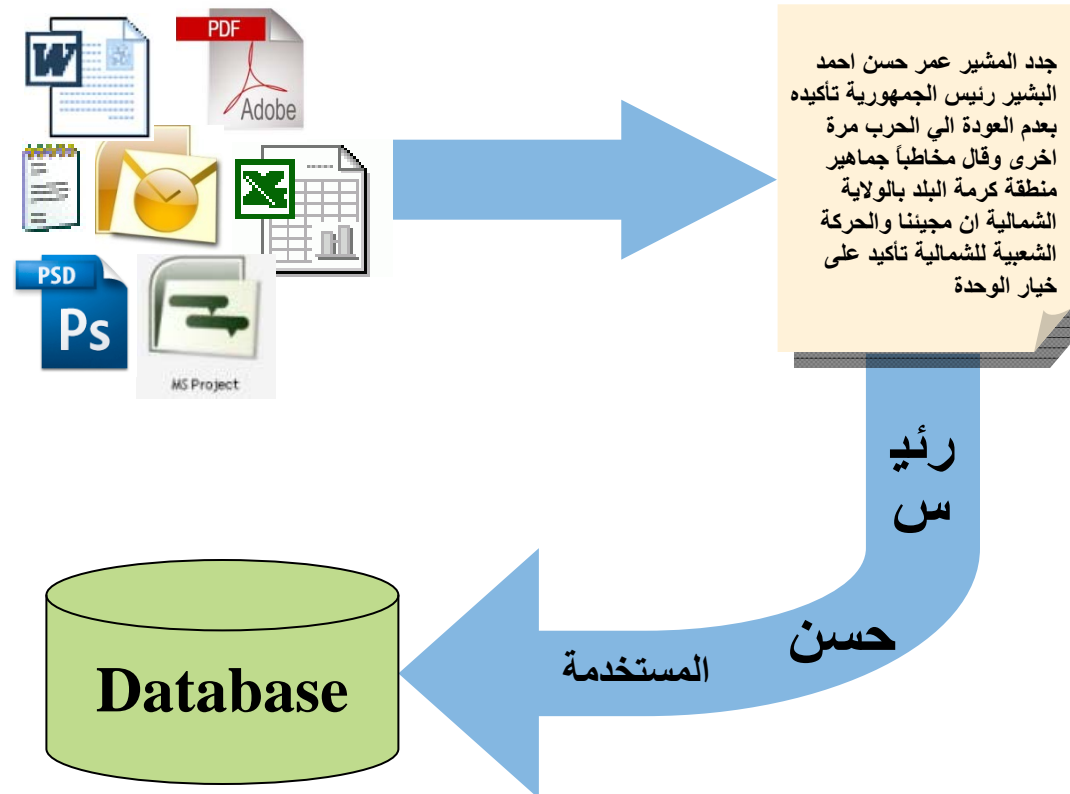


Motivation

- Need to get the text out of files before they can be indexed and searched.
- Arabic PDF files can be challenging.



PDF Basics

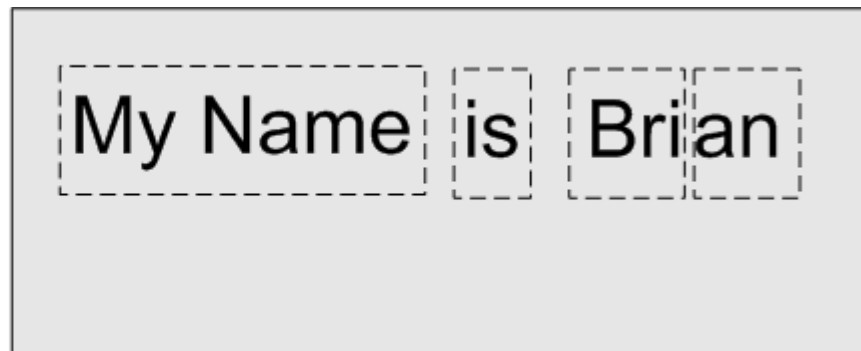
- Raw file contents are organized into objects.
- Each object stores a specific type of info:
 - Document (Root) object
 - Page objects
 - Font objects
- Basic structure of file is viewable text:

```
[...]  
7 0 obj  
<</Metadata 4 0 R/Pages 3 0 R/Type/Catalog/PageLabels 1 0 R>>  
endobj  
[...]
```

PDF Text

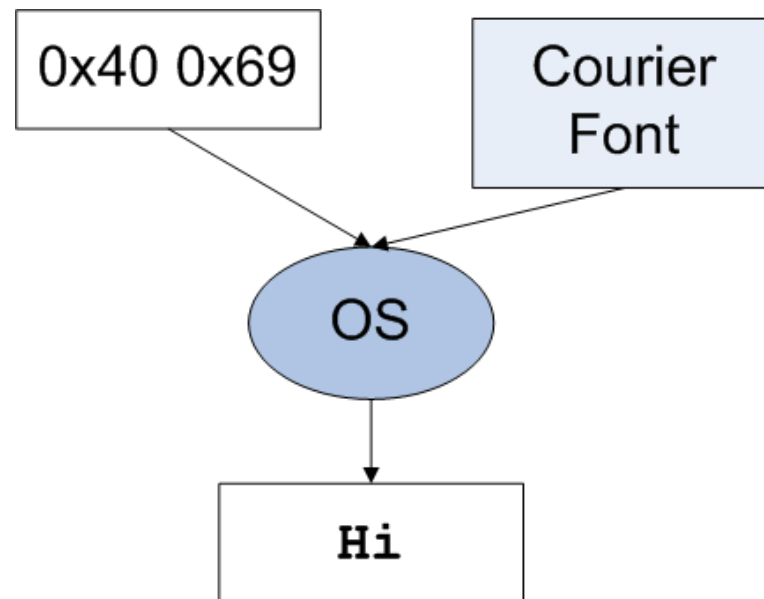
- Text is stored in chunks of one or more characters.
- Each chunk is located at a given X,Y coordinate
- Chunks can be stored in any order in the file

(48,20) Bri
(58,20) an
(10,20) My Name
(40,20) is



Typical Encoding and Rendering

- Files store text in an encoding:
 - ISO-8859-6 maps a 1 byte value to a Latin or Arabic character
 - Unicode maps values to characters in many languages
- The OS uses the encoding value and a specific font to find the correct glyph to display:



PDF Fonts and Encodings

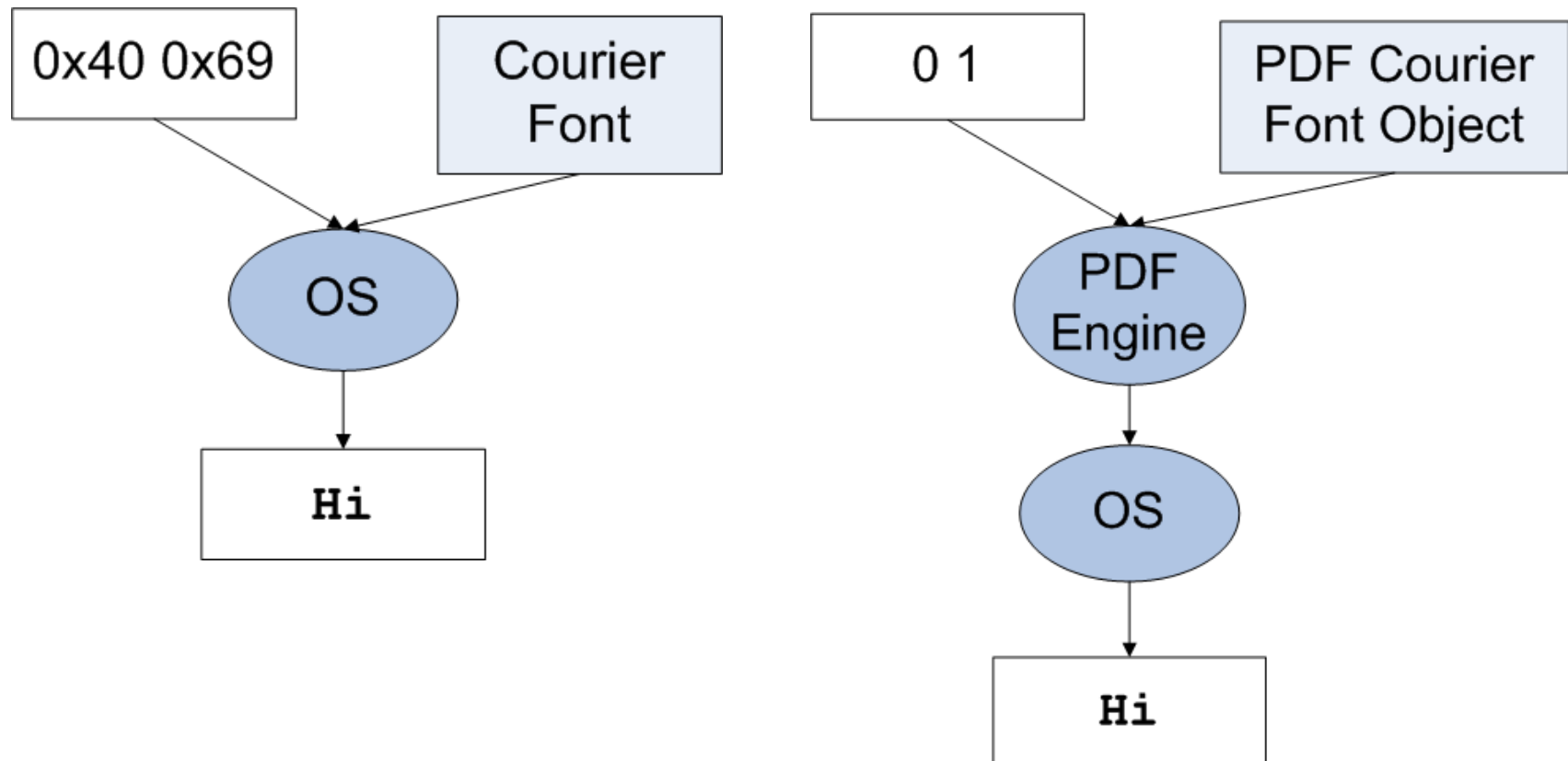
- PDF fonts typically store only the glyphs that are used.
- Text chunk stores an index into a PDF font object.
- Font object may map glyph to a Unicode value.

PDF Courier
Font Object

H	U+0040
i	U+0069

Rendering Difference

- Displaying a PDF requires the PDF Engine to map fonts
- Note that standard encoding values are not required.

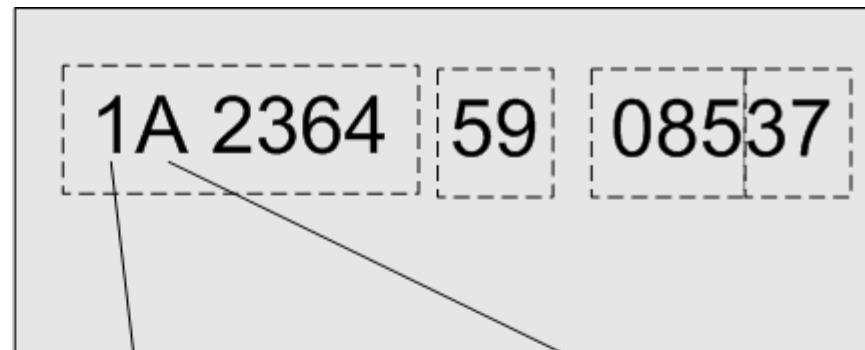


Basic Extraction Approach

1. Parse PDF file to identify page content objects
2. Parse page content stream into text chunks
3. Sort text chunks based on coordinates
4. Process chunks in order:
 1. Get index for each character
 2. Use font information to map index to Unicode (if defined)
 3. Add Unicode value to end of string

English Extraction Example

(48,20) 085
(58,20) 37
(10,20) 1A 2364
(40,20) 59



0	1	2	3	4	5	6	7	8	9	A
B	M	N	a	e	i	m	n	r	s	y

My Name is Brian

Arabic Glyphs

- Arabic characters have different shapes depending on their location in a word.

General: م

Isolated: م

Final: م

Medial: م

Initial: م

- Each shape is a different glyph in a font.

Arabic Extraction Example

(10,20) 0121
(40,20) 3456



0	1	2	3	4	5	6
د	م	ح	ي	م	س	إ

د	م	ح	م		ي	م	س	إ
---	---	---	---	--	---	---	---	---

Original:
إسمي محمد

Displayed:
دمحم يمسيا

Logical and Presentation Orders

- Text in computers is typically stored in logical order
 - First character stored is first character read or written
- Presentation order is based on screen layout
- Orders are same for Left to Right (LTR) Languages:

M	y		N	a	m	e		i	s		B	r	i	a	n
---	---	--	---	---	---	---	--	---	---	--	---	---	---	---	---

My Name is Brian

- Opposite for Right to Left (RTL) Languages:

د	م	ح	م		ي	م	س	!
---	---	---	---	--	---	---	---	---

إسمي محمد

Possible Order Solution

- PDF stores data in presentation (display) order.
- Text editors need the text in logical order though.
- Need to convert from presentation to logical order.
- Obvious solution:
 - After decoding each line, reverse the order of the Arabic text:

Got:

د	م	ح	م		ي	م	س	إ
---	---	---	---	--	---	---	---	---

Want:

إ	س	م	ي		م	ح	م	د
---	---	---	---	--	---	---	---	---

Bi-directional Text

- How should the following be logically stored?

إسمي محمد .2009

A) 2 0 0 9 . إ س م ي م ح م د

2009. إسمي محمد

B) إ س م ي م ح م د . 9 0 0 2

إسمي محمد 9002.

C) إ س م ي م ح م د . 2 0 0 9

إسمي محمد 2009.

D) إ س م ي م ح م د 2 0 0 9 .

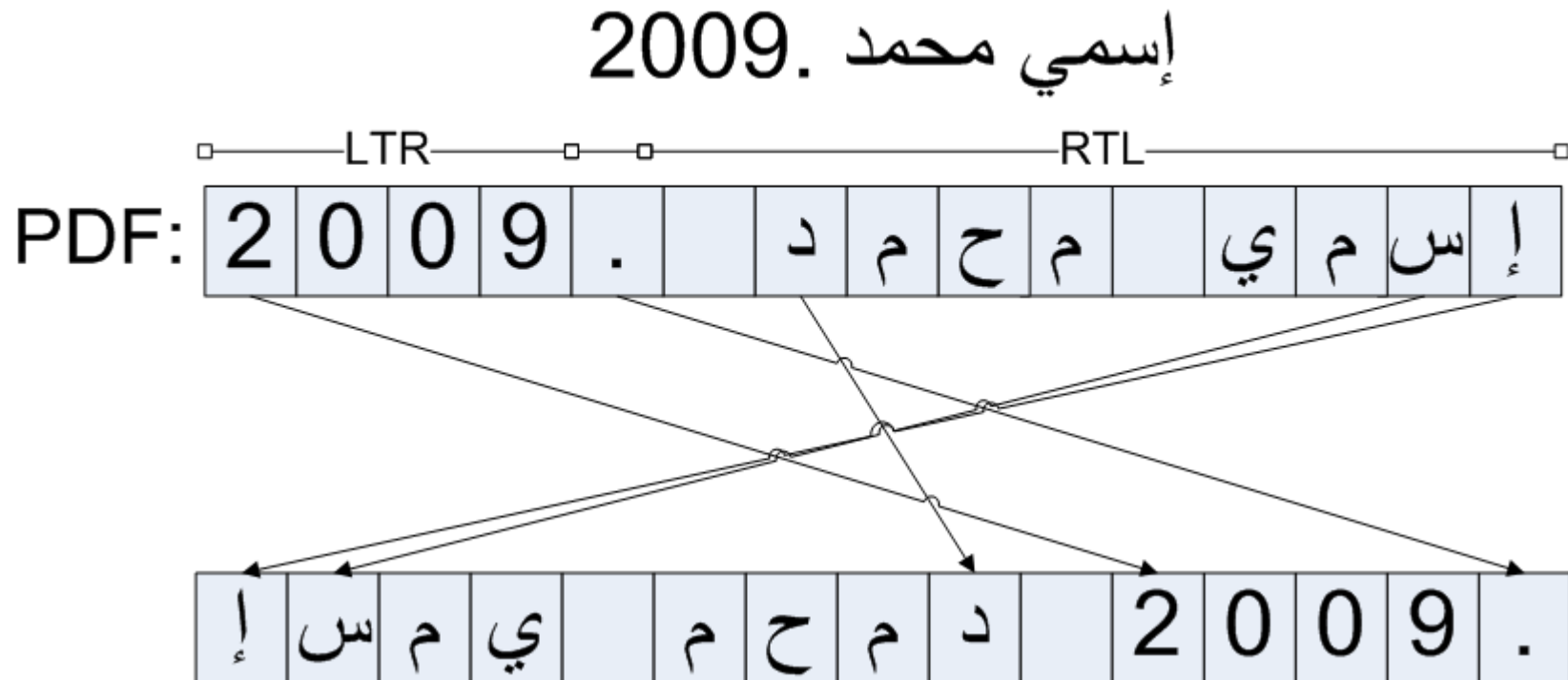
إسمي محمد .2009

Bi-directional Text

- Text can have both RTL and LTR characters and each should go in the correct direction
- Unicode Bi-directional Text (BiDi) algorithm defines how to order characters in a paragraph based on:
 - Dominant direction of text in paragraph
 - Direction of each character in text
 - Punctuation and neighboring characters
 - Implicit direction markers
- BiDi lets you convert from logical to presentation order.

Reverse Bi-directional Algorithm

- We need Reverse BiDi to convert from presentation to logical order.



Updated Extraction Approach

1. Parse PDF file to identify page content objects
2. Parse page content stream into text chunks
3. Sort text chunks based on coordinates
4. Determine dominant text direction
5. Process chunks in order and by line:
 1. Get index for each character
 2. Use font information to map index to Unicode
 3. Add Unicode value to end of "presentation order" string
 4. Apply reverse BiDi algorithm to "presentation order" string

Presentation Forms / Ligatures

- Encodings typically define only the general form of Arabic characters.
 - Unicode is an exception.
- The OS determines which glyph form to use (initial, medial, etc.) based on the context of the character.
- PDF stores the specific form of each Arabic character.
- Unicode presentation forms should not be used in a string and many tools cannot process them.
- Need to normalize text from presentation to general forms

Arabic Extraction Example 2

Original:
إسمي محمد

0123 4567

Font Mapping

0	1	2	3	4	5	6	7
د	م	ح	م	ي	م	س	أ

Presentation Order

د	م	ح	م		ي	م	س	أ
---	---	---	---	--	---	---	---	---

Logical Order

أ	س	م	ي		م	ح	م	د
---	---	---	---	--	---	---	---	---

Normalized

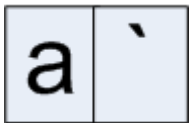
د	م	ح	م		ي	م	س	أ
---	---	---	---	--	---	---	---	---

Font-specific Ligature Implementations

- U+FDFA is the Unicode Arabic ligature for Allah (الله).
- The single ligature represents four characters:
 - “Alef, Lam, Lam, Heh”.
- Some fonts implement the ligature differently:
 - “Lam, Lam, Heh”
- They add a separate “Alef” before the ligature.
 - Alef (U+0627) Allah(U+FDFA)
- When decomposing using Unicode specs:
 - “Alef Alef Lam Lam Heh”

Diacritic Placement

- Vocalizations and diacritics can be separate glyphs
- With Unicode:
 - Diacritics are stored after the base character in logical order



- Diacritics are placed over the base character when rendered on screen
- With PDF:
 - Diacritics are stored in a separate text chunks
 - Coordinates cause them to overlap
 - Diacritic chunk can be before or after the chunk it modifies

Diacritic Insertion

(10,20) (20,20)
des `

Sort by starting coordinate:

d e s `

Sort and merge into existing text:

d e ` s

Need to add to other side of Arabic text
before order is reversed.

Spacing Estimation

- Spaces and newlines are not explicitly stored.
- Spacing is achieved by direct placement of text.
- Extraction requires guessing where spaces and newlines should exist.
 - Is this text chunk's X-value further away than we expected?
 - Is this text chunk's Y-value further away than we expected?
- Spacing estimation can be done by keeping track of average character width thus far.
- Newline estimation can be done by keeping track of character heights.

PDFBox

- PDFBox is an open source Apache Incubator project
- It worked well for many documents in LTR languages
- We enhanced it to:
 - Correct direction of RTL text
 - Normalize ligatures and presentation forms
 - Merge diacritics into text
 - Better estimate where to add spaces
 - Fix parsing issues
 - Deal with corrupt / non-compliant files
- Can be freely downloaded (in next release):
<http://incubator.apache.org/pdfbox/>

鉴于对人类家庭所有成员的固有尊严及其平等的和不移的权利的承认，乃是世界自由、正义与和平的基础，鉴于对人权的无视和侮蔑已发展为野蛮暴行，这些暴行玷污了人类的良心，而一个基于尊重和信仰自由并在平等和互利的世界的基础上



BASIS
TECHNOLOGY

GOVERNMENT USERS
CONFERENCE

**BRINGING HLT
TO THE WARFIGHTER**



Thank You!

внимания, что преследование и пресечение в правах человека привели к варварским актам, которые возмущают совесть человечества, и что создание такого мира, в котором люди будут иметь свободу слова и убеждений и будут свободны от страха и нужды, провозглашено как высокое стремление людей; и принимая во внимание, что необходимо, чтобы права человека охранялись властью закона в целях обеспечения того, чтобы человек не был вынужден прибегать к насильственному и незаконному средству, к восстанию против тирании и угнетения, и что в целях содействовать развитию дружественных отношений между народами Объединенных Наций и ценности человеческой личности и в равноправии и сотрудничестве между народами, и что прогрессу и улучшению условий жизни при более справедливом международном сотрудничестве и сотрудничестве между государствами-членами обязались всеобщему уважению и соблюдению прав человека и общему пониманию характер

لما كان الاعتراف بالكرامة المتأصلة في جميع أعضاء الأسرة البشرية وبحقوقهم المتساوية الثابتة هو أساس الحرية والعدل والسلام في العالم. ولما كان تناسي حقوق الإنسان وازدراؤها قد أفضيا إلى أعمال همجية آدت الضمير الإنساني. وكان غاية ما يرنو إليه عامة البشر انبثاق عالم يتمتع فيه الفرد بحرية القول والعقيدة ويتحرر من الفزع والفاقة. ولما كان من الضروري أن يتولى القانون حماية حقوق الإنسان لكيلا يضطر المرء آخر الأمر إلى التمرد على الاستبداد والظلم. ولما كانت شعوب الأمم المتحدة قد أكدت في الميثاق من جديد إيمانها بحقوق الإنسان الأساسية وبكرامة الفرد وقدره وبما للرجال والنساء من حقوق متساوية وحرمت أمرها على أن تدفع بالرفعي الاجتماعي قدماً وأن ترفع مستوى الحياة في جو من الحرية أفسح. ولما كانت الدول الأعضاء قد تعهدت بالتعاون مع الأمم المتحدة على ضمان إطراد مراعاة حقوق الإنسان والحريات الأساسية