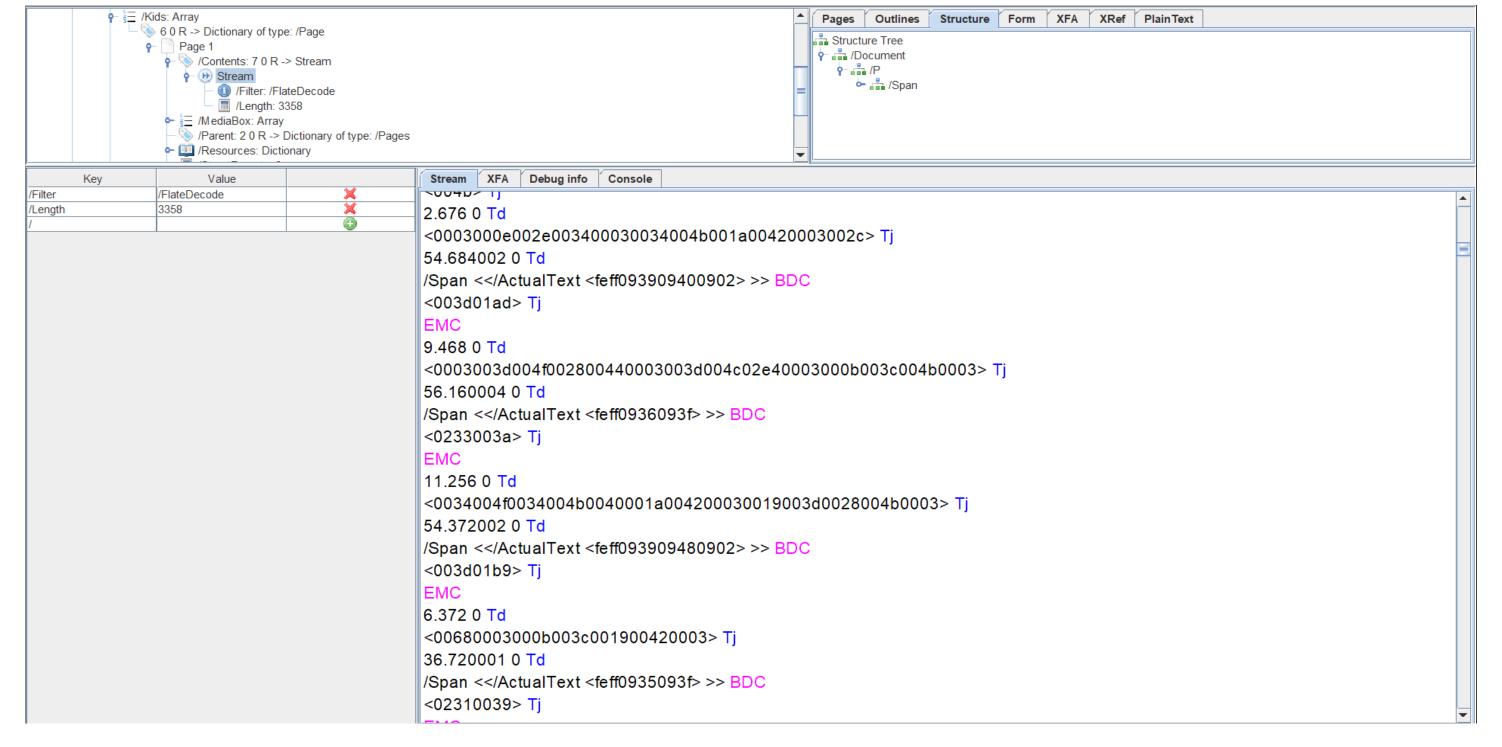
- iText 7 add-on: pdf2Data
  Extract data from templated
  PDF documents
- iText 7 add-on: pdfRender
  Generate images from PDF











\* Screenshot from iText RUPS <a href="https://github.com/itext/i7j-rups">https://github.com/itext/i7j-rups</a>



```
PdfDocument pdfDocument = new PdfDocument(new PdfWriter(PDF PATH));
Document document = new Document(pdfDocument);
document.add(new Paragraph("Good evening to OctoberPDFest participants!"));
Table greetingsTable = new Table(2)
     .setBackgroundColor(ColorConstants.LIGHT GRAY);
GREETINGS.forEach((language, text) ->
     greetingsTable.addCell(language).addCell(text));
document.add(greetingsTable.setHorizontalAlignment(HorizontalAlignment.CENTER));
document.close();
```



#### Good evening to OctoberPDFest participants!

French	Bonsoir
Italian	Buonasera
German	Guten Abend
Lithuanian	Labas vakaras



```
Collection<String> FONTS = Arrays.asList(ARABIC_FONT_PATH,

DEVANAGARI_FONT_PATH, THAI_FONT_PATH, KHMER_FONT_PATH);

FontProvider fontProvider = new FontProvider();

FONTS.forEach(fontProvider::addFont);

document.setFontFamily("Noto").setFontProvider(fontProvider);
```

```
FontProvider fontProvider = new FontProvider();
fontProvider.addSystemFonts();
document.setFontFamily("Noto Sans").setFontProvider(fontProvider);
```



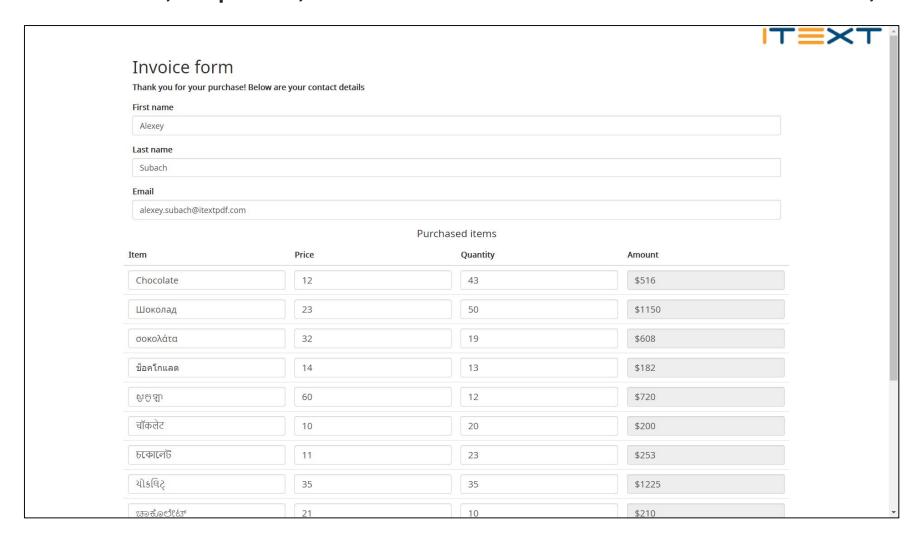
#### Good evening to OctoberPDFest participants!

Devanagari	सुसंध्या
Thai	สวัสดีตอนเย็น
Khmer	អរុណសូស្តី
Arabic	مساء الخير



## pdfHTML

- = HTML as more high-level syntax
- = Invoices, reports, contracts and other data from CRM, ERP systems





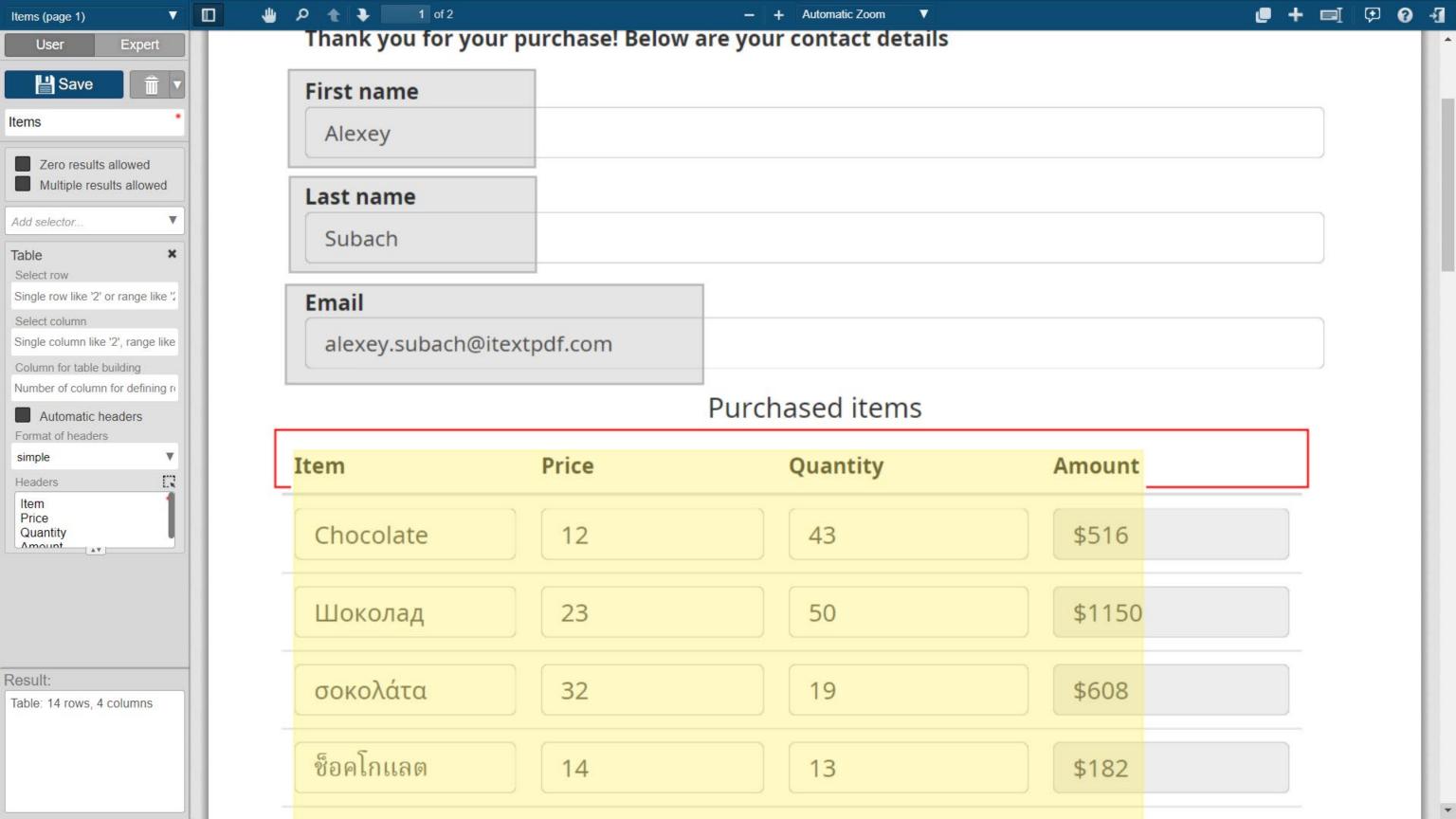
## pdfHTML: Code sample

HtmlConverter.convertToPdf(new File(HTML\_FILE\_PATH), new File(OUT\_FILE\_PATH));

- Single line of code
- = Fonts are loaded automatically
- Integrated support of complex scripts

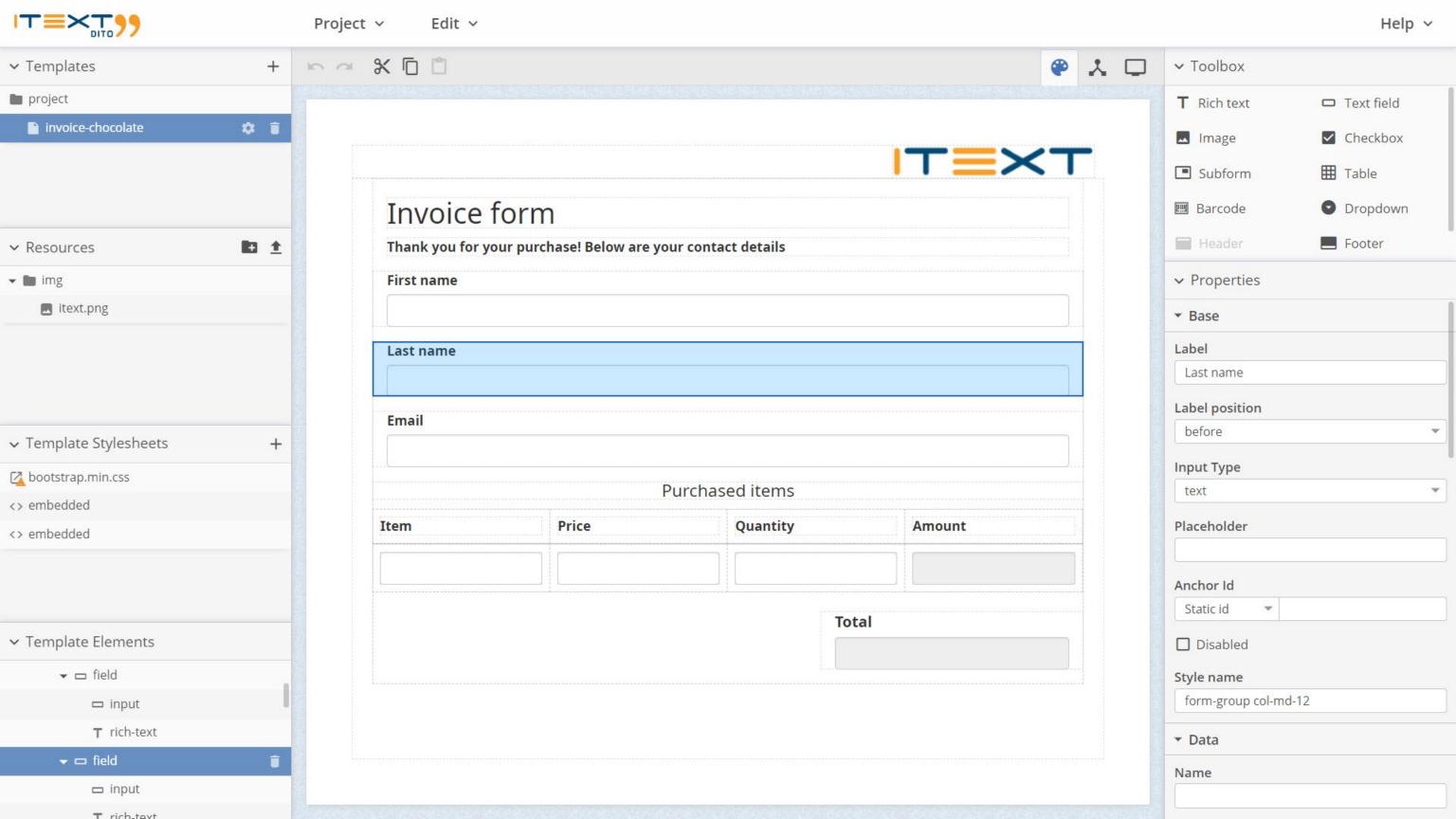


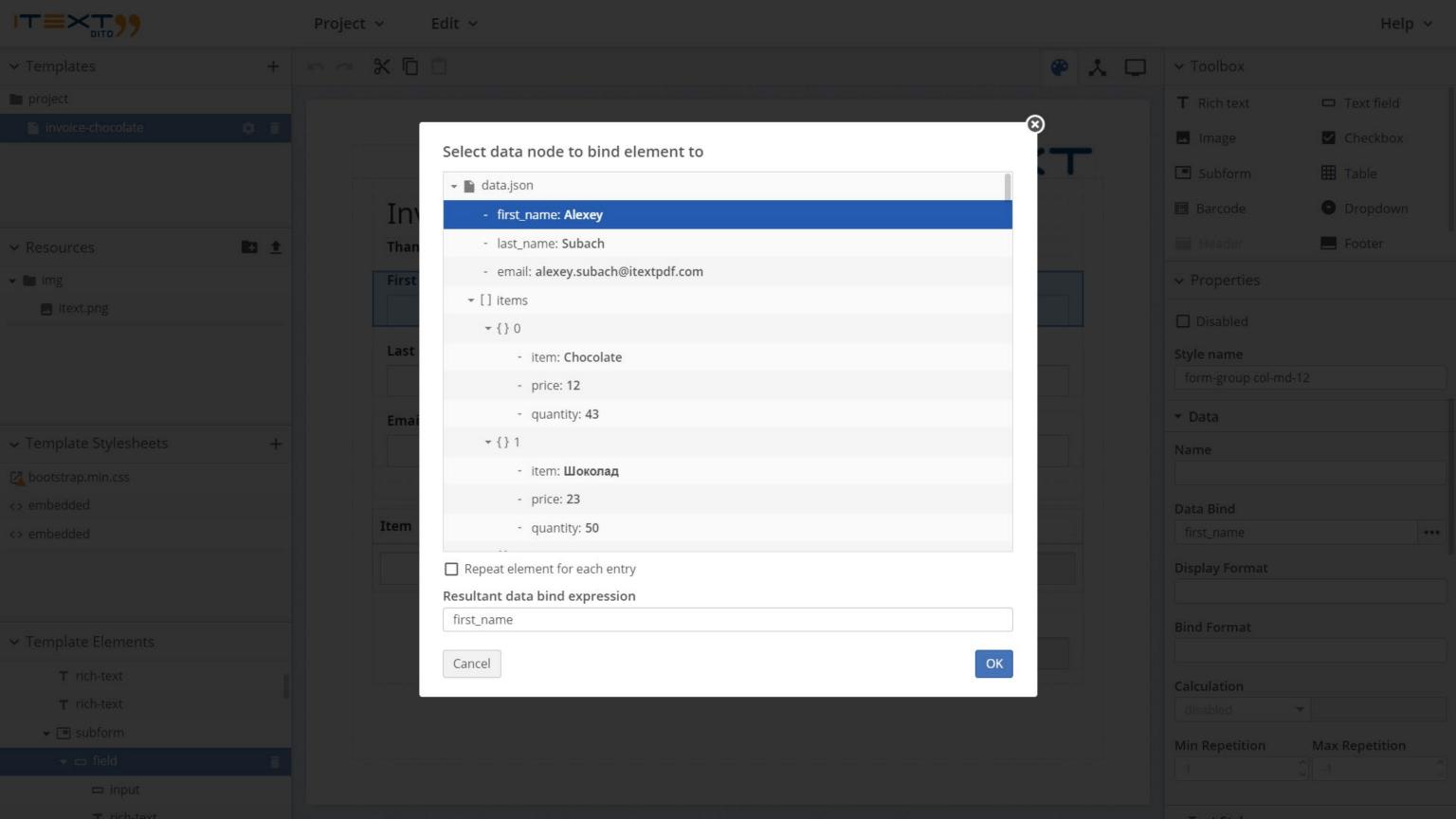




#### Data field Items, page 1

Selector	Property	Value			
Table	Headers: Item	ltem	Price	Quantity	Amount
	Price	Chocolate	12	43	\$516
	Quantity Amount	Шоколад	23	50	\$1150
		σοκολάτα	32	19	\$608
		ช็อคโกแลต	14	13	\$182
		សូកូឡា	60	12	\$720
		चॉकलेट	10	20	\$200
		চকোলেট	11	23	\$253
		યોકલિટ્	35	35	\$1225
		ಚಾಕೊಲೇಟ್	21	10	\$210
		ചോക്ലേറ്റ്	12	31	\$372
		ଟକୋଲେଟ୍	7	29	\$203
		சாக்லேட்	7	29	\$203



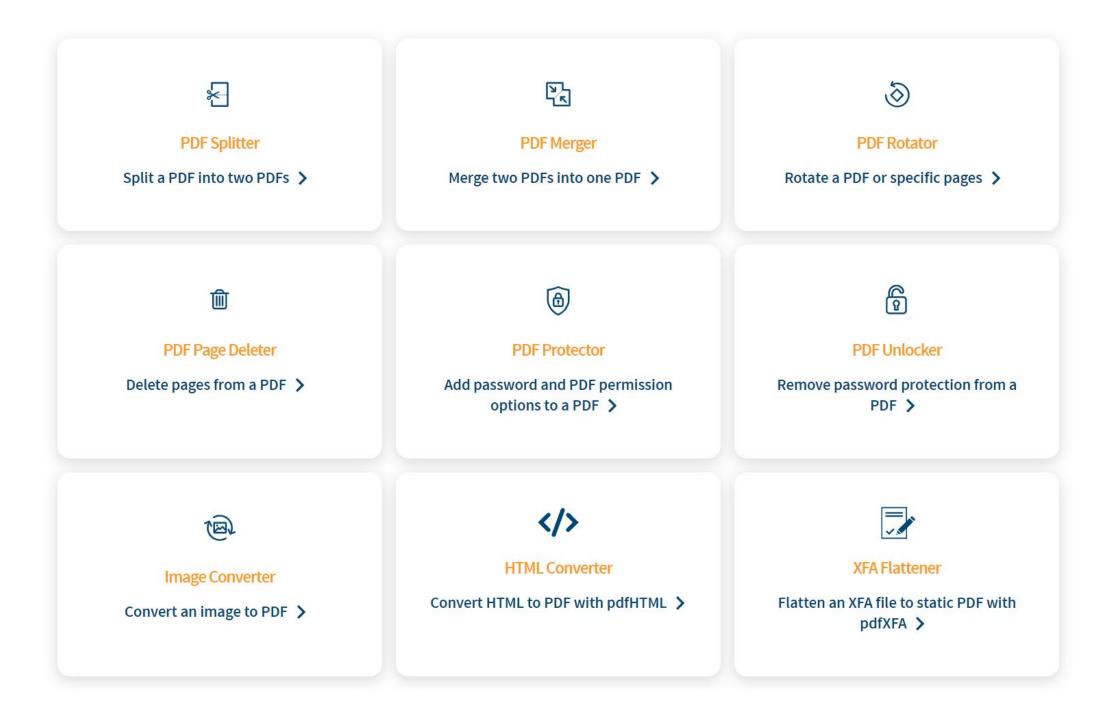


### iText DITO: Code sample



```
POST {server-address}/api/pdf-producer
Content-Type application/json
Body
     "templateProjectPath": "invoice.dito",
     "templateName": "invoice-chocolate",
    "data": {
         "first name": "Alexey",
         "last name": "Subach",
         "email": "alexey.subach@itextpdf.com",
         "items": [
     "path": "custom/path/to/file.pdf",
     "properties": { "pdfVersion": "2.0" }
```





https://itextpdf.com/demos



#### **TELEPHONE**

#### Page 1

Value	Screenshot	Correct?
+1 617 982 2646	+1 617 982 2646	ОК

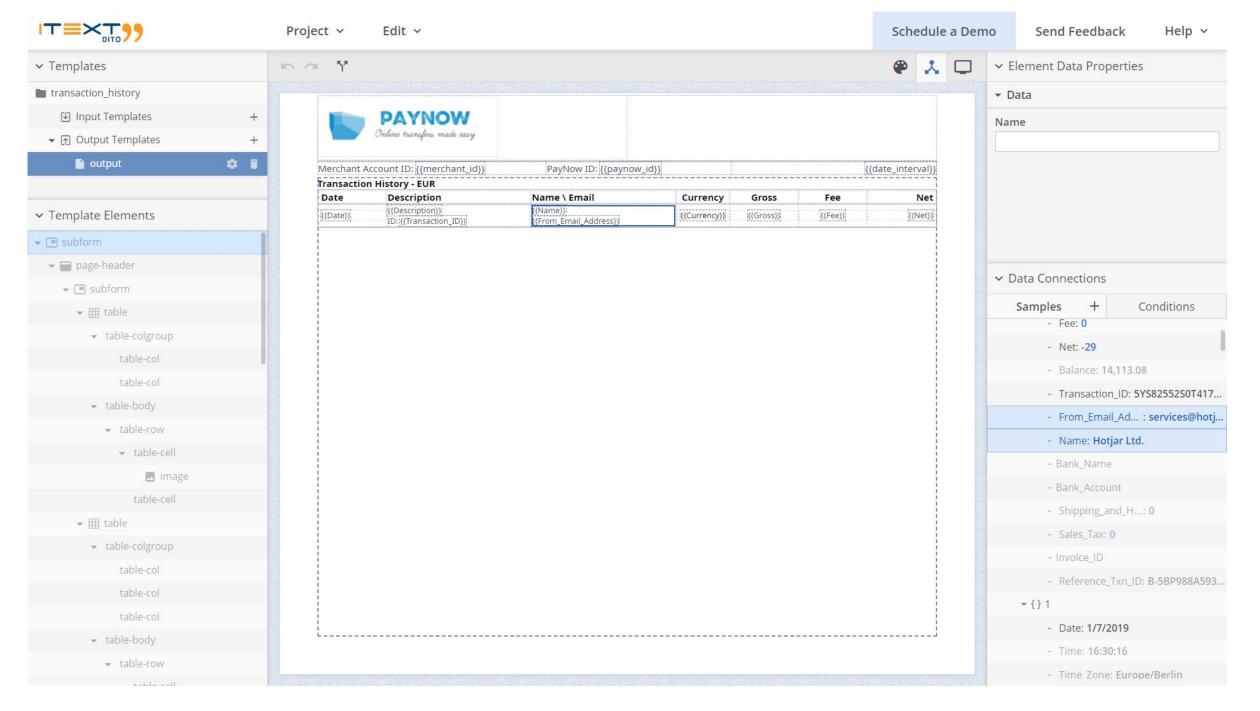
#### **TABLE**

#### Page 1

/alue			Screenshot			
Description	Quantity	Unit price	Discount	Total		C
		USD	%	USD	Machinal and an assessment on the control of the co	
Commercial Invisibility Cloak License .NET NO-CAN-SEE-1 Perpetual License - Unlimited use in Time - Unlimited Users	1.00	\$6,800.00	0.00 %	\$6,800.00	Value of Control Con	
Updates and Support Commercial Invisibility Cloak .NET NO-CAN-FAIL-1 Technical Support and Updates during 12 Months. Renewed on an Annual basis.	1.00 Year(s)	\$666.00	0.00 %	\$666.00		
Commercial Non-Production House Elf DBBY-OR-WNKY Perpetual License - Ready for Development - Testing - QA - Training - Disaster Recovery (Passive)	1.00 Server(s)	\$3,140.00	0.00 %	\$3,140.00		
Updates and Support Commercial Non- Production House Elf THE-CANE Technical Support and Updates during 12 Months	1.00 Year(s)	\$152.00	0.00 %	\$152.00		

https://pdf2data.online





https://itextdito.online





# Thank you! Questions?

alexey.subach@itextpdf.com

https://itextpdf.com

