IBM

Enterprise Content Management

IBM
Information
and Analytics
Group

PDF Indexing

Bud Paton
World Wide Technical Consultant

09/2015

Data Streams

- The main four data streams
  - **PDF** – Portable Document Format
  - **AFP** – Advanced Function Presentation
  - **XML** – eXtensible Markup Language
  - **Line Data**
    - Text
    - ASCII
    - EBCDIC
    - SCS

Agenda

- **A little bit about PDF's**
- **Six methods to load PDF's in CMOD**
  1. Generic Index – individual PDF's
  2. Generic Index – one big file with offsets
  3. GXIFF Indexer – loading using XML index file
  4. PDF Indexer – using PDF Document properties
  5. NEW PDF Indexer – indexing "X-Y" coordinates
  6. PDF Indexer – using Page-Piece dictionary (PPD's)
- **Demo - PDF loading example**

What is PDF?

- **Portable Document Format (PDF) created by Adobe Systems, Inc.**
  - A file format used to represent documents in a manner independent of application software, hardware and operating systems. Each PDF file encapsulates a complete description of a fixed-layout flat document, including the text, fonts, graphics, and other information needed to display it.
  - In 1991, Adobe co-founder John Warnock outlined a system called “Camelot” that evolved into PDF.

PDF Versions

- 1993 – PDF 1.0 / Acrobat 1.0
- 1994 – PDF 1.1 / Acrobat 2.0
- 1996 – PDF 1.2 / Acrobat 3.0
- 2000 – PDF 1.3 / Acrobat 4.0
- 2001 – PDF 1.4 / Acrobat 5.0
- 2003 – PDF 1.5 / Acrobat 6.0
- 2005 – PDF 1.6 / Acrobat 7.0
- 2006 – PDF 1.7 / Acrobat 8.0
- 2006 – PDF 1.7 / Acrobat 8.2
- 2008* – PDF 1.7, Adobe Extension Level 3 / Acrobat 9.0
- 2009 – PDF 1.7, Adobe Extension Level 5 / Acrobat 9.1
- PDF.next 2.0 coming!

* - ISO standard ISO 32000-1:2008

PDF Variations

- **PDF/X - PDF for Exchange**
  - The purpose of PDF/X is to facilitate graphics exchange, and it therefore has a series of printing related requirements which do not apply to standard PDF files
- **PDF/E - PDF for Engineering**
  - For the creation of documents used in geospatial, construction and manufacturing workflows
- **PDF/VT - PDF for exchange of variable data and transactional (VT) printing**
  - it is the first variable-data printing (VDP) format which ensures modern International Color Consortium-based (ICC) color management through the use of ICC Output Intents
- **PDF/UA - PDF for Universal Access**
  - Conformance with PDF/UA provides accessibility for people with disabilities who use assistive technology such as screen readers, screen magnifiers, joysticks and other technologies to navigate and read electronic content.
  - PDF/UA is not a separate file-format but simply a way to use PDF
- **PDF/A - PDF for Archive**
  - PDF/A differs from PDF by omitting features ill-suited to long-term archiving, such as font linking

PDF Non Optimized – Optimized

- **There are two layouts to the PDF files**
  - **Non-linear** (not "optimized")
    - Non-linear PDF files consume less disk space than their linear counterparts, though they are slower to access because portions of the data required to assemble pages of the document are scattered throughout the PDF file
  - **Linear ("optimized")**
    - Linear PDF files (also called "optimized" or "web optimized" PDF files) are constructed in a manner that enables them to be read in a Web browser plugin without waiting for the entire file to download, since they are written to disk in a linear (as in page order) fashion.

PDF Structure

- **A PDF file consists primarily of objects, of which there are eight types:**
  - Boolean values, representing true or false
  - Numbers
  - Strings
  - Names
  - Arrays, ordered collections of objects
  - **Dictionaries**, collections of objects indexed by Names
  - Streams, usually containing large amounts of data
  - The null object

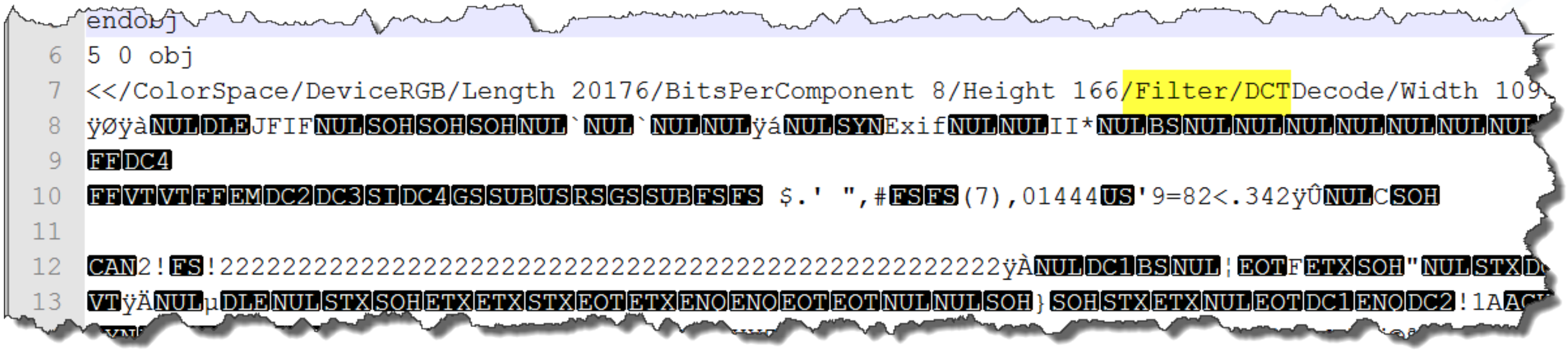
Compression(s) inside PDF’s

- **The “/Filter" keyword inside a pdf file is a indicator of the compression method used. Some of them are:**
    - Flate – used for compressing text as well as images
    - DCT - discrete cosine transform
    - CCITT G3/G4 – used for monochrome images
    - JPEG – a lossy algorithm that is used for images
    - JPEG2000 – a more modern alternative to JPEG, which is also used for compressing images
    - JBIG2 – an alternative to CCITT compression for monochrome images
    - LZW – used for compressing text as well as images but getting replaced by Flate
    - RLE – used for monochrome images
    - ZIP – used for grayscale or color images

PDF Compression algorithm's

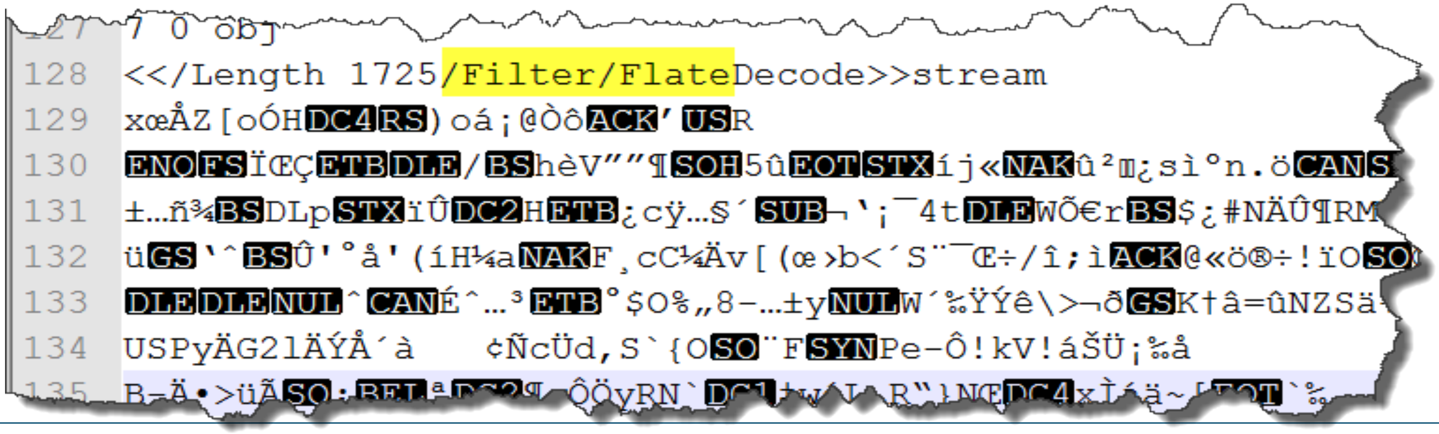
Edit of PDF showing different compression methods in one PDF

- Name following **/FILTER** is compression method



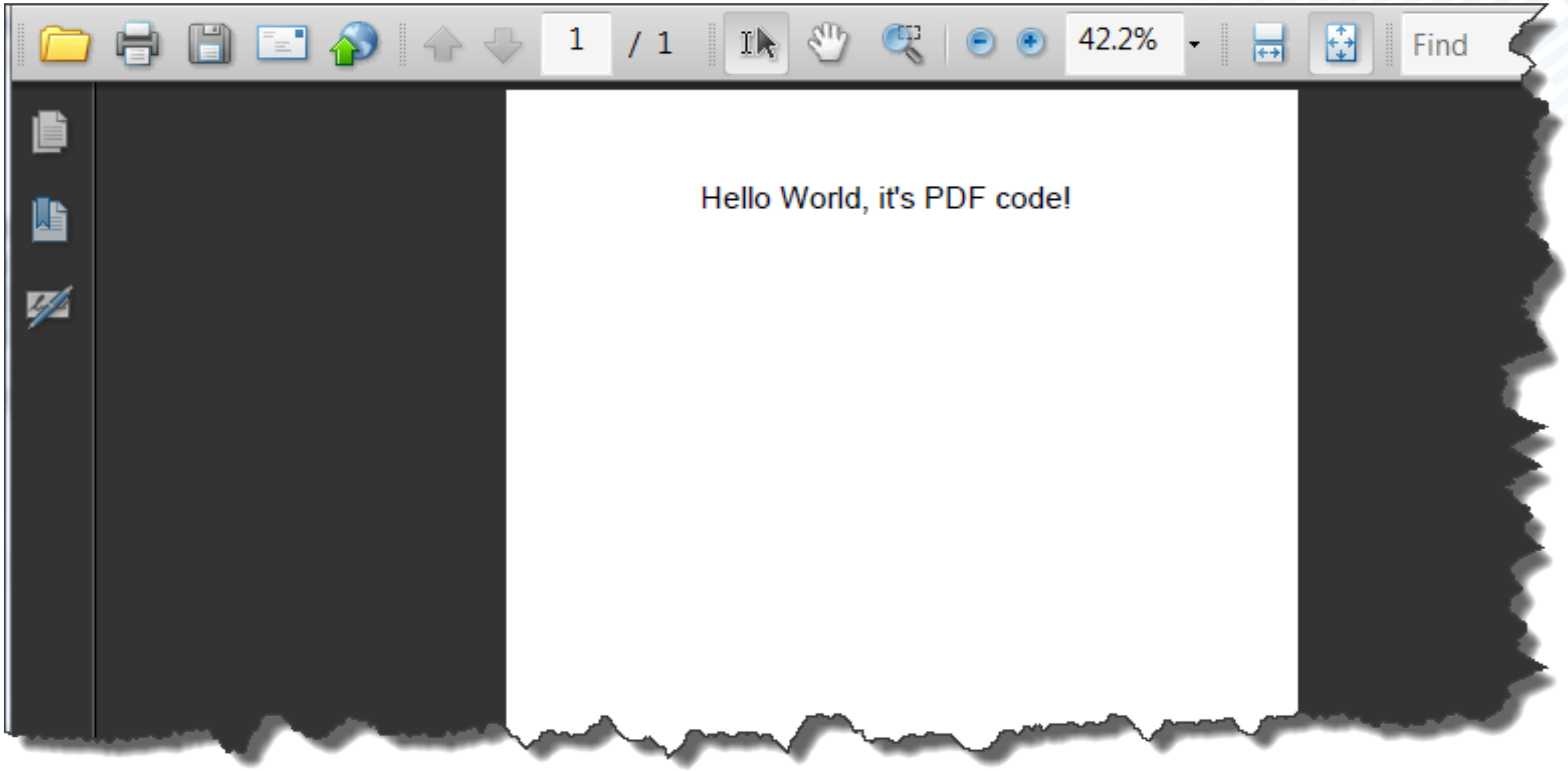
```
endobj
5 0 obj
<</ColorSpace/DeviceRGB/Length 20176/BitsPerComponent 8/Height 166/Filter/DCTDecode/Width 109
ÿØÿàNULDLEJFIFNULSOHSOHSOHNUL`NUL`NULNULÿáNULSYNExifNULNULII*NULBSNULNULNULNULNULNULNUL
FFDC4
FFVTVTFFEMDC2DC3SIDC4GSSUBUSRSGSSUBFSFS $.' ",#FSFS(7),01444US'9=82<.342ÿÛNULCSOH

CAN2!FS!22222222222222222222222222222222222222222222222222ÿÀNULDC1BSNUL¦EOTFETXSOH"NULSTXD
VTÿÄNULµDLENULSTXSOHETXETXSTXEOTETXENQENQEOTEOTNULNULSOH}SOHSTXETXNULEOTDC1ENQDC2!1AAC
```



```
7 0 obj
<</Length 1725/Filter/FlateDecode>>stream
xœÅZ[oÓHDC4RS)oá¡@ÒôACK'USR
ENQFSÏŒÇETBDLE/BShèV""¶SOH5ûEOTSTXíj«NAKû²¿sì°n.öCANS
±…ñ¾BSDLpSTXïÛDC2HETB¿cÿ…§´SUB¬'¡¯4tDLEWÕ€rBS$¿#NÄÛ¶RM
üGS'^BSÛ'°å'(íH¼aNAKF¸cC¼Äv[(œ›b<´S¨¯Œ÷/î;ìACK@«ö®÷!ïOSO
DLEDLENUL^CANÉ^…³ETB°$O%„8-…±yNULW´‰ŸÝê\>¬ðGSK†â=ûNZSä
USPyÄG21ÄÝÅ´à    ¢ÑcÜd,S`{OSO¨FSYNPe-Ô!kV!áŠÜ¡‰å
B-Ä•>üÃSO:BELªDC2¶ ÔÖyRN`DC1‡w I R"}NŒDC4xÌáä~[EOT`%
```

Hello World – PDF Sample output



Hello World, it's PDF code!

PDF code example

Heading

```
%PDF-1.4
1 0 obj
<< /Type /Catalog
/Outlines 2 0 R
/Pages 3 0 R
>>
endobj
```

Body

```
2 0 obj
<< /Type /Outlines
/Count 0
>>
endobj
3 0 obj
<< /Type /Pages
/Kids [4 0 R]
/Count 1
>>
Endobj
4 0 obj
<< /Type /Page
/Parent 3 0 R
/MediaBox [0 0 612 792]
/Contents 5 0 R
/Resources << /ProcSet 6 0 R
/Font << /F1 7 0 R >>
```

Body cont..

```
/PieceInfo
<</IBM-ODIndexes <</Private
<</Cname(Bud)
/BankNumber(0001)
/AcctNumber(123456)
/StmtDate(20120507)
>>
/LastModified(D:20130920000000Z)
endobj
5 0 obj
<< /Length 73 >>
stream
BT
/F1 24 Tf
100 100 Td
(Hello World it’s PDF Code!) Tj
ET
endstream
Endobj
6 0 obj
[/PDF /Text]
endobj
7 0 obj
<< /Type /Font
/Subtype /Type1
/Name /F1
/BaseFont /Helvetica
/Encoding /MacRomanEncoding
>>
endobj
```

XREF

```
xref
0 8
0000000000 65535 f
0000000009 00000 n
0000000074 00000 n
0000000120 00000 n
0000000179 00000 n
0000000364 00000 n
0000000466 00000 n
0000000496 00000 n
```

Trailer

```
trailer
<< /Size 8
/Root 1 0 R
>>
startxref
625
%%EOF
```

Enterprise Content Management

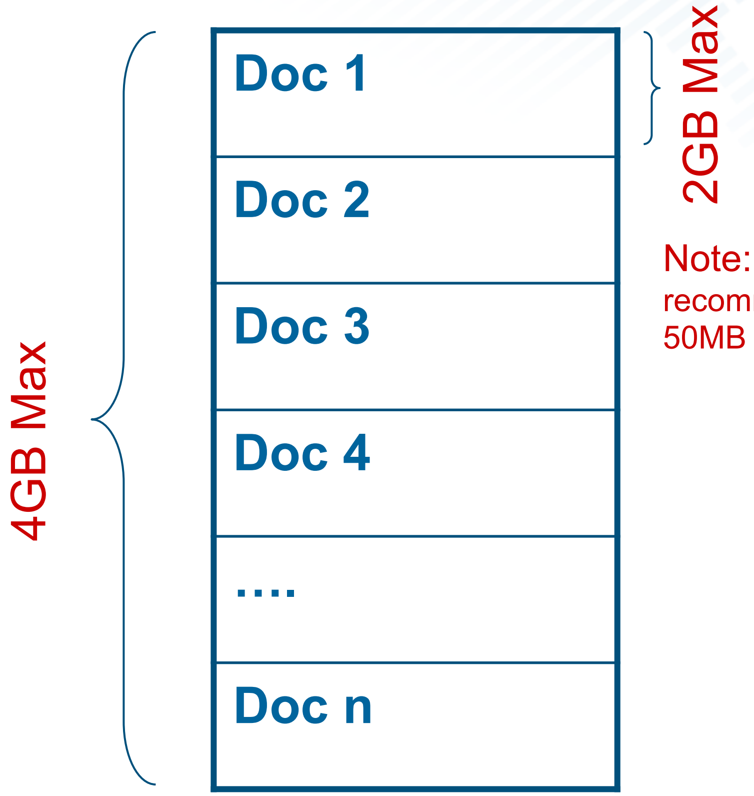
IBM
Information
and Analytics
Group

Six ways to load PDF’s into CMOD

PDF File Size Limit(s)

- **Input file size is 4GB (Adobe limit).**
  - One PDF file
- **NO limit on .out file size**
- **Max 2GB on one doc/segment**
- **Recommended no object larger than 50MB…**
- ***CMOD warnings on load***
- ***Need Memory to work on file!***

Begin PDF..



Note:
recommended > 50MB

End PDF..

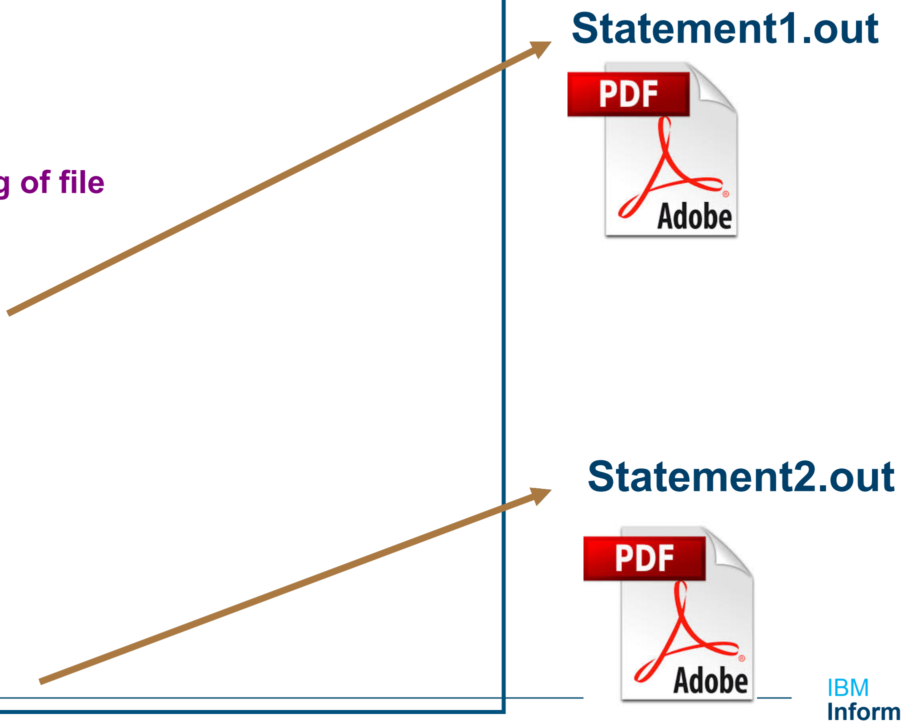
1. Generic Indexing – individual files

- **User or application supplies the indexes**
- **Must create XML "like" index file with indexes**
- **Must point to location of each file for loading**
- **IND file "like" XML, easy layout**
- **Very fast loading**
  - Does not look into files, just loads
  - Indexes already supplied
- **No PDF resource gathering because it never looks into files**

Generic Index - .IND file sample

- Loading two files -

```
COMMENT:
COMMENT: Generic Indexer Example 1
COMMENT: Different input file for each document
COMMENT:
COMMENT: Specify code page of the index data
CODEPAGE:819
COMMENT: Document #1
COMMENT: Index field #1
GROUP_FIELD_NAME:rdate
GROUP_FIELD_VALUE:07/13/99
COMMENT: Index field #2
GROUP_FIELD_NAME:studentID
GROUP_FIELD_VALUE:0012345678
COMMENT: document data starts at beginning of file
GROUP_OFFSET:0
COMMENT: document data goes to end of file
GROUP_LENGTH:0
GROUP_FILENAME:/arstmp/statement1.out
COMMENT: Document #2
COMMENT: Index field #1
GROUP_FIELD_NAME:rdate
GROUP_FIELD_VALUE:08/13/99
COMMENT: Index field #2
GROUP_FIELD_NAME:studentID
GROUP_FIELD_VALUE:0012345678
GROUP_OFFSET:0
GROUP_LENGTH:0
GROUP_FILENAME:/arstmp/statement2.out
```



Statement1.out

Statement2.out

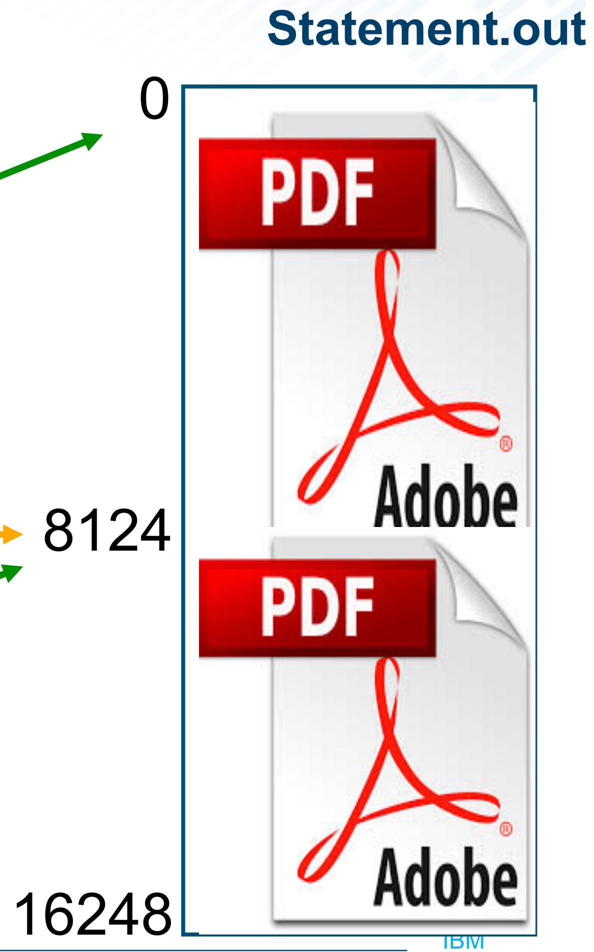
2. Generic Indexing – using offsets

- **User or application supplies the indexes, offsets, lengths Must create .IND file with indexes**
- **Input is individual PDF docs that are concatenated together**
- **Must point to location (offset – length) of each document in the file**
- **IND file "like" XML, easy layout**
- **Very fast loading**
  - Does not look into files, just loads
  - Indexes already supplied
  - Offsets and Lengths supplied
- **No PDF resource gathering because it never looks into files**

Generic Index - .IND file sample

- Two docs in one file -

```
COMMENT:
COMMENT: Generic Indexer Example 1
COMMENT: One file with many docs inside
COMMENT:
COMMENT: Specify code page of the index data
CODEPAGE:819
COMMENT: Document #1
COMMENT: Index field #1
GROUP_FIELD_NAME:rdate
GROUP_FIELD_VALUE:07/13/99
COMMENT: Index field #2
GROUP_FIELD_NAME:studentID
GROUP_FIELD_VALUE:0012345678
COMMENT: document data starts at beginning of file
GROUP_OFFSET:0
COMMENT: document length 8124 bytes
GROUP_LENGTH:8124
GROUP_FILENAME:/arstmp/statement.out
COMMENT: Document #2
COMMENT: Index field #1
GROUP_FIELD_NAME:rdate
GROUP_FIELD_VALUE:08/13/99
COMMENT: Index field #2
GROUP_FIELD_NAME:studentID
GROUP_FIELD_VALUE:0012345678
GROUP_OFFSET:8124
GROUP_LENGTH:8124
```



3. GXIFF - Generic XML Index File Format

- User or application supplies the indexes, offsets, lengths Must create real XML file with indexes
- Input is individual PDF docs or one PDF concatenated together
- Must point to location (offset – length) of each document in the file
- Real XML file like IND, easy layout
- Very fast loading
  - Does not look into files, just loads
  - Indexes already supplied
  - Offsets and Lengths supplied
- No PDF resource gathering because it never looks into files

Load any data type

- **Generic XML Index File Format**
  - Store any datatype into OnDemand
  - Same functionality as the GIFF (Generic Indexer File Format)

```
...
<oddoc>
   <oddataref file=“image1.jpg” offset=“0”
length=“0”>
   <odindex field=“account” value=“0452-6690”/>
   <odindex field=“title” value=“Customer letter”/>
   <odindex field=“date” value=“10/15/2014”/>
</oddoc>
<oddoc>
   <oddataref file=“image2.jpg” offset=“0”
length=“0”>
   <odindex field=“account” value=“0452-6690”/>
   <odindex field=“title” value=“Customer letter”/>
   <odindex field=“date” value=“10/15/2014”/>
</oddoc>
...
```

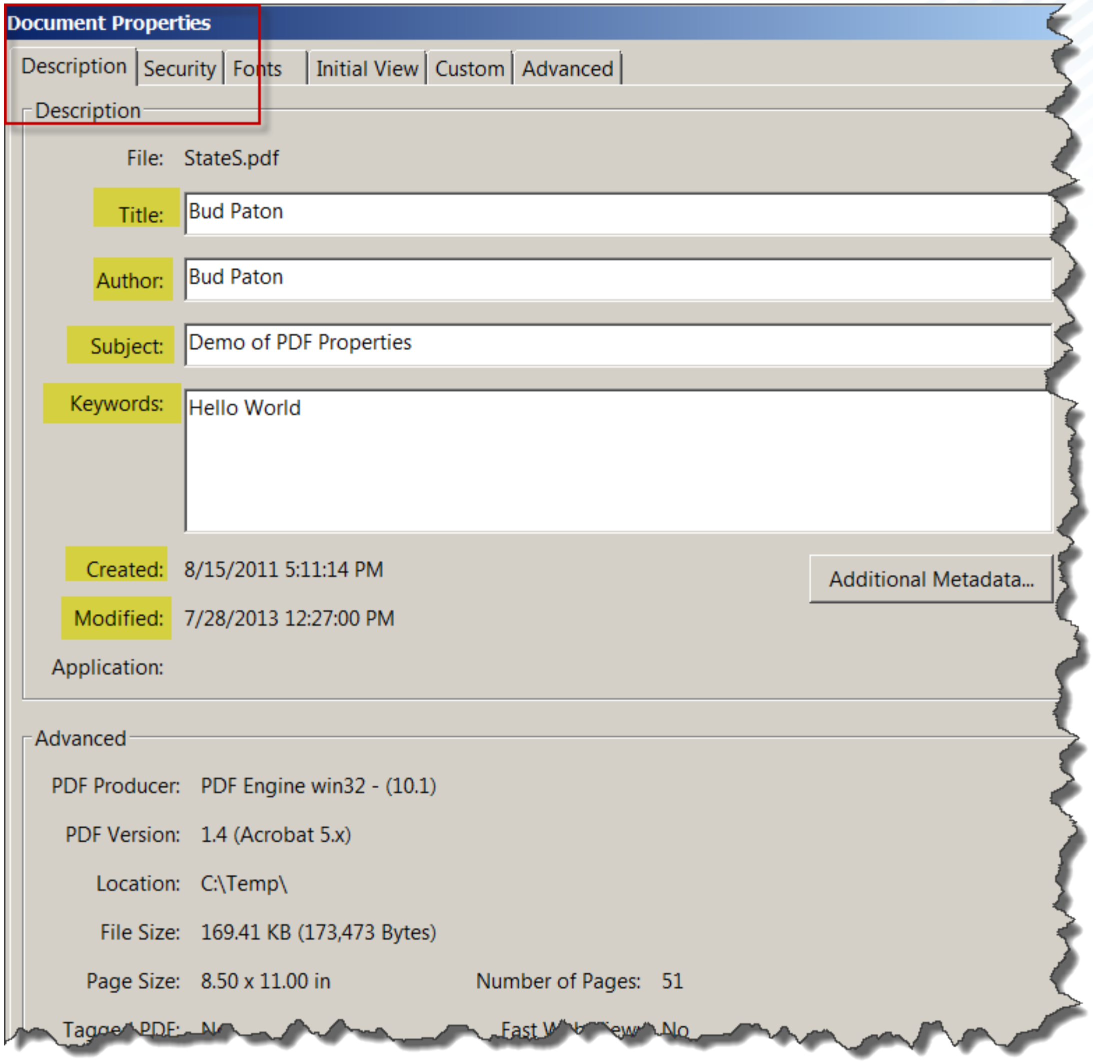
4. PDF Indexing – Using PDF Metadata

- User or application must place metadata (indexes) in PDF
- Metadata is only collected ONCE from PDF Properties
- Can gather PDF resources (fonts, images, forms)
- Very easy to set up in CMOD
- Good for one PDF file, no segmentation
- Mutually exclusive with other indexing
- Indexer looks into PDF, but only for metadata and resources
- Fast

PDF Indexing Metadata

- When INDEXMODE=METADATA
- The IBM CMOD PDF Indexer for Multiplatforms extracts fields from the document Information Dictionary that corresponds to the following metadata keywords, if they exist, and places their values into the .ind file:
  - Title
  - Author
  - Subject
  - Keywords
  - Creator
  - Producer
  - CreationDate
  - ModDate
  - Trapped

PDF Metadata – Document Properies



Description

File: StateS.pdf

Title: Bud Paton

Author: Bud Paton

Subject: Demo of PDF Properties

Keywords: Hello World

Created: 8/15/2011 5:11:14 PM

Modified: 7/28/2013 12:27:00 PM

Application:

Advanced

PDF Producer: PDF Engine win32 - (10.1)

PDF Version: 1.4 (Acrobat 5.x)

Location: C:\Temp\

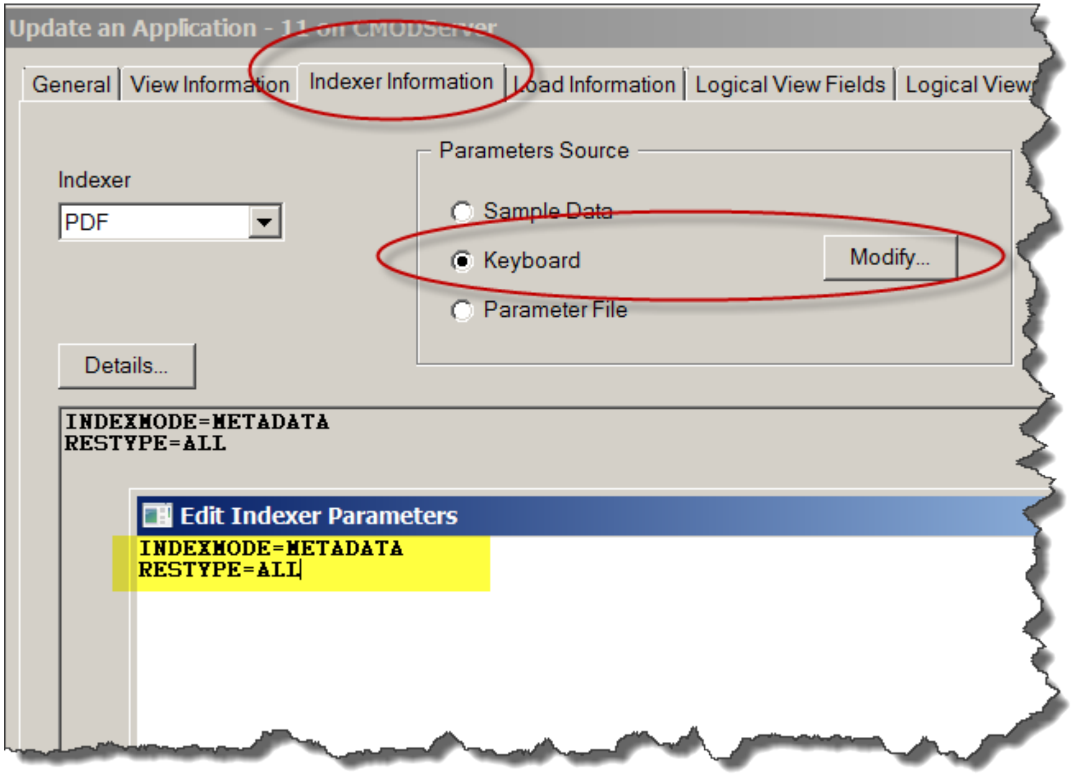
File Size: 169.41 KB (173,473 Bytes)

Page Size: 8.50 x 11.00 in Number of Pages: 51

Tagged PDF: No Fast Web View: No

PDF Metadata application setup

- Create new application group with fields matching metadata
- Create new application with INDEXMODE = METADATA



Indexer: PDF

Parameters Source

○ Sample Data

● Keyboard

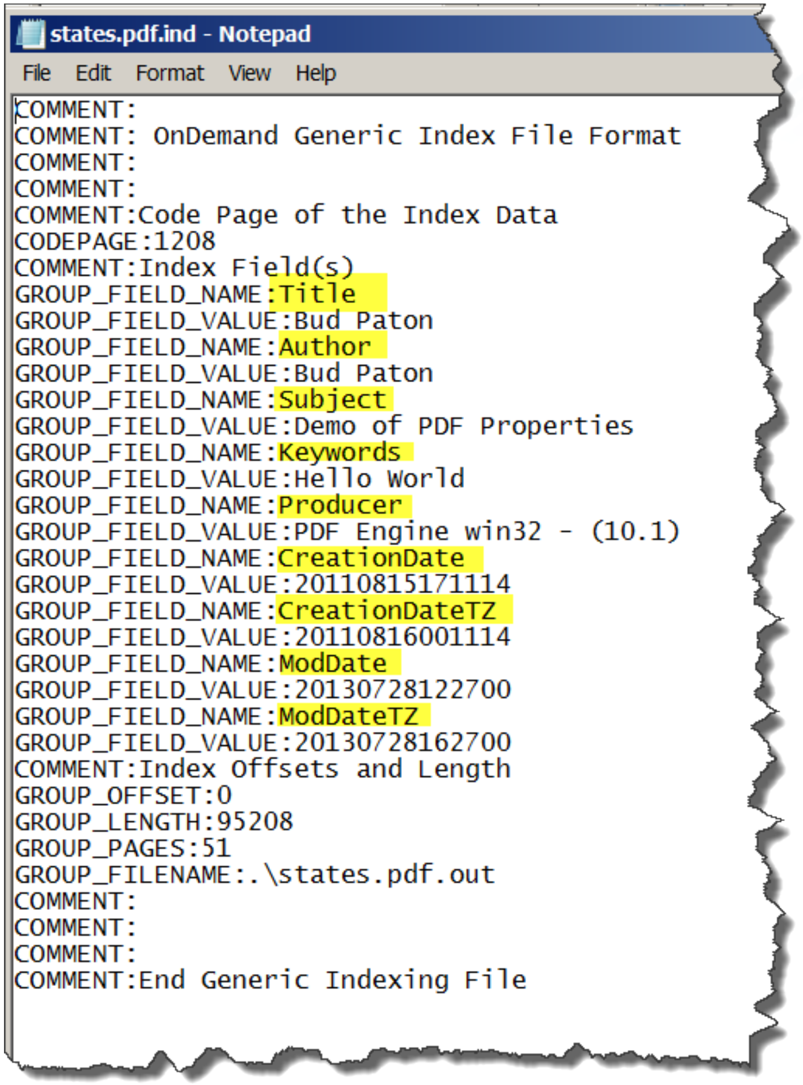
○ Parameter File

```
INDEXMODE=METADATA
RESTYPE=ALL
```

Edit Indexer Parameters

```
INDEXMODE=METADATA
RESTYPE=ALL
```

PDF Metadata .IND file created



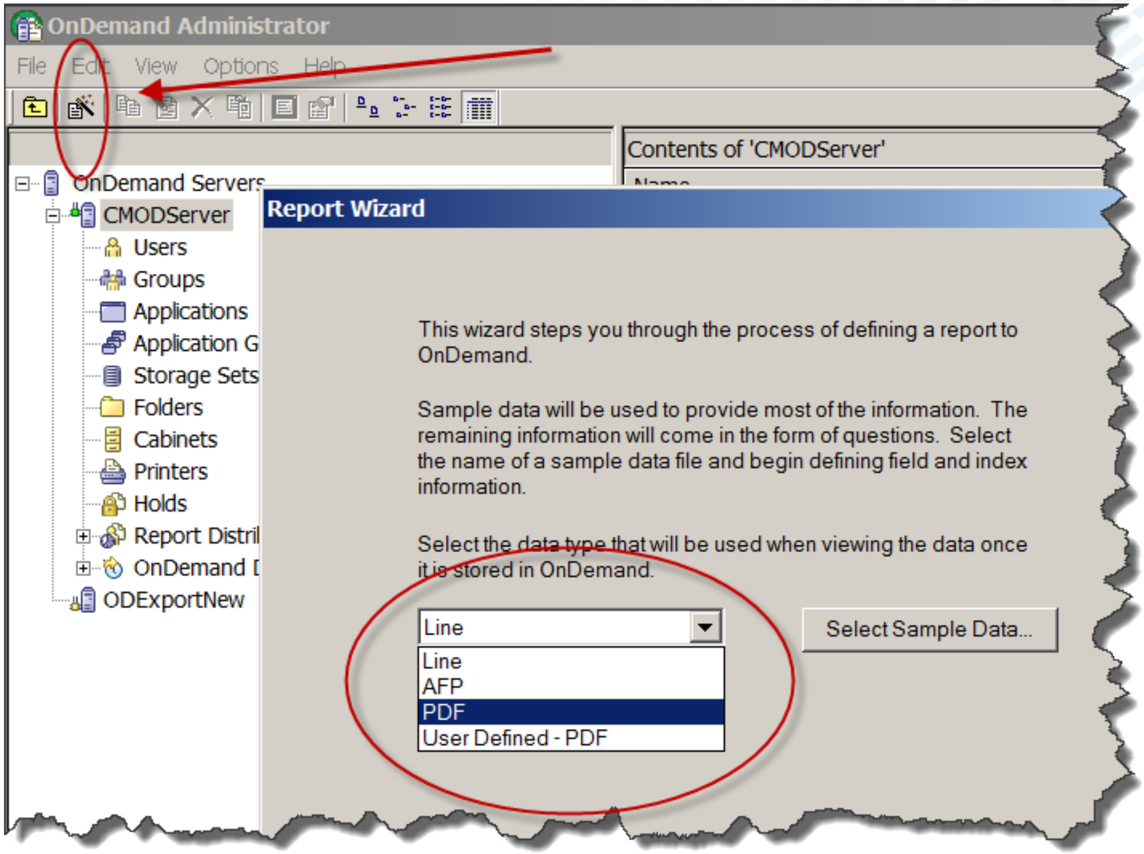
```
COMMENT:
COMMENT: OnDemand Generic Index File Format
COMMENT:
COMMENT:
COMMENT:Code Page of the Index Data
CODEPAGE:1208
COMMENT:Index Field(s)
GROUP_FIELD_NAME:Title
GROUP_FIELD_VALUE:Bud Paton
GROUP_FIELD_NAME:Author
GROUP_FIELD_VALUE:Bud Paton
GROUP_FIELD_NAME:Subject
GROUP_FIELD_VALUE:Demo of PDF Properties
GROUP_FIELD_NAME:Keywords
GROUP_FIELD_VALUE:Hello World
GROUP_FIELD_NAME:Producer
GROUP_FIELD_VALUE:PDF Engine win32 - (10.1)
GROUP_FIELD_NAME:CreationDate
GROUP_FIELD_VALUE:20110815171114
GROUP_FIELD_NAME:CreationDateTZ
GROUP_FIELD_VALUE:20110816001114
GROUP_FIELD_NAME:ModDate
GROUP_FIELD_VALUE:20130728122700
GROUP_FIELD_NAME:ModDateTZ
GROUP_FIELD_VALUE:20130728162700
COMMENT:Index Offsets and Length
GROUP_OFFSET:0
GROUP_LENGTH:95208
GROUP_PAGES:51
GROUP_FILENAME:.\states.pdf.out
COMMENT:
COMMENT:
COMMENT:
COMMENT:End Generic Indexing File
```

5. PDF Indexer – using X,Y Locations

- Must create indexing parms in CMOD for X,Y positions of indexes, triggers inside the PDF
- Will segment the file into individual docs
- Can use Graphical Wizard or ARSPDUMP for X,Y positions
- Will gather PDF resources (fonts, images, forms)

Using the Graphic Wizard for PDF Indexer?

- A utility used to create indexing parameters for the PDF Indexer
- Triggers and fields are identified by actual physical position of words on a page
- The Graphical Wizard parses each word on a page to provide x,y coordinates
- Those coordinates identify the trigger and field locations on a page



This wizard steps you through the process of defining a report to OnDemand.

Sample data will be used to provide most of the information. The remaining information will come in the form of questions. Select the name of a sample data file and begin defining field and index information.

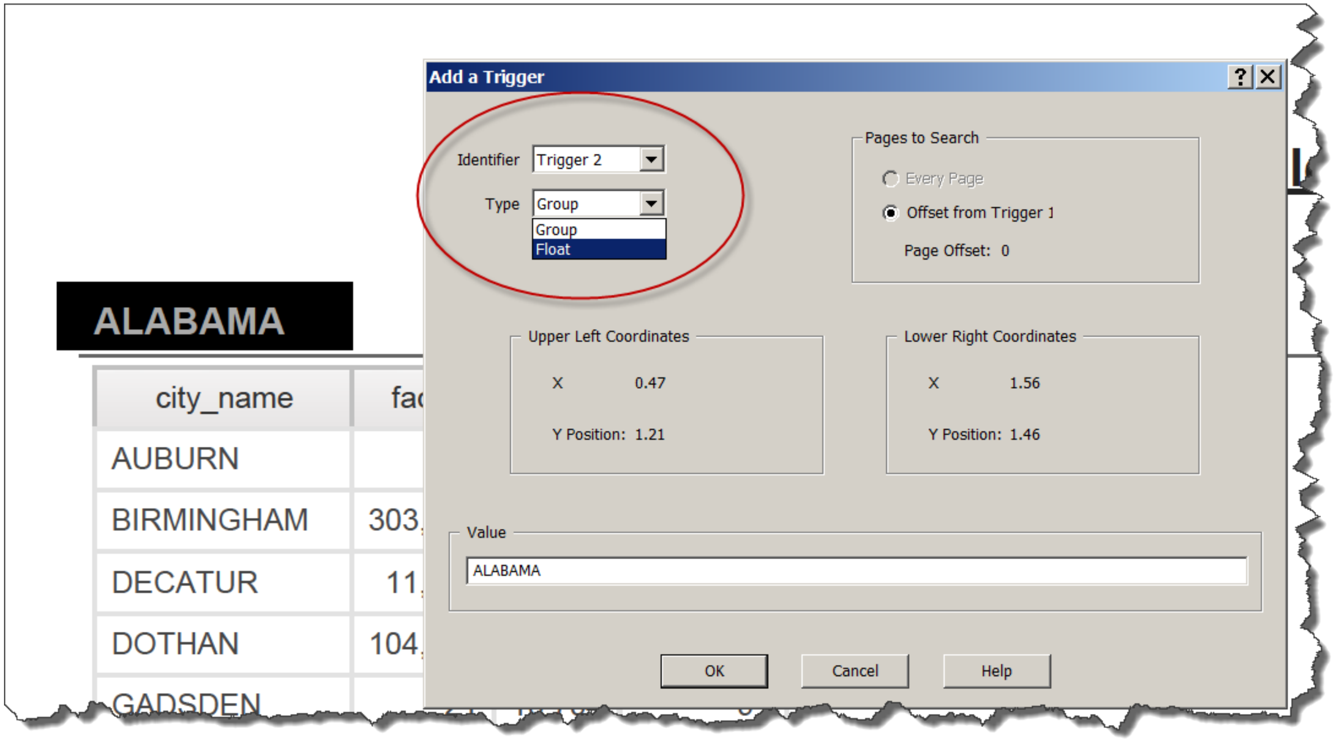
Select the data type that will be used when viewing the data once it is stored in OnDemand.

Line

NOTE: Must have PDF Indexer license

Graphical Wizard Continued

- First Trigger must be a Group Trigger
- Floating Triggers now supported with CMOD V9
- Up to 16 total Group or Floating Triggers



Add a Trigger

Identifier: Trigger 2

Type: Group

Pages to Search

○ Every Page

● Offset from Trigger 1

Page Offset: 0

Upper Left Coordinates

X 0.47

Y Position: 1.21

Lower Right Coordinates

X 1.56

Y Position: 1.46

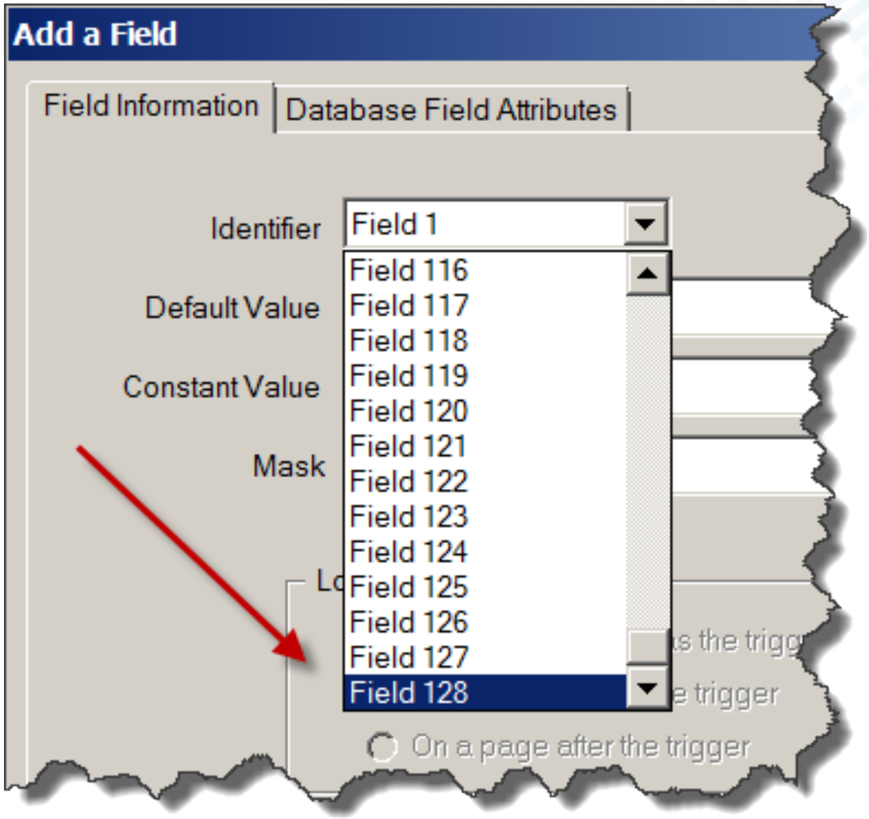
Value: ALABAMA

ALABAMA

city_name	fa
AUBURN	
BIRMINGHAM	303
DECATUR	11
DOTHAN	104
GADSDEN	

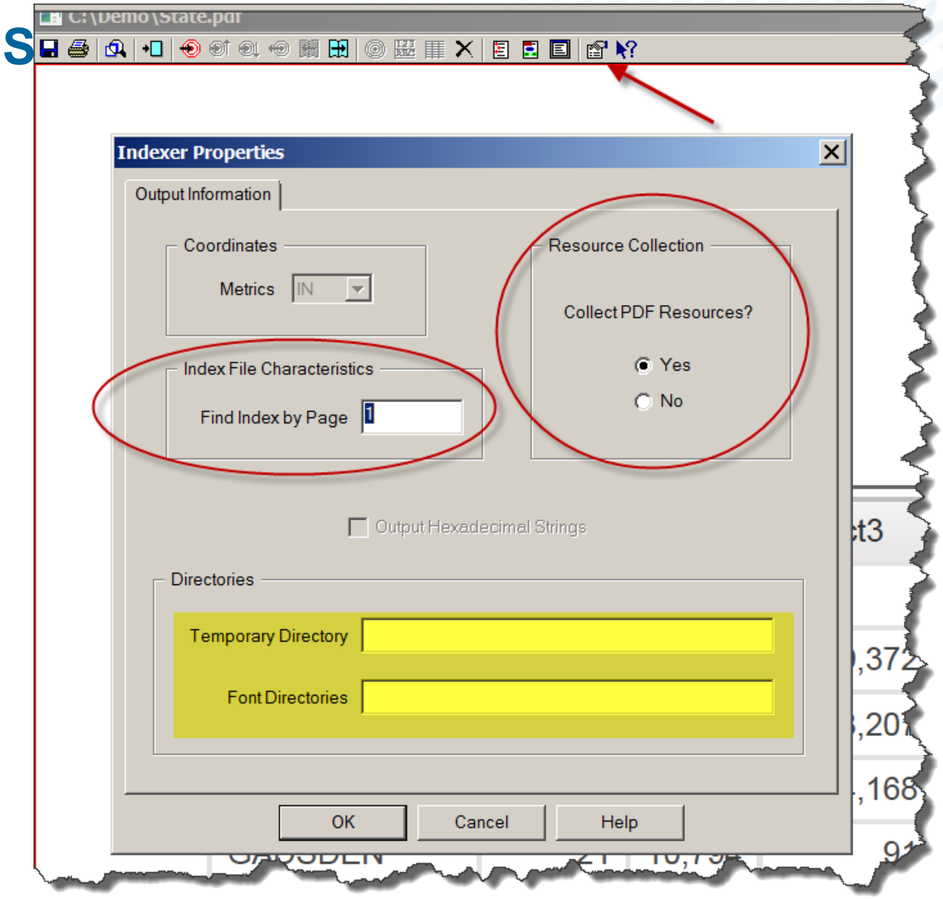
Graphical Wizard Continued

- Now up to 128 Fields



Graphical Wizard Index Properties

- Set Resource Collection to YES (default)
- Find by Page “1” (default)
- Temp Dir. Very important for performance!
- Font Dir. for external called fonts



C:\Demo\State.pdf

Indexer Properties

Output Information

Coordinates

Metrics: IN

Resource Collection

Collect PDF Resources?

● Yes

○ No

Index File Characteristics

Find Index by Page: 1

☐ Output Hexadecimal Strings

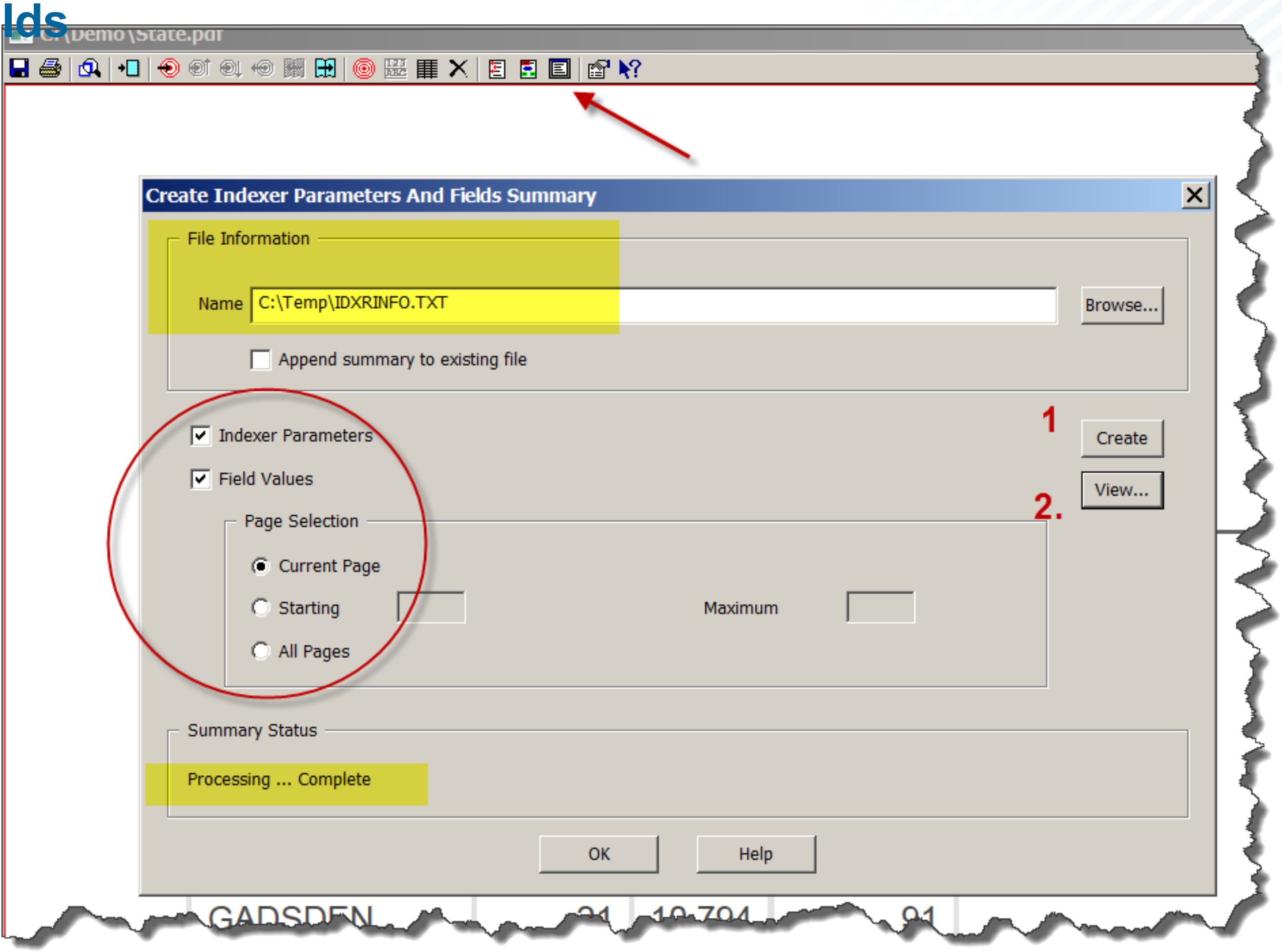
Directories

Temporary Directory:

Font Directories:

Graphical Wizard Indexer Reports

- Use this utility to quickly “see” your indexing parms and X,Y positions of all fields



Create Indexer Parameters And Fields Summary

File Information

Name C:\Temp\IDXRINFO.TXT

□ Append summary to existing file

☑ Indexer Parameters

☑ Field Values

Page Selection

● Current Page

○ Starting

Maximum

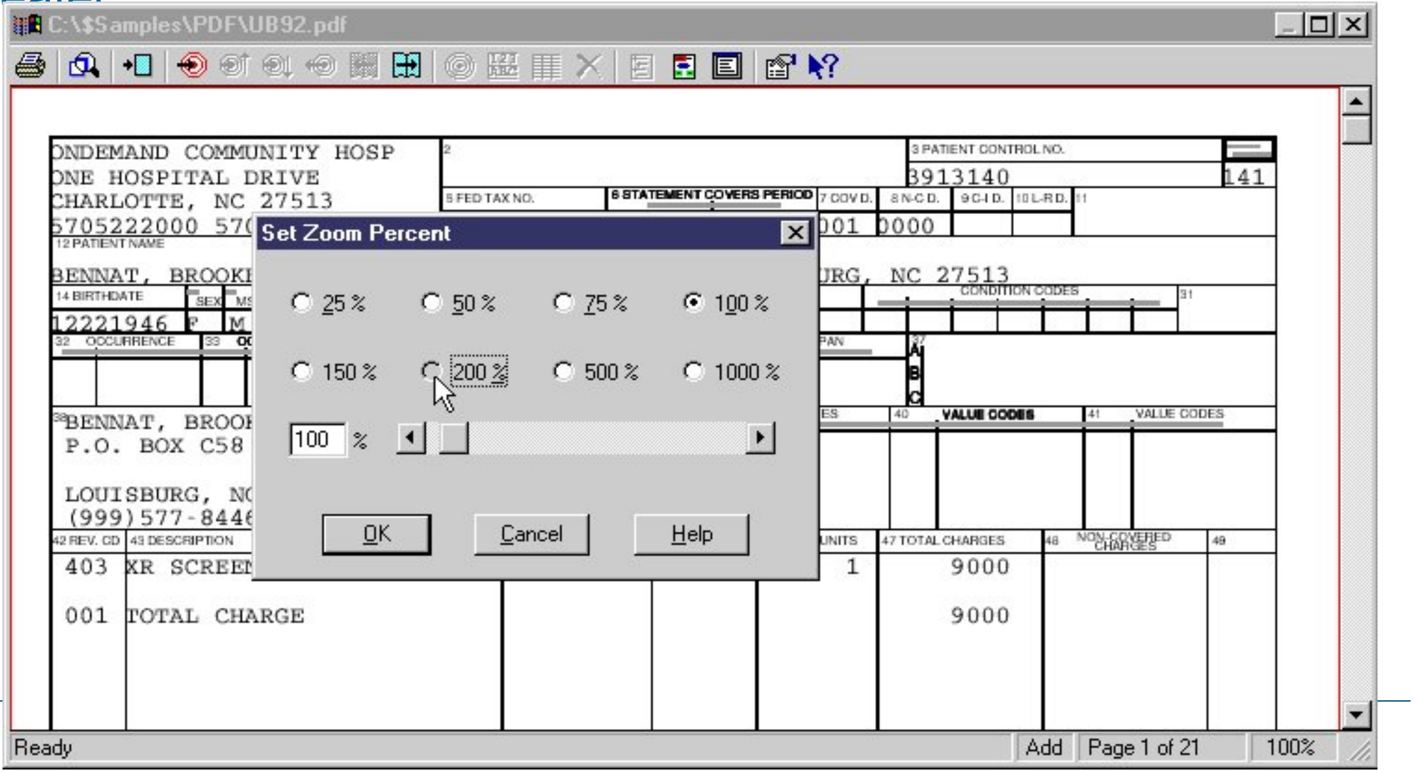
○ All Pages

Summary Status

Processing ... Complete

Graphical Wizard Markup

- **Markup is similar to spooled files except that the locations of triggers and fields are determined by physical location on the page (x and y coordinates), not by row and column number**
  - Small text, lines, and boxes can be hard to markup, we suggest you zoom in when working with PDF files
  - Suggest creating PDF docs with fonts greater than a size of 6 point or greater



C:\$Samples\PDF\UB92.pdf

ONDEMAND COMMUNITY HOSP
ONE HOSPITAL DRIVE
CHARLOTTE, NC 27513
5705222000 570

2

3 PATIENT CONTROL NO.
3913140
141

5 FED TAX NO. | 6 STATEMENT COVERS PERIOD | 7 COV D. | 8 N-C D. | 9 C-I D. | 10 L-R D. | 11

001 0000

12 PATIENT NAME
BENNAT, BROOKE
URG, NC 27513

14 BIRTHDATE | SEX | MS
12221946 F M

CONDITION CODES | 31

32 OCCURRENCE | 33 OC | PAN | 37 A B C

38 BENNAT, BROOK
P.O. BOX C58

LOUISBURG, NC
(999)577-8446

ES | 40 VALUE CODES | 41 VALUE CODES

42 REV. CD | 43 DESCRIPTION | UNITS | 47 TOTAL CHARGES | 48 NON-COVERED CHARGES | 49
403 XR SCREEN | 1 | 9000
001 TOTAL CHARGE | | 9000

Set Zoom Percent

○ 25 %　○ 50 %　○ 75 %　● 100 %

○ 150 %　○ 200 %　○ 500 %　○ 1000 %

100 %

OK　Cancel　Help

Ready　Add　Page 1 of 21　100%

6. PDF Indexer – Using Page Piece Dictionary

- New with CMOD V9
- User or application must insert indexes into the Page Piece Dictionary, using tools or API's.
  - Page-Piece Dictionary must be named: **IBM-ODIndexes**
- PDF Indexer will find the PPD and collect indexes
- Will segment the file into individual docs
- Will gather PDF resources (fonts, images, forms)

Page Piece Dictionary (from PDF Reference)

- A *page-piece dictionary (PDF 1.3)* can be used to hold private application data.
- Starting with PDF 1.4, private data may also be associated with the PDF document itself, by means of the PieceInfo entry in the document catalog

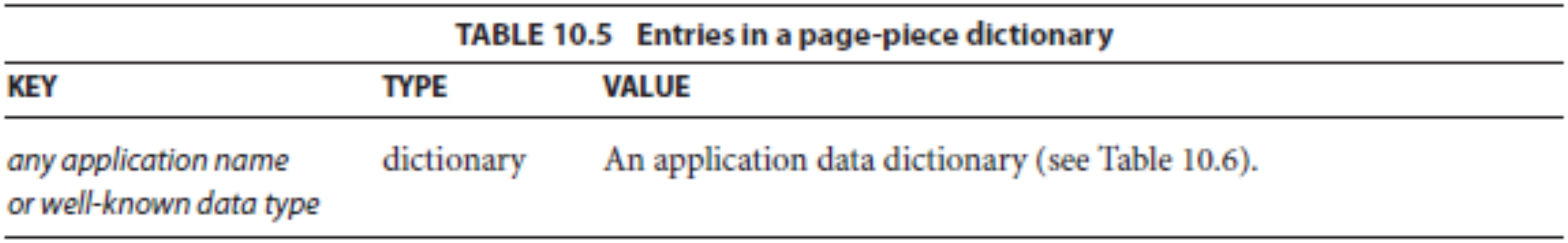


TABLE 10.5 Entries in a page-piece dictionary

KEY	TYPE	VALUE
any application name or well-known data type	dictionary	An application data dictionary (see Table 10.6).

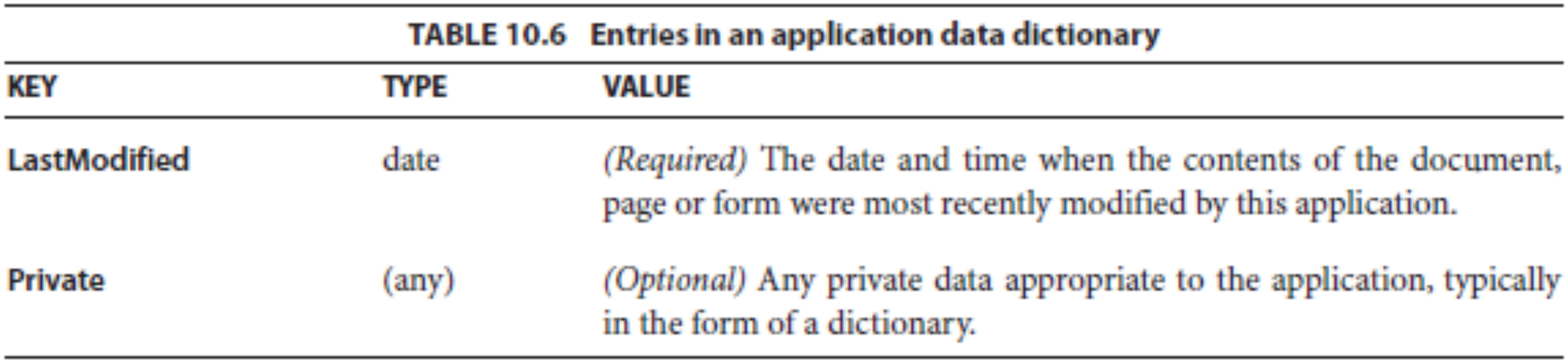


TABLE 10.6 Entries in an application data dictionary

KEY	TYPE	VALUE
LastModified	date	*(Required)* The date and time when the contents of the document, page or form were most recently modified by this application.
Private	(any)	*(Optional)* Any private data appropriate to the application, typically in the form of a dictionary.

PDF Page-Piece Dictionary example

- /PieceInfo <</IBM-ODIndexes <</Private is required.
- The first index has an index name of “Cname" and the value is “BUD".
- The <u>LastModified</u> date is <u>required</u> by the PDF architecture.

```
/PieceInfo <</IBM-ODIndexes <</Private
   <</Cname(BUD)
      /BankNumber(0000000001)
      /AcctNumber(18730845)
      /NoticeType(W)
      /StmtDate(20120507)
      /Branch(2 )
      >>
      /LastModified(D:20120619000000Z)
      >>
      >>
```

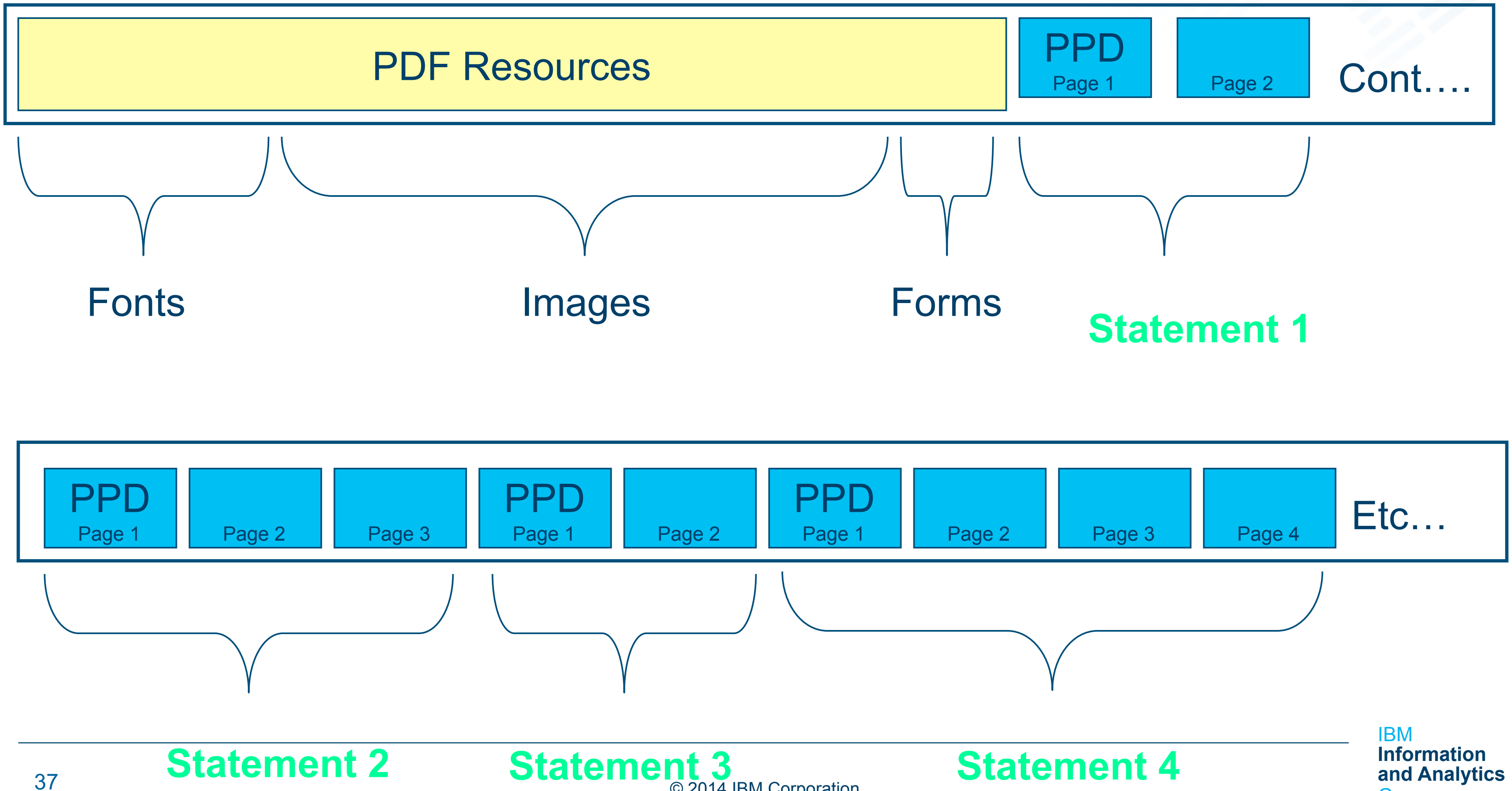
PDF Page-Piece Dictionary Setup

- **To place index values into the Page-Piece Dictionary use the following Adobe APIs:**
  - Use CosNewDict to create the Page-Piece dictionary.
  - Use CosNewString to create the PDF string object from the index value.
  - Use CosDictPutKeyString to put the name/value pair into the dictionary.
  - Use ASDateNew and ASDateGetTimeString to create the LastModified date.
  - Note: AFP sample program is available on request
- **Use the following indexing parameters to extract internal indexes. No other parameters are needed.**
  - **INDEXMODE=internal,xx (default indexes = 32)**
    - **If there are less than 32 indexes (most common) you only need: INDEXMODE=internal**

PDF Merged File

***NOTE:** In reality, pages and resources are spread throughout the file*

Large merged PDF file



PDF Resources

PPD Page 1

Page 2

Cont….

Fonts

Images

Forms

Statement 1

PPD Page 1

Page 2

Page 3

PPD Page 1

Page 2

PPD Page 1

Page 2

Page 3

Page 4

Etc…

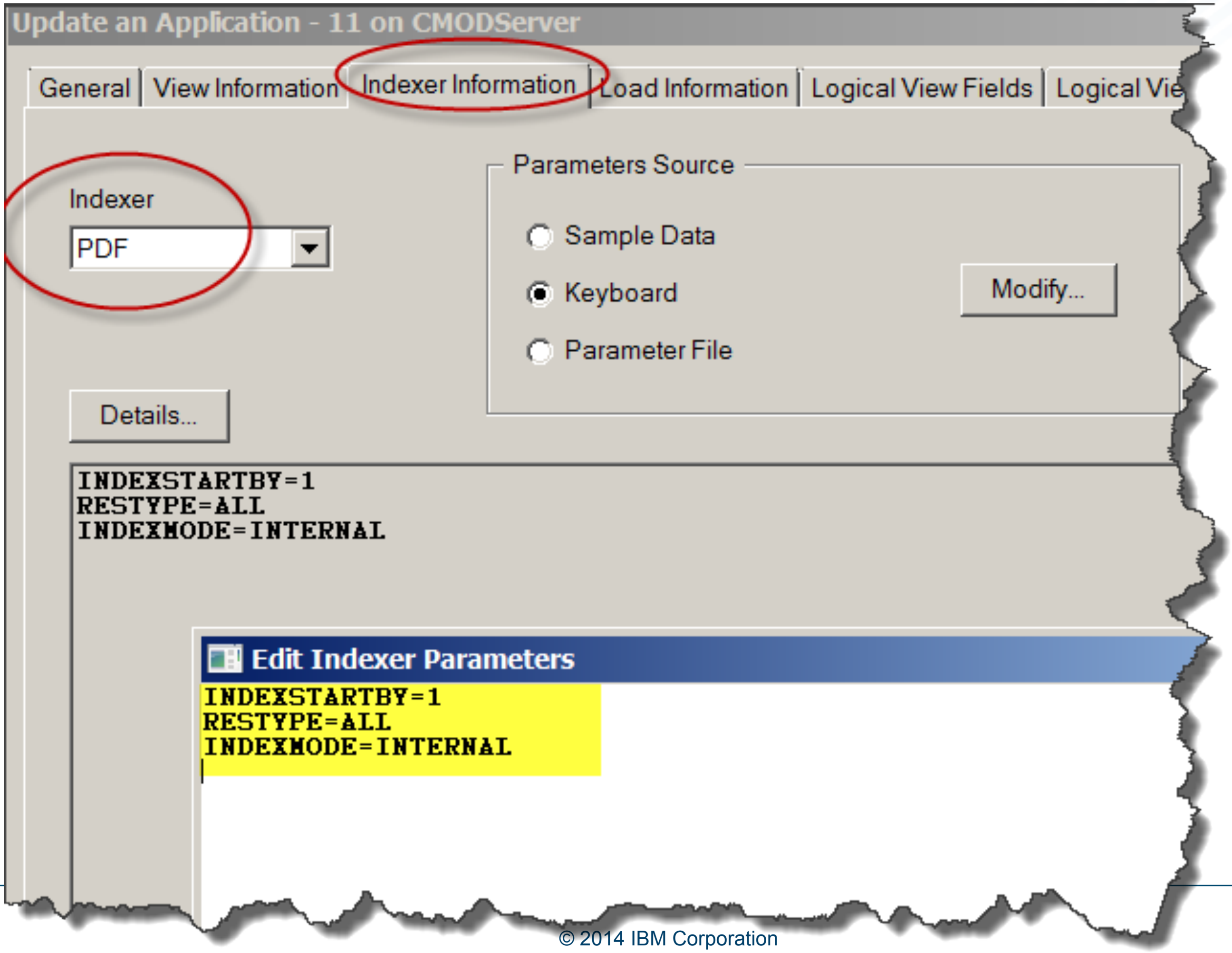
Statement 2

Statement 3

Statement 4

PDF Page-Piece Dictionary Application Setup

- Create new application group with fields that match PPD
- Create new application with INDEXMODE = INTERNAL



Update an Application - 11 on CMODServer

Indexer: PDF

Parameters Source

○ Sample Data
● Keyboard
○ Parameter File

```
INDEXSTARTBY=1
RESTYPE=ALL
INDEXMODE=INTERNAL
```

Edit Indexer Parameters

```
INDEXSTARTBY=1
RESTYPE=ALL
INDEXMODE=INTERNAL
```

Command to create PPD’s in PDF

- **This is used to create PPD’s for testing a PDF with PPD’s**
- **Must use the PDF X,Y indexing method to define indexes**
- **indexmode=add**
  - When indexmode=add, no output file is produced
  - The indexes are added to the input file, and a new input file with a 0 appended is created
  - Example would be sample.pdf becomes sample.pdf0 (with PPD’s added)
  - An empty index file is created
  - Use arspdoci.exe
    - Command arspdoci parmdd=myparms.txt

NOTE: *This command IS NOT DOCUMENTED and is NOT SUPPORTED!*

Document Composition vendors that create PPD’s

- GMC - Inspire (example)
- Assimilated Information Systems – DocFusion (example)
- ICON – DOPiX
- Crawford Technologies – (via transform)
- Ricoh (via transform)
- HP – Exstream (accepted - working on it)
- Pitney Bowes – DOC1 (working on it)
- Thunderhead – (in discussions)
- LaserNet – (working on it)

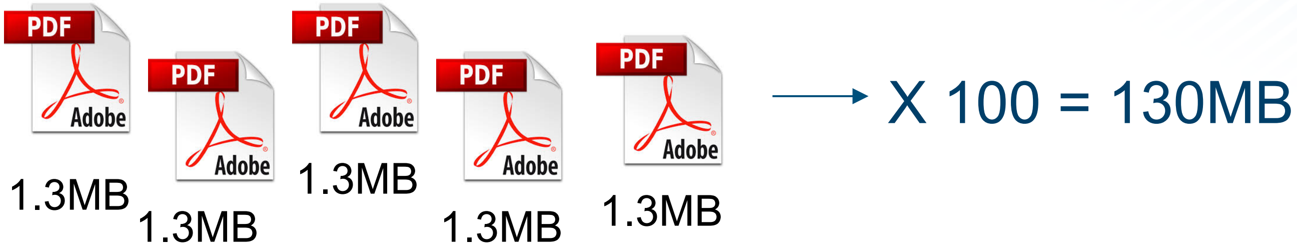
PPD Benefits

- **Storage**
  - instead of individual PDF's, now they can send one large PDF with multiple statements, and CMOD will break them up, remove the PDF resources (fonts, images, forms). See example on slide #9
- **Very fast indexing**
  - since CMOD has only to find PPD's it is very fast indexing
- **Different formats if indexes are similar**
  - you can have different documents inside one PDF and CMOD will still index and ingest them.
- **Easy CMOD administration**
  - CMOD applications only requires one statement - INDEXMODE=INTERNAL
- **PDF's can be dynamic vs. static**
  - because indexer is not looking at content only indexes
- **No more trial and error with XY positioning**

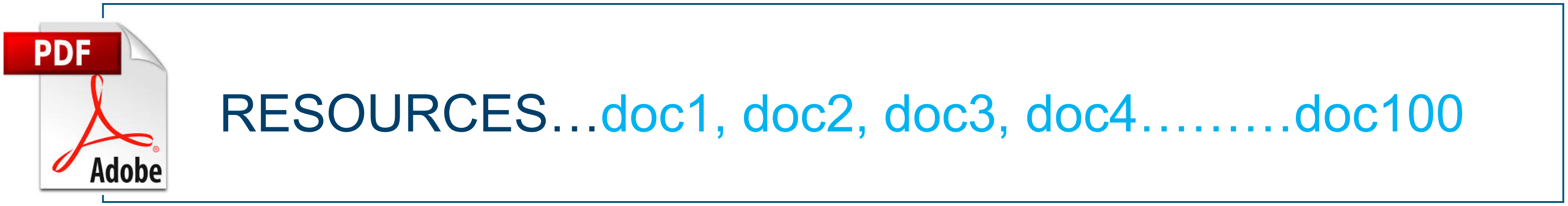
Seven PDF Load Methods Matrix

Method	One large file Many docs	Concatenated files	Gather Resources	User/Application Supplied Indexes	Speed
Generic Index Separate files	NO	NO	NO	YES	*****
Generic Index using offsets	NO	YES	NO	YES	****
GXIFF	NO	YES	NO	YES	*****
PDF Indexer Using Metadata	NO	NO	YES	YES	*****
PDF Indexer using X,Y positions	YES	NO	YES	NO	**
PDF Indexer using Page Piece Dictionary	YES	NO	YES	YES	*****

PDF One File vs. Many
CASE I



VS.



RESOURCES…doc1, doc2, doc3, doc4………doc100

1.6MB before CMOD – 1MB in CMOD

Effective Compression rate 130:1

PDF One File vs. Many
CASE II

.5MB .5MB .5MB .5MB .5MB → X 3840 = 2.1GB

GRNBank

Your Account Statement

0800-555-9999
Call GRN PhoneBank anytime for account information, current rates and answers to your questions.

Mr Dale Dalton
159 Hainworth Wood Road
Keighley
BD21 5DF
UK

Statement Date
Apr 01 to Apr 30, 2015

Your assets, at a glance

	Value on Apr 01	Value on Apr 30
🟧 Current - 31901689	8,000.00	9,322.25
🟨 Savings - 4130111	13,990.00	15,000.00
⬜ Investment - 518101-B	19,900.00	20,000.00
🟦 Home Equity	199,000.00	200,000.00
Total	£240,890.00	£244,322.25

Since Apr 01 your assets have grown by 3,432.25

Interest paid to you this statement period

Account	Interest Paid
Current - 31901689	5.00
Savings - 4130111	10.00
Investment - 518101-B	100.00
Mortgage - 613-0114	0.00
Total	£115.00

Value of assets on Apr 30

Apply for a Diamond Rewards Gold Card

Now you can use your Diamond Rewards Gold Card everywhere to earn points fast. Transform everything you spend into your next family holiday, dinner out with friends - even a flat screen television! With 1 point for every pound you spend on your card, 500 bonus points on your first purchase and rewards beginning at just 2,500 points, you'll be ready to redeem in no time!

How can we help you?
We treat our customers the way we'd want to be treated.

Call Us with Questions about your account

You can reach us 24 hours a day, 7 days a week, at 0800-555-9999.

These example statements are completely fictitious and are intended for illustrative purposes only. GRN Bank, the individual names and account information contained in these statements have been made up by Pitney Bowes. Any resemblance to real companies, people or accounts is strictly a coincidence.

Demonstration application was created on PBBI DOC1 and enhanced using Pitney Bowes Production Intelligence® output management software. It was printed on the Pitney Bowes IntelliJet™ Printing System, Model 30 onto 90 gsm, 98% brightness, untreated paper.

VS.

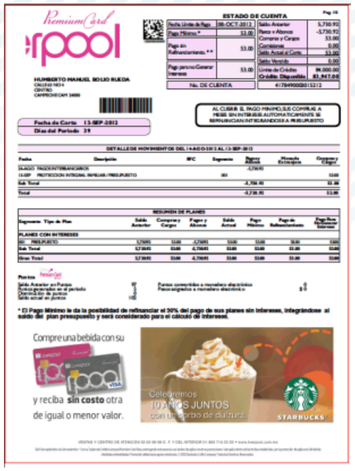
RESOURCES…doc1, doc2, doc3, doc4………doc100

165MB before CMOD – 30MB in CMOD

Effective Compression rate 71:1

PDF One File vs. Many
Case III

2.4MB 2.4MB 2.4MB 2.4MB 2.4MB → X 4889 = 12.4GB



PremiumCard

rpool

Pag. 1/1

ESTADO DE CUENTA

Fecha Límite de Pago	08-OCT-2012	Saldo Anterior	5,730.92
Pago Mínimo *	53.00	Pagos y Abonos	-5,730.92
Pago sin Refinanciamiento **	53.00	Compras y Cargos	53.00
		Comisiones	0.00
		Saldo Actual al Corte	53.00
		Saldo Vencido	0.00
Pago para no Generar Intereses	53.00	Límite de Crédito	84,000.00
		Crédito Disponible	83,947.00
No. DE CUENTA		4178490000015212	

HUMBERTO MANUEL BOLIO RUEDA
CALLE 63 NO 4
CENTRO
CAMPECHE CAM 24000

Fecha de Corte 13-SEP-2012
Días del Periodo [illegible]

AL CUBRIR EL PAGO MINIMO, SUS COMPRAS A MESES SIN INTERESES AUTOMATICAMENTE SE REFINANCIAN INTEGRANDOSE A PRESUPUESTO

DETALLE DE MOVIMIENTOS DEL 14-AGO-2012 AL 13-SEP-2012

Fecha	Descripción	RFC	Segmento	Pagos y Abonos	Moneda Extranjera	Compras y Cargos
24-AGO	PAGOS INTERBANCARIOS			-5,730.92		
13-SEP	PROTECCION INTEGRAL FAMILIAR/PRESUPUESTO		001			53.00
Sub Total				-5,730.92		53.00
Total				-5,730.92		53.00

RESUMEN DE PLANES

Segmento Tipo de Plan	Saldo Anterior	Compras y Cargos	Pagos y Abonos	Saldo Actual	Pago Mínimo	Pago de Refinanciamiento	Pago Para No Generar Intereses
PLANES CON INTERESES							
001 PRESUPUESTO	5,730.92	53.00	-5,730.92	53.00	53.00	53.00	53.00
Sub Total	5,730.92	53.00	-5,730.92	53.00	53.00	53.00	53.00
Gran Total	5,730.92	53.00	-5,730.92	53.00	53.00	53.00	53.00

Puntos PremiumCard

Saldo Anterior en Puntos	[illegible]	Puntos convertidos a monedero electrónico	0
Puntos generados en el periodo	[illegible]	Pesos asignados a monedero electrónico	$0
Disminución de puntos	[illegible]		
Saldo actual en puntos	[illegible]		

* El Pago Mínimo le da la posibilidad de refinanciar el 30% del pago de sus planes sin intereses, integrándose al saldo del plan presupuesto y será considerado para el cálculo de intereses.

Compre una bebida con su y reciba **sin costo** otra de igual o menor valor.

Celebremos 10 AÑOS JUNTOS con un sorbo de dulzura.

STARBUCKS

VENTAS Y CENTRO DE ATENCIÓN 52 62 99 99 D. F. Y DEL INTERIOR 01 800 713 55 55 • www.liverpool.com.mx

VS.

RESOURCES…doc1, doc2, doc3, doc4………doc4889

46MB before CMOD – 33MB in CMOD

Effective Compression rate 363:1

PDF Compression

- *When using PDF Indexer X,Y method*

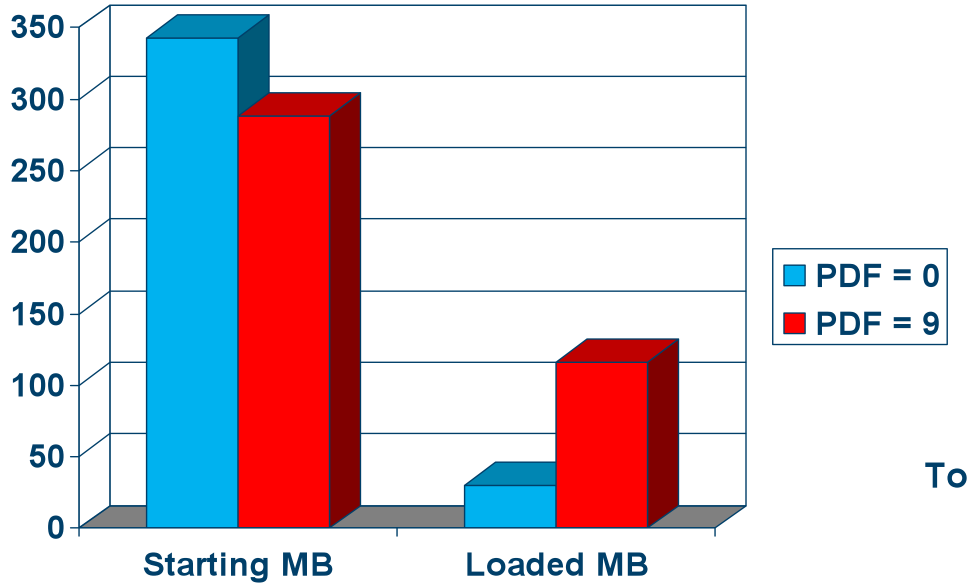
- **ADOBE PDF has 9 levels of compression 0-9**
  - 0 = no compression
  - 9= highest compression
- **Hard for PDF Indexer to work with compressed PDF**
- **Most applications turn on PDF compression as default**
  - TURN IT OFF
  - TURN IT OFF!
  - TURN IT OFF!!!!!

PDF Compression vs NO PDF Compression

- **Example:**
  - 3000 PDF statements - 1 file
    - Compressed file (9) 288MB = 116MB within CMOD compression
    - Time to process = 02:10:30 (index), :20 sec to load, Total 02:10:50
    - Compression ratio 2:1
  - Non- Compressed file (0) 343 MB= 30 MB within CMOD compression
    - Time to load = 00:01:27 (index), :05 sec to load, Total 00:01:31
    - Compression ratio 17:1 *

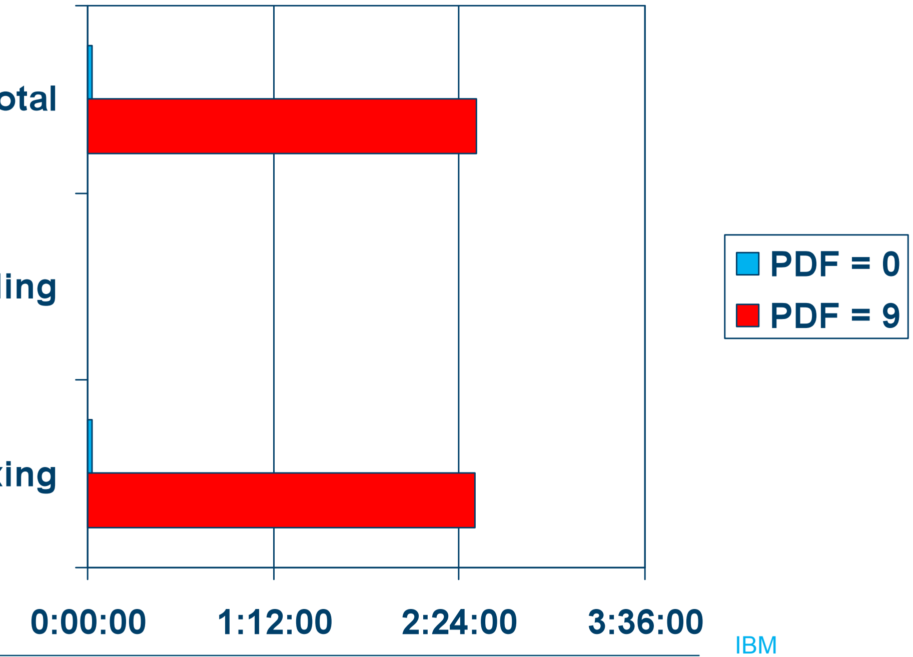
* NOTE: with CMOD removing resources. Test were run on Windows PC

PDF Compression Comparison



File Sizes

Time



IBM

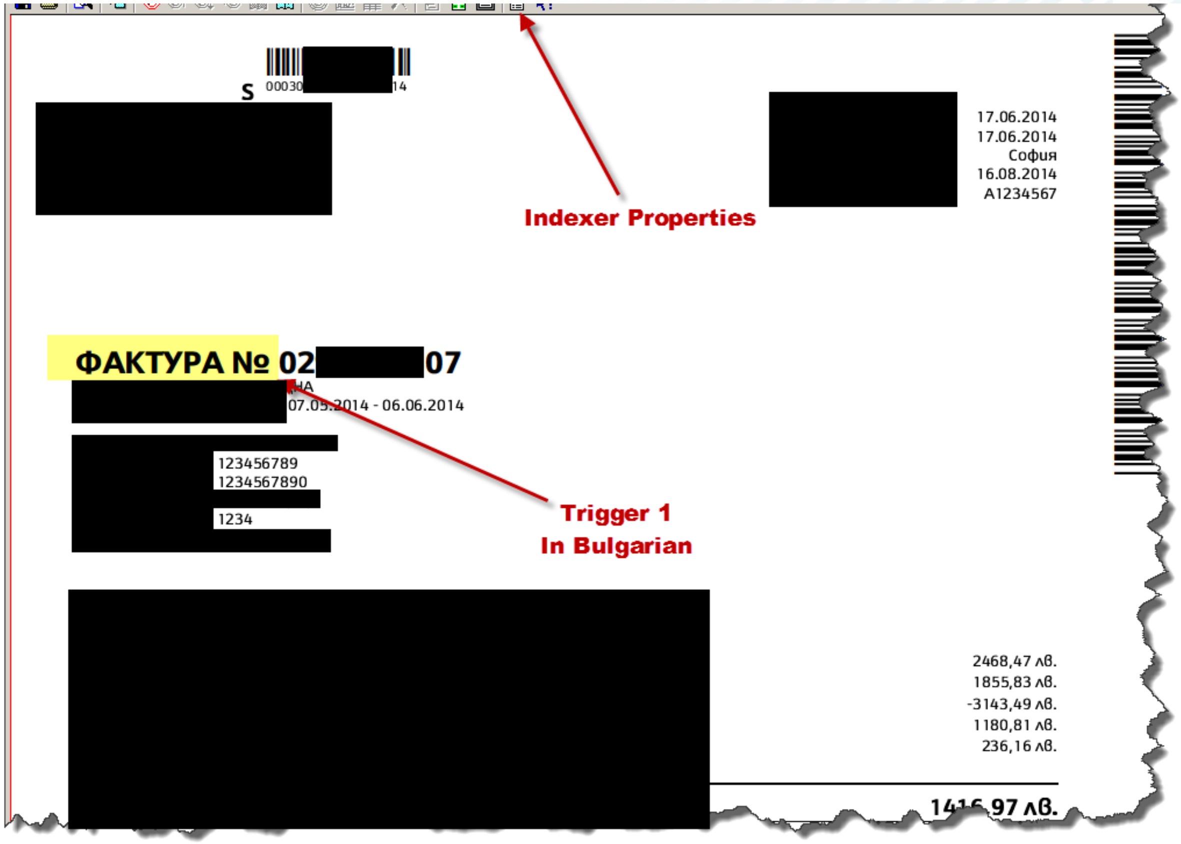
Enterprise Content Management

IBM
Information
and Analytics
Group

International Character Support

Working with International Languages – UTF-8

- **In this example – Bulgarian**
  - You need to tell the PDF Indexer to use (UTF-8) code page
  - Click on Indexer Properties icon on tool bar



S 00030[redacted]14

[redacted]

[redacted]

17.06.2014
17.06.2014
София
16.08.2014
A1234567

ФАКТУРА № 02[redacted]07

[redacted]

НА
07.05.2014 - 06.06.2014

[redacted]
123456789
1234567890
1234

[redacted]

2468,47 лв.
1855,83 лв.
-3143,49 лв.
1180,81 лв.
236,16 лв.

1416,97 лв.

Indexer Properties

Trigger 1
In Bulgarian

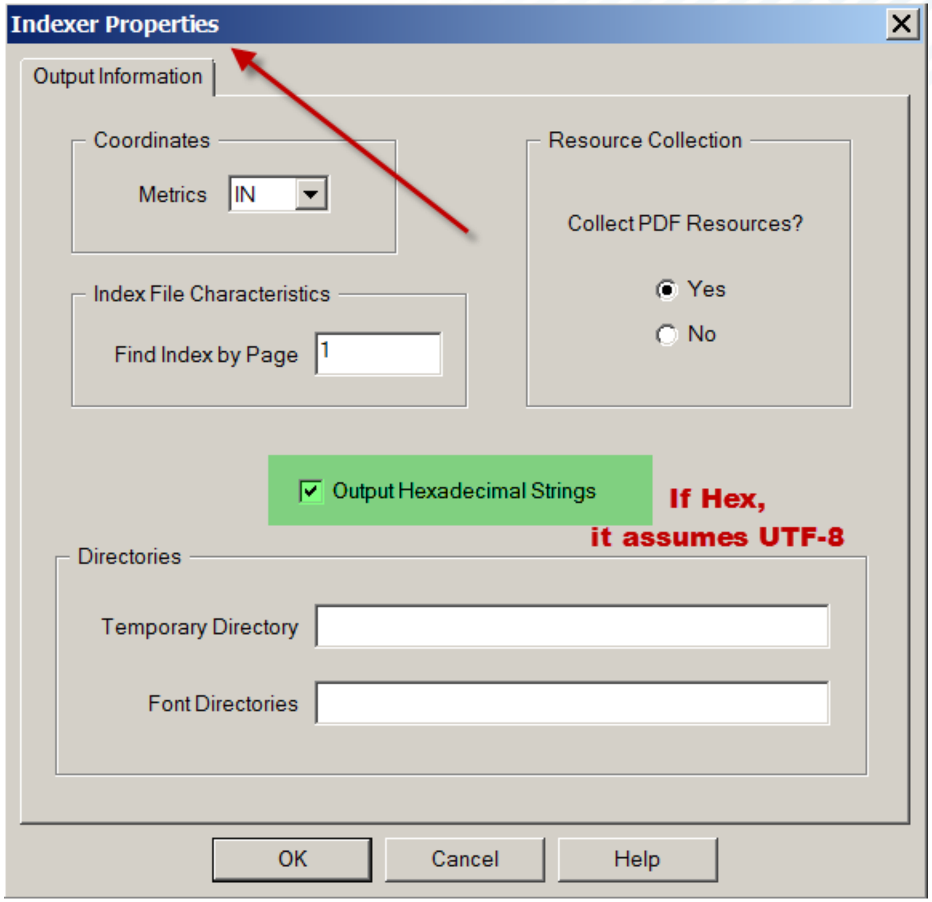
Working with International Languages – UTF-8 cont…

- **In Indexer Properties**
  - Click Output Hexadecimal Strings to use UTF-8
- **Now Indexer will output all strings as Hex**
- **Example:**

ФАКТУРА № (Invoice No.)

Becomes :

X'D0A4D090D09AD0A2D0A3D0A0D09020E28496'



Indexer Properties

Output Information

Coordinates

Metrics: IN

Index File Characteristics

Find Index by Page: 1

Resource Collection

Collect PDF Resources?

● Yes

○ No

☑ Output Hexadecimal Strings

If Hex, it assumes UTF-8

Directories

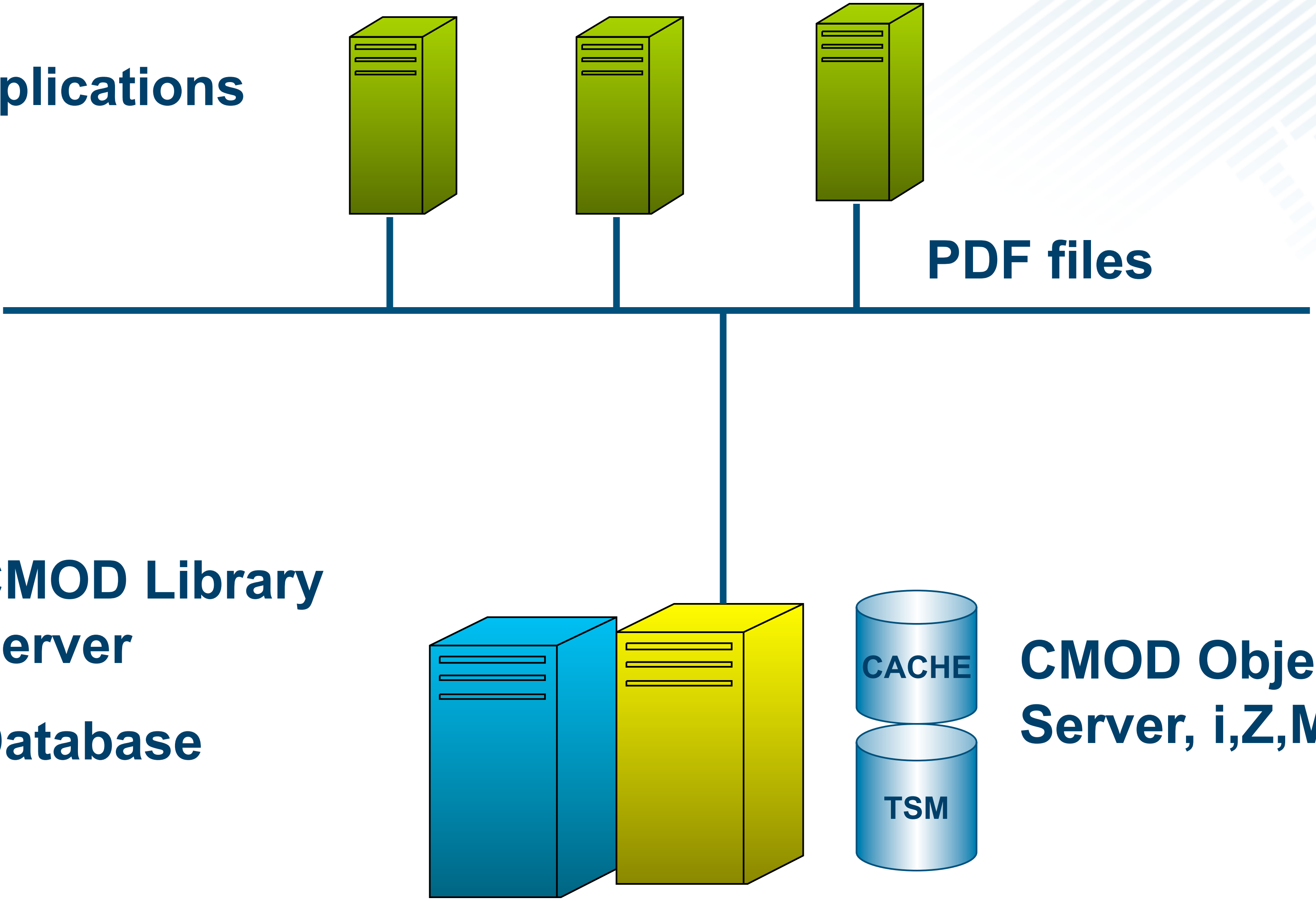
Temporary Directory:

Font Directories:

Enterprise Content Management

IBM
Information
and Analytics
Group

Using More Than One PDF Indexer



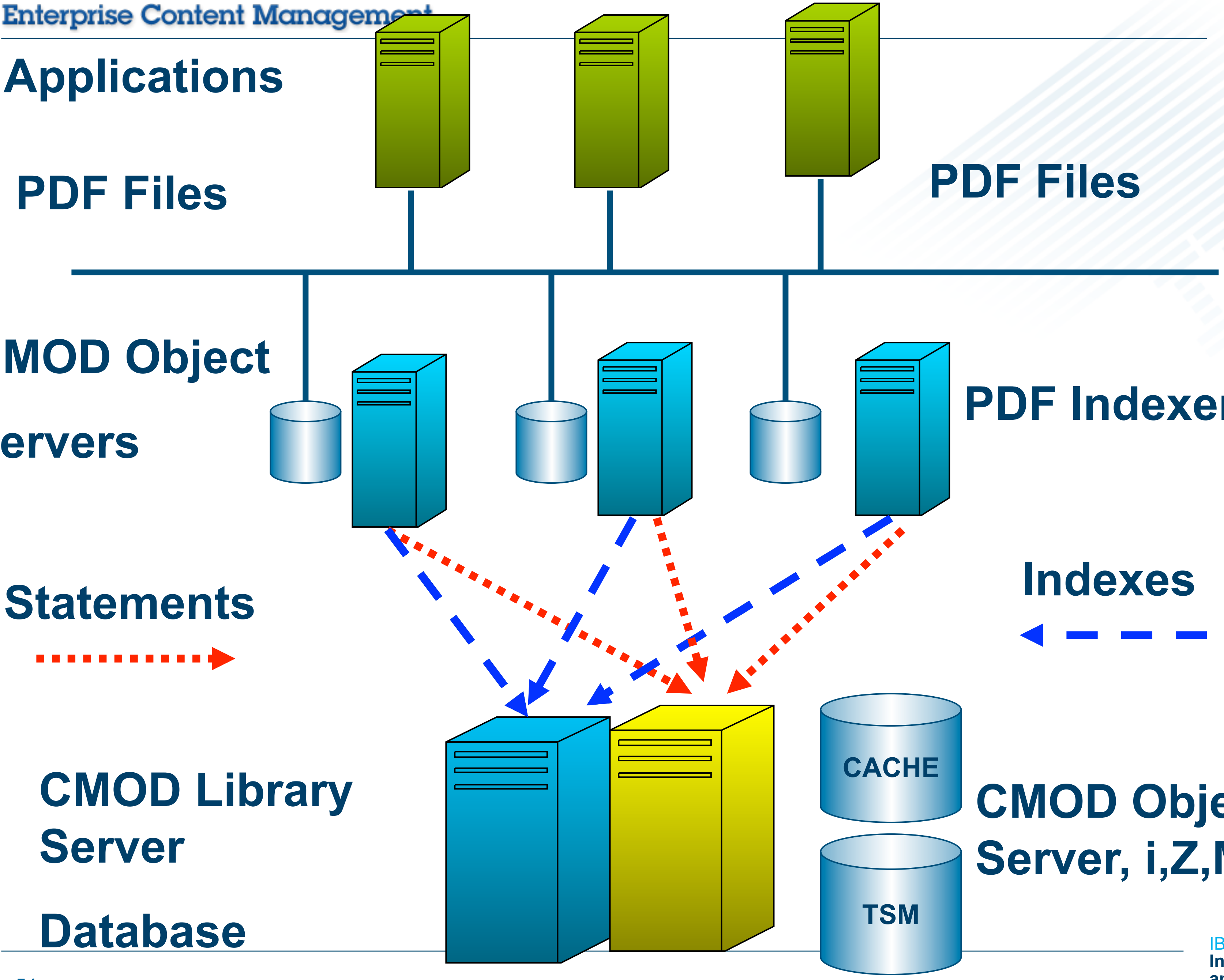
Applications

PDF files

CMOD Library Server

Database

CMOD Object Server, i,Z,MP



Applications

PDF Files

PDF Files

CMOD Object

Servers

PDF Indexer(s)

Statements

Indexes

CMOD Library
Server

Database

CMOD Object
Server, i,Z,MP