Let’s write a PDF file *r2*

A simple walk-through to learn
the basics of the PDF format
(at your rhythm)

PDF = Portable Document Format

ANGE ALBERTINI

reverse engineering

VISUAL DOCUMENTATION

@angealbertini

ange@corkami.com

http://www.corkami.com

Goal:
write a “Hello World” in PDF

PDF is text-based,
with some binary in specific cases.

But not in this example,
so just open a text editor.

Statements are separated by white space.

(any extra white space is ignored)

```
Any of these:
0x00 Null              0x0C Form Feed
0x09 Tab               0x0D Carriage Return
0x0A Line feed         0x20 Space
(yes, you can mix EOL style :( )
```

Delimiters don’t require white space before.

( ) < > [ ] { } /

Let’s *start!*

%PDF-_

A PDF starts with a %PDF-? signature followed by a version number.

1.0 <= version number <= 1.7

(it doesn’t really matter here)

%PDF-1.3

–

Ok, we have a valid signature ☺

%PDF-1.3

%_

A comment starts with %
until the end of the line.

%PDF-1.3

%file body

After the signature, comes the file body.

(we'll see about it later)

%PDF-1.3

%file body

xref

_

After the file body,
comes the cross reference table.

It starts with the **xref** keyword, on a separated line.

%PDF-1.3

%file body

xref

%xref table here

After the **xref** keyword, comes the actual table.

(we'll see about it later)

%PDF-1.3

%file body

xref

%xref table here

trailer_

After the table,
comes the trailer...

It starts with a **trailer** keyword.

%PDF-1.3

%file body

xref

%xref table here

trailer

%trailer contents

...and its contents.

(we'll see that later too…)

%PDF-1.3

%file body

xref

%xref table here

trailer

%trailer contents

startxref

–

Then, a pointer
to the xref table...

(with **startxref**)

%PDF-1.3

%file body

xref

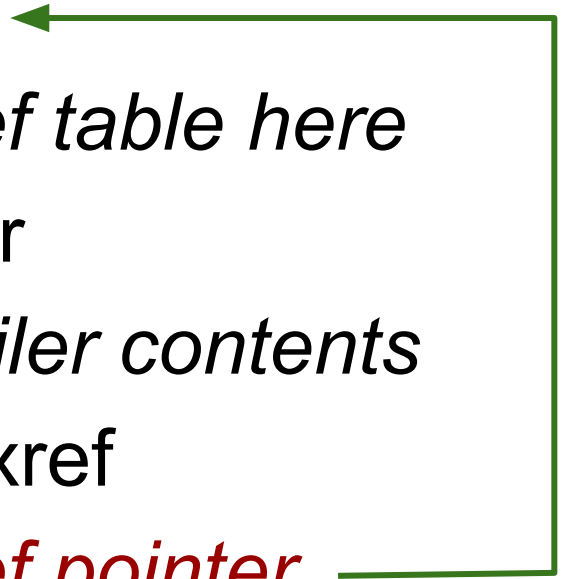
%xref table here

trailer

%trailer contents

startxref

%xref pointer



(later, too...)

%PDF-1.3

%file body

xref

%xref table here

trailer

%trailer contents

startxref

%xref pointer

%%EOF_

Lastly, to mark the end of the file...

...an **%%EOF** marker.

%PDF-1.3

%file body

xref

%xref table here

trailer

%trailer contents

startxref

%xref pointer

%%EOF

That’s the overall layout of a PDF document!

Easy ;)

%PDF-1.3

%file body

xref

%xref table here

trailer

%trailer contents

startxref

%xref pointer

%%EOF

Now, we just need to fill in the rest :)

Study time

Def: name objects

A.k.a. “strings starting with a slash”

/Name

A slash, then an alphanumeric string (no whitespace)

Case sensitive

/Name != /name

Names with incorrect case are just ignored
(no error is triggered)

Def: dictionary object

Sequence of ***keys*** and ***values***
(no delimiter in between)
enclosed in **<<** and **>>**
sets each ***key*** to ***value***

Syntax

<<
key value key value
[key value]*…
>>

Keys are always name objects

<< /Index 1>> sets /Index to 1

<< Index 1 >> is invalid
(the key is not a name)

Dictionaries can have any length

```
<< /Index 1
/Count /Whatever >>
```

sets /Index to 1
and /Count to /Whatever

Extra white space is ignored

(as usual)

```
<< /Index 1
/Count
/Whatever        >>
```

is equivalent to

```
<< /Index 1 /Count /Whatever >>
```

Dictionaries can be nested.

<< /MyDict << >> >>

sets /MyDict to << >> (empty dictionary)

White space before delimiters is not required.

<< /Index 1 /MyDict << >> >>

equivalent to

<</Index 1/MyDict<<>>>>

Def: indirect object

an object number (>0), a generation number (0*)
the **obj** keyword
the object content
the **endobj** keyword

* 99% of the time

Example

1 0 obj
3
endobj

is object #1, generation 0, containing “3”

Def: object reference

object number, object generation, **R**
number number R

ex: 1 0 R

Object reference

Refers to an indirect object as a value

ex: << /Root 1 0 R >> refers to
object number 1 generation 0
as the /Root

Used only as values in a dictionary

<< /Root 1 0 R >> is OK.
<< 1 0 R /Catalog>> isn’t.

Be careful with the syntax!

“1 0 3” is a sequence of 3 numbers 1 0 3

“1 0 R” is a single reference to an object number 1 generation 0

Def: file body

sequence of indirect objects
object order doesn’t matter

Example

1 0 obj 3 endobj
2 0 obj << /Index 1 >> endobj

defines 2 objects with different contents

%PDF-1.3

%file body

xref

%xref table here

trailer

%trailer contents

startxref

%xref pointer

%%EOF

Remember this?

A PDF document is defined by a tree of objects.

%PDF-1.3

%file body

xref

%xref table here

trailer

%trailer contents

startxref

%xref pointer

%%EOF

Now, let’s start!

%PDF-1.3

%file body

xref

%xref table here

trailer

<< _ >>

startxref

%xref pointer

%%EOF

The trailer is a dictionary.

%PDF-1.3

%file body

xref

%xref table here

trailer

<< /Root_ >>

startxref

%xref pointer

%%EOF

It defines a /Root name...

%PDF-1.3

%file body

xref

%xref table here

trailer

<< /Root 1 0 R_ >>

startxref

%xref pointer

%%EOF

...that refers to an object...

%PDF-1.3

%file body

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...that will be in the file body.

(like all the the other objects)

Recap:

the trailer is a dictionary that refers to a root object.

%PDF-1.3

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

Let’s create our first object...

%PDF-1.3

1 0 obj

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

…(with the standard object declaration)...

%PDF-1.3

1 0 obj

<< _ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...that contains a dictionary.

(like most objects)

%PDF-1.3

1 0 obj

<< /Type_ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...and its /Type is...

%PDF-1.3

1 0 obj

<< /Type /Catalog_ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...defined as /Catalog...

%PDF-1.3

1 0 obj

<< /Type /Catalog _ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

the /Root object also refers to the *page tree*...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages_ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...via a /Pages name...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R_>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...that refers to another object...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...which we'll create.

Recap:

object 1 is a catalog, and refers to a Pages object.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

Let’s create object 2.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

—

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

The usual declaration.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< _

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

It’s a dictionary too.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages_

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

The pages' object /Type has to be defined as … /Pages ☺

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids_

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

This object defines its children via /Kids...

Def: array

enclosed in [ ]
values separated by whitespace
ex: [1 2 3 4] is an array of 4 integers 1 2 3 4

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [_]

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...which is an array...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R_]

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

… of references
to each page object.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

_ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

One last step...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1_>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...the number of kids has to be set in /Count...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...and now
object 2 is complete!

Recap:

object 2 is /Pages;

it defines Kids + Count (pages of the document).

```
%PDF-1.3
1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj
2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

We can add our only Kid...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

—

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

…(a single page)...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< _ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

… a dictionary…

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type_ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

… defining a /Type...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page_ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

… as /Page.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent_ >>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

This grateful kid properly recognizes its own parent...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R_>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

… as you would expect ☺

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

—

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

Our page requires resources.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources_

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

Let’s add them...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << _ >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...as a dictionary:

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font_ >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

In this case, fonts...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font << _ >> >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...as a dictionary.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font <<

>> >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

We define one font...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font <<

/F1_

>> >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...by giving it a name...

```
%PDF-1.3
1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj
2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj
3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font <<
/F1 << _ >>
>> >>
>>
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...and setting its parameters:

```
%PDF-1.3
1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj
2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj
3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font <<
/F1 << /Type_ >>
>> >>
>>
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

its type is ...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font <<

/F1 << /Type /Font_ >>

>> >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

```
%PDF-1.3
1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj
2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj
3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font <<
/F1 << /Type /Font /Subtype_ >>
>> >>
>>
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

Its font type is...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font <<

/F1 << /Type /Font /Subtype /Type1_

>> >> >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

…(Adobe) Type1...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font <<

/F1 << /Type /Font /Subtype /Type1

/BaseFont_ >> >> >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

...and its name is...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font <<

/F1 << /Type /Font /Subtype /Type1
/BaseFont /Arial_>> >> >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font << /F1 << /Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

One thing is missing in our page...

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font << /F1 << /Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>

/Contents_

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

The actual page contents...

```
%PDF-1.3
1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj
2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj
3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 << /Type /Font
/Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R_
>>
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

… as a reference to another object.

%PDF-1.3

1 0 obj

<< /Type /Catalog /Pages 2 0 R >>

endobj

2 0 obj

<< /Type /Pages

/Kids [3 0 R]

/Count 1 >>

endobj

3 0 obj

<< /Type /Page /Parent 2 0 R

/Resources << /Font << /F1 << /Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>

/Contents 4 0 R

>>

endobj

xref

%xref table here

trailer

<< /Root 1 0 R >>

startxref

%xref pointer

%%EOF

That’s all for our page object.

Recap:

object 3 defines a /Page,
its /Parent, /Resources (fonts)
and its /Contents is
in another object.

(thank you Mario!)

Study time

Def: stream objects

So far, everything is text.
How do you store binary data (images,...) ?

Stream objects are objects.

They start and they end like any other object:

Ex: 1 0 obj

...

endobj

Stream objects contain a stream.

between ***stream*** and ***endstream*** keywords

1 0 obj

stream

<stream content>

endstream

endobj

Streams can contain *anything*

Yes, really!
Even binary, other file formats...
(except the **endstream** keyword)

Stream parameters are stored before the stream.

a dictionary
after **obj**, before **stream**
required: stream length
optional: compression algorithm, etc…

Example

1 0 obj
<< /Length 10 >>
stream
0123456789
endstream
endobj

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```



```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

4 0 obj

–

endobj

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

We create
a /Content object...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

4 0 obj

stream

–

endstream

endobj

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...that is a stream object...

Study time

Page contents syntax

parameters sequence then operator

ex: param1 param2 operator

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream

endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

Text objects are delimited by ***BT*** and ***ET***...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream

BT

ET

endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

...(BeginText & EndText).

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream
BT
Tf_
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

We need to set a font, with ***Tf***.

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
stream
BT
_ Tf
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

It takes 2 parameters:
a font name...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
stream
BT
/F1_ Tf
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...(from the page's resources)...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
stream
BT
/F1 100_Tf
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...and a font size.

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream
BT
/F1 100 Tf

ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

We move the cursor...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream
BT
/F1 100 Tf
Td_
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

...with the ***Td*** operator...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream
BT
/F1 100 Tf
_ Td
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...that takes 2 parameters...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
stream
BT
/F1 100 Tf
10 400_Td
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...***x*** and ***y*** coordinates.

(default page size: 612x792)

Study time

Def: literal strings

enclosed in parentheses
Ex: (Hi Mum)

Can contain parentheses

(Hello() World(((()

Can contain white space

(Hello

World !
)

Standard escaping is supported

(Hello \
World \r\n)

Escaping is in *octal*

(Hell**157** World)

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream
BT
/F1 100 Tf
10 400 Td

ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
stream
BT
/F1 100 Tf
10 400 Td
Tj_
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...is done with the ***Tj*** operator...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
stream
BT
/F1 100 Tf
10 400 Td
_ Tj
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...that takes a single parameter...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
stream
BT
/F1 100 Tf
10 400 Td
(_) Tj
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...a literal string.

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream
BT
/F1 100 Tf
10 400 Td
(Hello World_) Tj
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

Our contents stream is complete...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj

<<
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

One last thing...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
<<  _  >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...we need to set its parameters...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
<< /Length_ >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

… the stream length...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

…including white space (new lines characters…).

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj

<< /Length 44 >>

stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

Our stream parameters are finished...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj
```

```
xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...so our page contents object is finished.

Recap:

obj 4 is a stream object with a set length, defining the page's contents:
declare text, set a font and size,
move cursor, display text.

The whole document is defined.
We need to polish the structure.

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

Our PDF defines 4 objects, starting at index 1...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...but PDFs always have an object `0`, that is null...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
%xref table here
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...so 5 objects, starting at `0`.

Warning: offsets & EOLs

We have to define offsets,
which are affected by the EOL conventions:
1 char under Linux/Mac, 2 under Windows.
(I use 1 char newlines character here)

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj
```

xref

```
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

Let’s edit the XREF table!

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref

0_

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

The next line defines the starting index...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref

0 5_

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

...and the number of objects.

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

```
4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj
```

```
xref
0 5
```

```
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

Then, one line per object...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

...following the

xxxxxxxxxx yyyyy a format

(10 digits, 5 digits, 1 letter).

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

The first parameter is the offset (in decimal) of the object...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref

0 5

0000000000_

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

...(for the null object, it's 0).

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 _

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

Then, the generation number (that is almost always 0)...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref

0 5

0000000000 65535_

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

...but for object 0, it's 65535.

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f_
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

Then, a letter, to tell if this entry is free (***f***) or in use (***n***).

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f_
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

Lastly, each line should take 20 bytes, including EOL...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

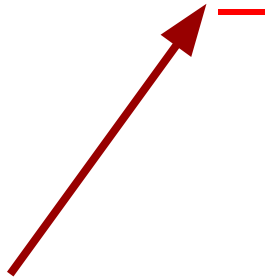
4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref

0 5

0000000000 65535 f

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF



...so add a trailing space.

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

Next line (the first real object)...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010_

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

…object offset, in decimal...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000_
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

…generation number...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n_

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

…and declare the object index in use (***n***)...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

…and the trailing space 

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

Do the same with the other objects...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
00000 n
00000 n
00000 n _

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

…knowing that all lines will end with “ 00000 n ”,...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

...set all offsets.

```
4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj
```

```
xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n

trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

The cross-reference table is finished.

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R >>
startxref
%xref pointer
%%EOF
```

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R >>
startxref

%%EOF
```

We set the startxref pointer...

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R >>
startxref
364_
%%EOF
```

...as xref's offset, in decimal (no prepending 0s).

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R >>
startxref
364
%%EOF
```

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R _ >>
startxref
364
%%EOF
```

We also need to update the trailer dictionary...

%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [3 0 R]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

...with the number of objects...

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R /Size_ >>
startxref
364
%%EOF

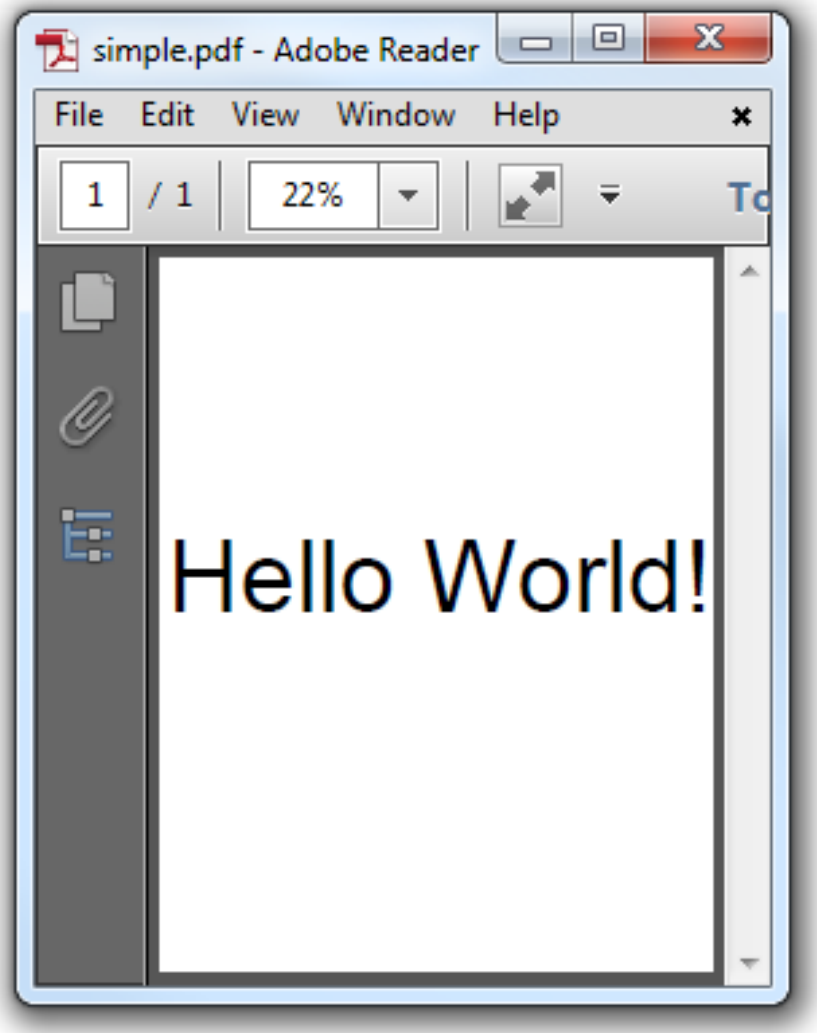
```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj
```

… in the PDF (including object 0).

```
4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n
trailer
<< /Root 1 0 R /Size 5_ >>
startxref
364
%%EOF
```

```
%PDF-1.3

1 0 obj
<< /Type /Catalog /Pages 2 0 R >>
endobj

2 0 obj
<< /Type /Pages
/Kids [ 3 0 R ]
/Count 1 >>
endobj

3 0 obj
<< /Type /Page /Parent 2 0 R
/Resources << /Font << /F1 <<
/Type /Font /Subtype /Type1 /BaseFont /Arial >> >> >>
/Contents 4 0 R
>>
endobj

4 0 obj
<< /Length 44 >>
stream
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
endstream
endobj

xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000060 00000 n
0000000120 00000 n
0000000269 00000 n

trailer
<< /Root 1 0 R /Size 5 >>

startxref
364
%%EOF
```

Our PDF is now complete.



Congratulations!

Disclaimer:

this is a minimal PDF.

Most PDF documents are much bigger, and contain many more elements.

Our PDF:	**A standard generated “Hello World”:**
528 bytes	15 kiloBytes
4 objects	20 objects
text only	text and binary (embedded fonts…)

Hint: use “`mutool clean`” to fix offsets and lengths.

No need to type them yourself!

⇒ mutool version

Slightly different content,
but same rendering.

```
%PDF-1.3
%%µü

1 0 obj
<</Type/Catalog/Pages 2 0 R>>
endobj

2 0 obj
<</Type/Pages/Kids[3 0 R]/Count 1>>
endobj

3 0 obj
<</Type/Page/Parent 2 0 R/Resources 5 0 R/Contents 4 0 R>>
endobj

4 0 obj
<</Length 49>>
stream
q
BT
/F1 100 Tf
10 400 Td
(Hello World!) Tj
ET
Q
endstream
endobj

5 0 obj
<</Font<</F1<</Type/Font/Subtype/Type1/BaseFont/Arial>>>>>>
endobj

xref
0 6
0000000000 65536 f
0000000018 00000 n
0000000064 00000 n
0000000116 00000 n
0000000191 00000 n
0000000288 00000 n

trailer
<</Size 6/Root 1 0 R>>

startxref
364
%%EOF
```

Hint: you can directly extract the PDF sources.

use “`pdftotext --layout`” on the slide deck

http://www.foolabs.com/xpdf/home.html

One more thing...

This one is important for self study.

Def: stream filters

streams can be encoded and/or compressed
algorithms can be cascaded
ex: compression, then ASCII encoding

New stream parameter: /Filter

ex: encode the stream in ASCII

```
1 0 obj
<< /Length 12 >>
stream
Hello World!
endstream
endobj
```

⇔

```
1 0 obj
<< /Length 24 /Filter /ASCIIHexDecode>>
stream
48656C6C6F20576F726C6421
endstream
endobj
```

Ex: compression

(deflate = ZIP compression)

```
1 0 obj
<< /Length 12 >>
stream
Hello World!
endstream
endobj
```

```
1 0 obj
<< /Length 20 /Filter /FlateDecode>>
stream
x£¾H═╔╔¤/╩IQ♦ ∟I♦>
endstream
endobj
```

Filters can be cascaded.

Ex: compressed, then encoded in ASCII

```
1 0 obj
<< /Length 12 >>
stream
Hello World!
endstream
endobj
```

⇔

```
1 0 obj
<< /Length 40 /Filter [/ASCIIHexDecode /FlateDecode] >>
stream
789CF348CDC9C95708CF2FCA495104001C49043E
endstream
endobj
```

Hint: “`mutool clean -d`” to remove any stream filter.

(if you want to explore PDFs by yourself)

http://www.mupdf.com/

Want more?

pdf101.corkami.com

HEADER

```
%PDF-1.1
```
SIGNATURE & VERSION INFORMATION

BODY

DICTIONARY `<< [ID VALUE]* >>`

OBJECT REFERENCE: <OBJECT NUMBER> <REVISION NUMBER> R

IDENTIFIER (WITH /)

ARRAY

STREAM PARAMETERS: LENGTH, COMPRESSION....

STRING

```
1 0 obj
<<
  /Pages 2 0 R
>>
endobj

2 0 obj
<<
  /Type /Pages
  /Count 1
  /Kids [3 0 R]
>>
endobj

3 0 obj
<<
  /Type /Page
  /Contents 4 0 R
  /Parent 2 0 R
  /Resources <<
    /Font <<
      /F1 <<
        /Type /Font
        /Subtype /Type1
        /BaseFont /Arial
      >>
    >>
  >>
>>
endobj

4 0 obj
<< /Length 50 >>
stream
BT
  /F1 110 Tf
  10 400 Td
  (Hello World!)Tj
ET
endstream
endobj
```

BEGIN TEXT
FONT F1 (ARIAL) SET TO SIZE 110
MOVE TO COORDINATE 10, 400
OUTPUT TEXT "HELLO WORLD!"
END TEXT

XREF TABLE

CROSS REFERENCE

```
xref
0 5
0000000000 65535 f
0000000010 00000 n
0000000047 00000 n
0000000111 00000 n
0000000313 00000 n
```

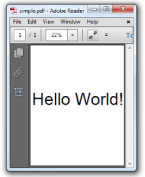
CROSS REFERENCES
5 OBJECTS, STARTING AT INDEX 0
(STANDARD FIRST EMPTY OBJECT 0
OFFSET TO OBJECT 1, REV 0
TO OBJECT 2...
3...
4

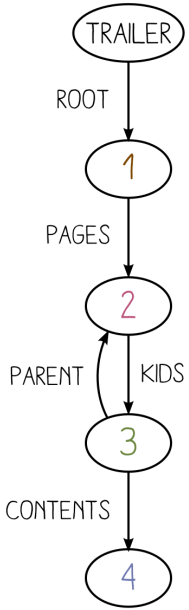
TRAILER

```
trailer
<<
  /Root 1 0 R
>>

startxref
413
%%EOF
```

PARSING

`%PDF-1.?` IS CHECKED
`startxref` POINTS TO XREF
`xref` POINTS TO EACH OBJECT
`trailer` IS PARSED
REFERENCES ARE FOLLOWED
DOCUMENT IS RENDERED





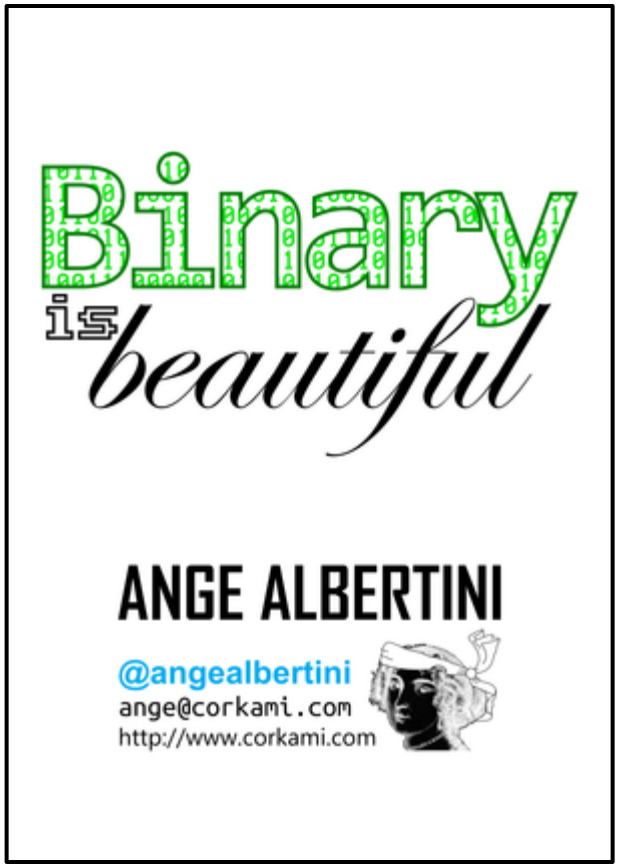
Questions?

(you can download this poster at http://pics.corkami.com)

ACK

@Doegox @ChrisJohnRiley
@PDFKungFoo

To be continued...?



https://leanpub.com/binaryisbeautiful

@angealbertini

corkami.com

Let's write a PDF file *r2*