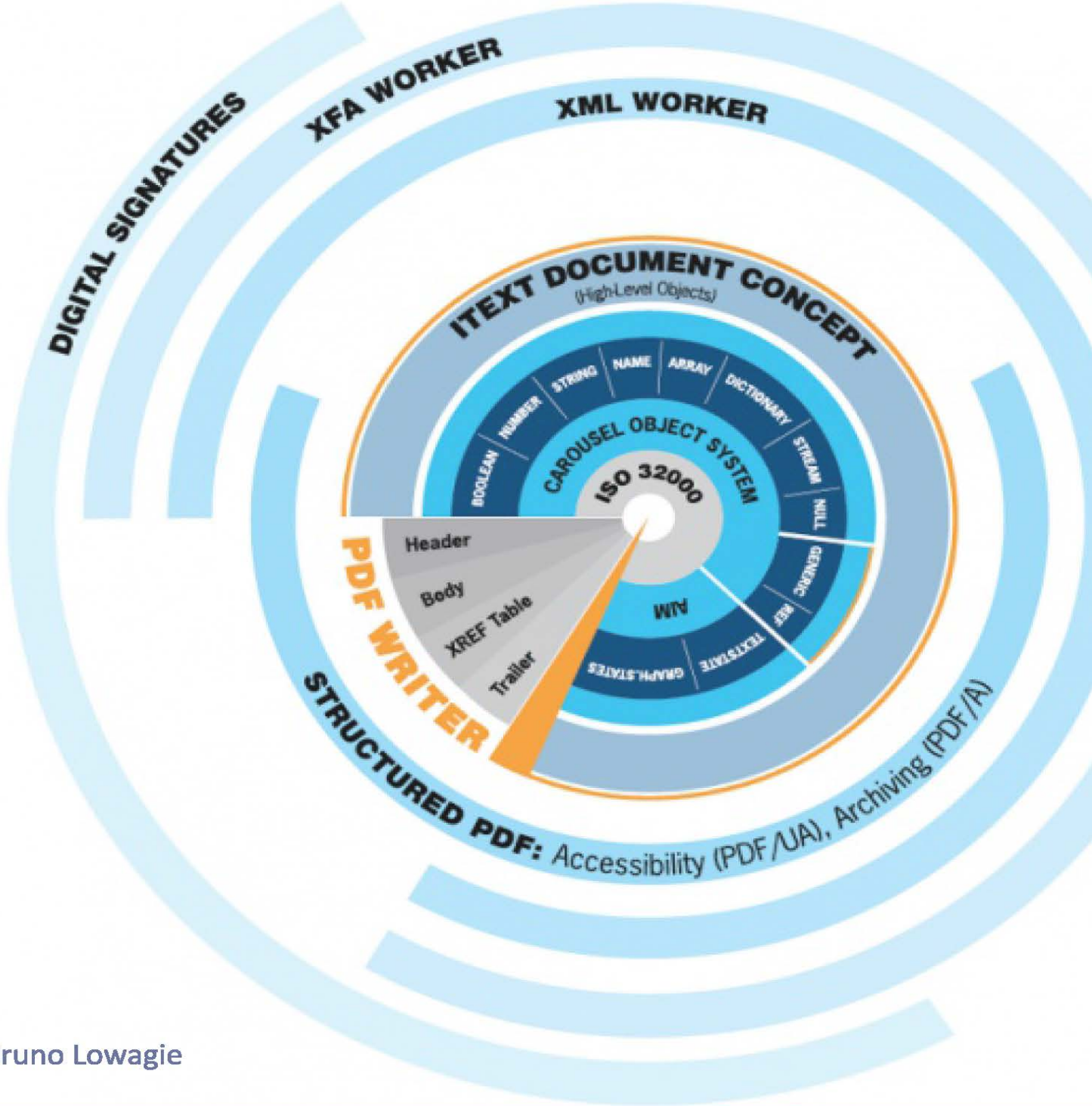
The ABC of PDF with iText

PDF Syntax essentials



by Bruno Lowagie

The ABC of PDF with iText

PDF Syntax essentials

iText Software

This book is for sale at http://leanpub.com/itext_pdfabc

This version was published on 2014-03-01

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Tweet This Book!

Please help iText Software by spreading the word about this book on Twitter!

The suggested tweet for this book is:

@iText: I just bought The ABC of PDF with iText

The suggested hashtag for this book is #itext_pdfabc.

Find out what other people are saying about the book by clicking on this link to search for this hashtag on Twitter:

https://twitter.com/search?q=#itext_pdfabc

Contents

Introduction

This book is a vademecum for the other iText books entitled "Create your PDFs with iText[1]," "Update your PDFs with iText[2]," and "Sign your PDFs with iText[3]."

In the past, I used to refer to ISO-32000 whenever somebody asked me questions such as *"why can't I use PDF as a format for editing documents"* or whenever somebody wanted to use a feature that wasn't supported out-of-the-box.

I soon realized that answering *"read the specs"* is lethal when the specs consist of more than a thousand pages. In this iText tutorial, I'd like to present a short introduction to the syntax of the Portable Document Format. It's not the definitive guide, but it should be sufficient to help you out when facing a PDF-related problem.

You'll find some simple iText examples in this book, but the heavy lifting will be done in the other iText books.

[1] https://leanpub.com/itext_pdfcreate
[2] https://leanpub.com/itext_pdfupdate
[3] https://leanpub.com/itext_pdfsign

I Part 1: The Carousel Object System

The Portable Document Format (PDF) specification, as released by the International Organization for Standardization (ISO) in the form of a series of related standards (ISO-32000-1 and -2, ISO-19005-1, -2, and -3, ISO-14289-1,...), was originally created by Adobe Systems Inc.

Carousel was the original code name for what later became Acrobat. The name Carousel was already taken by Kodak, so a marketing consultant was asked for an alternative name. These were the names that were proposed:

- *Adobe Traverse*– didn't make it,
- *Adobe Express*– sounded nice, but there was already that thing called Quark Express,
- *Adobe Gates*– was never an option, because there was already somebody with that name at another company,
- *Adobe Rosetta*– couldn't be used, because there was an existing company that went by that name.
- *Adobe Acrobat*– was a name not many people liked, but it was chosen anyway.

Although Acrobat exists for more than 20 years now, the name Carousel is still used to refer to the way a PDF file is composed, and that's what the first part of this book is about.

In this first part, we'll:

- Take a look at the basic PDF objects,
- Find out how these objects are organized inside a file, and
- Learn how to read a file by navigating from object to object.

At the end of this chapter, you'll know how PDF is structured and you'll understand what you see when opening a PDF in a text editor instead of inside a PDF viewer.

1 PDF Objects

There are eight basic types of objects in PDF. They're explained in sections 7.3.2 to 7.3.9 of ISO-32000-1.

1.1 The basic PDF objects

These eight objects are implemented in iText as subclasses of the abstract `PdfObject` class. Table 1.1 lists these types as well as their corresponding objects in iText.

Table 1.1: Overview of the basic PDF objects

PDF Object	iText object	Description
Boolean	`PdfBoolean`	This type is similar to the Boolean type in programming languages and can be `true` or `false`.
Numeric object	`PdfNumber`	There are two types of numeric objects: integer and real. Numbers can be used to define coordinates, font sizes, and so on.
String	`PdfString`	String objects can be written in two ways: as a sequence of literal characters enclosed in parentheses `( )` or as hexadecimal data enclosed in angle brackets `< >`. Beginning with PDF 1.7, the type is further qualified as text string, PDFDocEncoded string, ASCII string, and byte string, depending upon how the string is used in each particular context.
Name	`PdfName`	A name object is an atomic symbol uniquely defined by a sequence of characters. Names can be used as keys for a dictionary, to define an explicit destination type, and so on. You can easily recognize names in a PDF file because they're all introduced with a forward slash: `/`.
Array	`PdfArray`	An array is a one-dimensional collection of objects, arranged sequentially between square brackets. For instance, a rectangle is defined as an array of four numbers: `[0 0 595 842]`.
Dictionary	`PdfDictionary`	A dictionary is an associative table containing pairs of objects known as dictionary entries. The key is always a name; the value can be (a reference to) any other object. The collection of pairs is enclosed by double angle brackets: `<<` and `>>`.
Stream	`PdfStream`	Like a string object, a stream is a sequence of bytes. The main difference is that a PDF consumer reads a string entirely, whereas a stream is best read incrementally. Strings are used for small pieces of data; streams are used for large amounts of data.

Table 1.1: Overview of the basic PDF objects

PDF Object	iText object	Description
		Each stream consists of a dictionary followed by zero or more bytes enclosed between the keywords `stream` (followed by a newline) and `endstream`.
Null object	`PdfNull`	This type is similar to the `null` object in programming languages. Setting the value of a dictionary entry to to `null` is equivalent to omitting the entry.

If you look inside iText, you'll find subclasses of these basic PDF implementations created for specific purposes.

- `PdfDate` extends `PdfString` because a date is a special type of string in the Portable Document Format.
- `PdfRectangle` is a special type of `PdfArray`, consisting of four number values: `[llx, lly, urx, ury]` representing the coordinates of the lower-left and upper-right corner of the rectangle.
- `PdfAction`, `PdfFormField`, `PdfOutline` are examples of subclasses of the `PdfDictionary` class.
- `PRStream` is a special implementation of `PdfStream` that needs to be used when extracting a stream from an existing PDF document using `PdfReader`.

When creating or manipulating PDF documents with iText, you'll use high-level objects and convenience methods most of the time. This means you probably won't be confronted with these basic objects very often, but it's interesting to take a look under the hood of iText.

1.2 iText's PdfObject implementations

Let's take a look at some simple code samples for each of the basic types.

1.2.1 PdfBoolean

As there are only two possible values for the `PdfBoolean` object, you can use a static instance instead of creating a new object.

Code sample 1.1: C0101_BooleanObject

```
1  public static void main(String[] args) {
2      showObject(PdfBoolean.PDFTRUE);
3      showObject(PdfBoolean.PDFFALSE);
4  }
5  public static void showObject(PdfBoolean obj) {
6      System.out.println(obj.getClass().getName() + ":");
7      System.out.println("-> boolean? " + obj.isBoolean());
8      System.out.println("-> type: " + obj.type());
9      System.out.println("-> toString: " + obj.toString());
10     System.out.println("-> booleanvalue: " + obj.booleanValue());
11 }
```

In code sample 1.1, we use `PdfBoolean`'s constant values `PDFTRUE` and `PDFFALSE` and we inspect these objects in the `showObject()` method. We get the fully qualified name of the class. We use the `isBoolean()` method that will return `false` for all objects that aren't derived from `PdfBoolean`. And we display the `type()` in the form of an `int` (this value is 1 for `PdfBoolean`).

All `PdfObject` implementations have a `toString()` method, but only the `PdfBoolean` class has a `booleanValue()` method that allows you to get the value as a primitive Java `boolean` value.

The output of the `showObject` method looks like this:

```
com.itextpdf.text.pdf.PdfBoolean:
-> boolean? true
-> type: 1
-> toString: true
-> booleanvalue: true
com.itextpdf.text.pdf.PdfBoolean:
-> boolean? true
-> type: 1
-> toString: false
-> booleanvalue: false
```

We'll use the `PdfBoolean` object in the tutorial Update your PDFs with iText[1] when we'll update properties of dictionaries to change the behavior of a PDF feature.

1.2.2 PdfNumber

There are many different ways to create a `PdfNumber` object. Although PDF only has two types of numbers (integer and real), you can create a `PdfNumber` object using a `String`, `int`, `long`, `double` or `float`.

This is shown in code sample 1.2.

Code sample 1.2: C0102_NumberObject

```
1  public static void main(String[] args) {
2      showObject(new PdfNumber("1.5"));
3      showObject(new PdfNumber(100));
4      showObject(new PdfNumber(1001));
5      showObject(new PdfNumber(1.5));
6      showObject(new PdfNumber(1.5f));
7  }
8  public static void showObject(PdfNumber obj) {
9      System.out.println(obj.getClass().getName() + ":");
10     System.out.println("-> number? " + obj.isNumber());
11     System.out.println("-> type: " + obj.type());
12     System.out.println("-> bytes: " + new String(obj.getBytes()));
```

[1] https://leanpub.com/itext_pdfupdate

```
13      System.out.println("-> toString: " + obj.toString());
14      System.out.println("-> intValue: " + obj.intValue());
15      System.out.println("-> longValue: " + obj.longValue());
16      System.out.println("-> doubleValue: " + obj.doubleValue());
17      System.out.println("-> floatValue: " + obj.floatValue());
18  }
```

Again we display the fully qualified classname. We check for number objects using the `isNumber()` method. And we get a different value when we asked for the type (more specifically: 2).

The `getBytes()` method returns the bytes that will be stored in the PDF. In the case of numbers, you'll get a similar result using `toString()` method. Although iText works with `float` objects internally, you can get the value of a `PdfNumber` object as a primitive Java `int`, `long`, `double` or `float`.

```
com.itextpdf.text.pdf.PdfNumber:
-> number? true
-> type: 2
-> bytes: 1.5
-> toString: 1.5
-> intValue: 1
-> longValue: 1
-> doubleValue: 1.5
-> floatValue: 1.5
com.itextpdf.text.pdf.PdfNumber:
-> number? true
-> type: 2
-> bytes: 100
-> toString: 100
-> intValue: 100
-> longValue: 100
-> doubleValue: 100.0
-> floatValue: 100.0
```

Observe that you lose the decimal part if you invoke the `intValue()` or `longValue()` method on a real number. Just like with `PdfBoolean`, you'll use `PdfNumber` only if you hack a PDF at the lowest level, changing a property in the syntax of an existing PDF.

1.2.3 PdfString

The `PdfString` class has four constructors:

- An empty constructor in case you want to create an empty `PdfString` object (in practice this constructor is only used in subclasses of `PdfString`),
- A constructor that takes a Java `String` object as its parameter,

- A constructor that takes a Java `String` object as well as the encoding value (`TEXT_PDFDOCENCODING` or `TEXT_UNICODE`) as its parameters,
- A constructor that takes an array of bytes as its parameter in which case the encoding will be `PdfString.NOTHING`. This method is used by iText when reading existing documents into PDF objects.

You can choose to store the PDF string object in hexadecimal format by using the `setHexWriting()` method:

Code sample 1.3: C0103_StringObject

```
1  public static void main(String[] args) {
2      PdfString s1 = new PdfString("Test");
3      PdfString s2 = new PdfString("\u6d4b\u8bd5", PdfString.TEXT_UNICODE);
4      showObject(s1);
5      showObject(s2);
6      s1.setHexWriting(true);
7      showObject(s1);
8      showObject(new PdfDate());
9  }
10 public static void showObject(PdfString obj) {
11     System.out.println(obj.getClass().getName() + ":");
12     System.out.println("-> string? " + obj.isString());
13     System.out.println("-> type: " + obj.type());
14     System.out.println("-> bytes: " + new String(obj.getBytes()));
15     System.out.println("-> toString: " + obj.toString());
16     System.out.println("-> hexWriting: " + obj.isHexWriting());
17     System.out.println("-> encoding: " + obj.getEncoding());
18     System.out.println("-> bytes: " + new String(obj.getOriginalBytes()));
19     System.out.println("-> unicode string: " + obj.toUnicodeString());
20 }
```

In the output of code sample 1.3, we see the fully qualified name of the class. The `isString()` method returns `true`. The type value is 3. In this case, the `toBytes()` method can return a different value than the `toString()` method. The `String` `"\u6d4b\u8bd5"` represents two Chinese characters meaning “test”, but these characters are stored as four bytes.

Hexademical writing is applied at the moment the bytes are written to a PDF `OutputStream`. The encoding values are stored as `String` values, either `"PDF"` for PdfDocEncoding, `"UnicodeBig"` for Unicode, or `""` in case of a pure byte string.

The `getOriginalBytes()` method only makes sense when you get a `PdfString` value from an existing file that was encrypted. It returns the original encrypted value of the string object.

The `toUnicodeString()` method is a safer method than `toString()` to get the PDF string object as a Java `String`.

```
com.itextpdf.text.pdf.PdfString:
-> string? true
-> type: 3
-> bytes: Test
-> toString: Test
-> hexWriting: false
-> encoding: PDF
-> original bytes: Test
-> unicode string: Test
com.itextpdf.text.pdf.PdfString:
-> string? true
-> type: 3
-> bytes: □□mK□□
-> toString: □□
-> hexWriting: false
-> encoding: UnicodeBig
-> original bytes: □□mK□□
-> unicode string: □□
com.itextpdf.text.pdf.PdfString:
-> string? true
-> type: 3
-> bytes: Test
-> toString: Test
-> hexWriting: true
-> encoding: PDF
-> original bytes: Test
-> unicode string: Test
com.itextpdf.text.pdf.PdfDate:
-> string? true
-> type: 3
-> bytes: D:20130430161855+02'00'
-> toString: D:20130430161855+02'00'
-> hexWriting: false
-> encoding: PDF
-> original bytes: D:20130430161855+02'00'
-> unicode string: D:20130430161855+02'00'
```

In this example, we also create a `PdfDate` instance. If you don't pass a parameter, you get the current date and time. You can also pass a Java `Calendar` object if you want to create an object for a specific date. The format of the date conforms to the international Abstract Syntax Notation One (ASN.1) standard defined in ISO/IEC 8824. You recognize the pattern `YYYYMMDDHHmmSSOHH' mm` where `YYYY` is the year, `MM` the month, `DD` the day, `HH` the hour, `mm` the minutes, `SS` the seconds, `OHH` the relationship to Universal Time (UT), and `' mm` the offset from UT in minutes.

1.2.4 PdfName

There are different ways to create a `PdfName` object, but you should only use one. The constructor that takes a single `String` as a parameter guarantees that your name object conforms to ISO-32000-1 and -2.

You probably wonder why we would add constructors that allow people names that don't conform with the PDF specification. With iText, we did a great effort to ensure the creation of documents that comply. Unfortunately, this can't be said about all PDF creation software. We need some `PdfName` constructors that accept any kind of value when reading names in documents that are in violation with the PDF ISO standards.

In many cases, you don't need to create a `PdfName` object yourself. The `PdfName` object contains a large set of constants with predefined names. One of these names is used in code sample 1.4.

Code sample 1.4: C0104_NameObject

```
public static void main(String[] args) {
    showObject(PdfName.CONTENTS);
    showObject(new PdfName("CustomName"));
    showObject(new PdfName("Test #1 100%"));
}
public static void showObject(PdfName obj) {
    System.out.println(obj.getClass().getName() + ":");
    System.out.println("-> name? " + obj.isName());
    System.out.println("-> type: " + obj.type());
    System.out.println("-> bytes: " + new String(obj.getBytes()));
    System.out.println("-> toString: " + obj.toString());
}
```

The `getClass().getName()` part no longer has secrets for you. We use `isName()` to check if the object is really a name. The type is 4. And we can get the value as bytes or as a `String`.

```
com.itextpdf.text.pdf.PdfName:
-> name? true
-> type: 4
-> bytes: /Contents
-> toString: /Contents
com.itextpdf.text.pdf.PdfName:
-> name? true
-> type: 4
-> bytes: /CustomName
-> toString: /CustomName
com.itextpdf.text.pdf.PdfName:
-> name? true
-> type: 4
```

```
-> bytes: /Test#20#231#20100#25
-> toString: /Test#20#231#20100#25
```

Note that names start with a forward slash, also know as a *solidus*. Also take a closer look at the name that was created with the `String` value `"Test #1 100%"`. iText has escaped values such as `' '`, `'#'` and `'%'` because these are forbidden in a PDF name object. ISO-32000-1 and -2 state that a name is a sequence of 8-bit values and iText's interprets this literally. If you pass a string containing multibyte characters (characters with a value greater than 255), iText will only take the lower 8 bits into account. Finally, iText will throw an `IllegalArgumentException` if you try to create a name that is longer than 127 bytes.

1.2.5 PdfArray

The `PdfArray` class has six constructors. You can create a `PdfArray` using an `ArrayList` of `PdfObject` instances, or you can create an empty array and add the `PdfObject` instances one by one (see code sample 1.5). You can also pass a byte array of `float` or `int` values as parameter in which case you create an array consisting of `PdfNumber` objects. Finally you can create an array with a single object if you pass a `PdfObject`, but be carefull: if this object is of type `PdfArray`, you're using the copy constructor.

Code sample 1.5: C0105_ArrayObject

```
1  public static void main(String[] args) {
2      PdfArray array = new PdfArray();
3      array.add(PdfName.FIRST);
4      array.add(new PdfString("Second"));
5      array.add(new PdfNumber(3));
6      array.add(PdfBoolean.PDFFALSE);
7      showObject(array);
8      showObject(new PdfRectangle(595, 842));
9  }
10 public static void showObject(PdfArray obj) {
11     System.out.println(obj.getClass().getName() + ":");
12     System.out.println("-> array? " + obj.isArray());
13     System.out.println("-> type: " + obj.type());
14     System.out.println("-> toString: " + obj.toString());
15     System.out.println("-> size: " + obj.size());
16     System.out.print("-> Values:");
17     for (int i = 0; i < obj.size(); i++) {
18         System.out.print(" ");
19         System.out.print(obj.getPdfObject(i));
20     }
21     System.out.println();
22 }
```

Once more, we see the fully qualified name in the output. The `isArray()` method tests if this class is a `PdfArray`. The value of the array type is 5.

The elements of the array are stored in an `ArrayList`. The `toString()` method of the `PdfArray` class returns the `toString()` output of this `ArrayList`: the values of the separate objects delimited with a comma and enclosed by square brackets. The `getBytes()` method returns `null`.

You can ask a `PdfArray` for its size, and use this size to get the different elements of the array one by one. In this case, we use the `getPdfObject()` method. We'll discover some more methods to retrieve elements from an array in section 1.3.

```
com.itextpdf.text.pdf.PdfArray:
-> array? true
-> type: 5
-> toString: [/First, Second, 3, false]
-> size: 4
-> Values: /First Second 3 false
com.itextpdf.text.pdf.PdfRectangle:
-> array? true
-> type: 5
-> toString: [0, 0, 595, 842]
-> size: 4
-> Values: 0 0 595 842
```

In our example, we created a `PdfRectangle` using only two values `595` and `842`. However, a rectangle needs four values: two for the coordinate of the lower-left corner, two for the coordinate of the upper-right corner. As you can see, iText added two zeros for the coordinate of the lower-left coordinate.

1.2.6 PdfDictionary

There are only two constructors for the `PdfDictionary` class. With the empty constructor, you can create an empty dictionary, and then add entries using the `put()` method. The constructor that accepts a `PdfName` object will create a dictionary with a `/Type` entry and use the name passed as a parameter as its value. This entry identifies the type of object the dictionary describes. In some cases, a `/SubType` entry is used to further identify a specialized subcategory of the general type.

In code sample 1.6, we create a custom dictionary and an action.

Code sample 1.6: C0106_DictionaryObject

```
1  public static void main(String[] args) {
2      PdfDictionary dict = new PdfDictionary(new PdfName("Custom"));
3      dict.put(new PdfName("Entry1"), PdfName.FIRST);
4      dict.put(new PdfName("Entry2"), new PdfString("Second"));
5      dict.put(new PdfName("3rd"), new PdfNumber(3));
6      dict.put(new PdfName("Fourth"), PdfBoolean.PDFFALSE);
7      showObject(dict);
8      showObject(PdfAction.gotoRemotePage("test.pdf", "dest", false, true));
9  }
```

```
10  public static void showObject(PdfDictionary obj) {
11      System.out.println(obj.getClass().getName() + ":");
12      System.out.println("-> dictionary? " + obj.isDictionary());
13      System.out.println("-> type: " + obj.type());
14      System.out.println("-> toString: " + obj.toString());
15      System.out.println("-> size: " + obj.size());
16      for (PdfName key : obj.getKeys()) {
17          System.out.print(" " + key + ": ");
18          System.out.println(obj.get(key));
19      }
20  }
```

The `showObject()` method shows us the fully qualified names. The `isDictionary()` returns `true` and the `type()` method returns 6.

Just like with `PdfArray`, the `getBytes()` method returns `null`. iText stores the objects in a `HashMap`. The `toString()` method of a `PdfDictionary` doesn't reveal anything about the contents of the dictionary, except for its type if present. The type entry is usually optional. For instance: the `PdfAction` dictionary we created in code sample 1.6 doesn't have a `/Type` entry.

We can ask a dictionary for its number of entries using the `size()` method and get each value as a `PdfObject` by its key. As the entries are stored in a `HashMap`, the keys aren't shown in the same order we used to add them to the dictionary. That's not a problem. The order of entries in a dictionary is irrelevant.

```
com.itextpdf.text.pdf.PdfDictionary:
-> dictionary? true
-> type: 6
-> toString: Dictionary of type: /Custom
-> size: 4
 /3rd: 3
 /Entry1: /First
 /Type: /Custom
 /Fourth: false
 /Entry2: Second
com.itextpdf.text.pdf.PdfAction:
-> dictionary? true
-> type: 6
-> toString: Dictionary
-> size: 4
 /D: dest
 /F: test.pdf
 /S: /GoToR
 /NewWindow: true
```

As explained in table 1.1, a PDF dictionary is stored as a series of key value pairs enclosed by << and >>. The action created in code sample 1.6 looks like this when viewed in a plain text editor:

```
<</D(dest)/F(test.pdf)/S/GoToR/NewWindow true>>
```

The basic `PdfDictionary` object has plenty of subclasses such as `PdfAction`, `PdfAnnotation`, `PdfCollection`, `PdfGState`, `PdfLayer`, `PdfOutline`, etc. All these subclasses serve a specific purpose and they were created to make it easier for developers to create objects without having to worry too much about the underlying structures.

1.2.7 PdfStream

The `PdfStream` class also extends the `PdfDictionary` object. A stream object always starts with a dictionary object that contains at least a `/Length` entry of which the value corresponds with the number of stream bytes.

For now, we'll only use the constructor that accepts a `byte[]` as parameter. The other constructor involves a `PdfWriter` instance, which is an object we haven't discussed yet. Although that constructor is mainly for internal use —it offers an efficient, memory friendly way to write byte streams of unknown length to a PDF document—, we'll briefly cover this alternative constructor in the Create your PDFs with iText[2] tutorial.

Code sample 1.7: C0107_StreamObject

```
1  public static void main(String[] args) {
2      PdfStream stream = new PdfStream(
3       "Long stream of data stored in a FlateDecode compressed stream object"
4       .getBytes());
5      stream.flateCompress();
6      showObject(stream);
7  }
8  public static void showObject(PdfStream obj) {
9      System.out.println(obj.getClass().getName() + ":");
10     System.out.println("-> stream? " + obj.isStream());
11     System.out.println("-> type: " + obj.type());
12     System.out.println("-> toString: " + obj.toString());
13     System.out.println("-> raw length: " + obj.getRawLength());
14     System.out.println("-> size: " + obj.size());
15     for (PdfName key : obj.getKeys()) {
16         System.out.print(" " + key + ": ");
17         System.out.println(obj.get(key));
18     }
19 }
```

In the lines following the fully qualified name, we see that the `isStream()` method returns `true` and the `type()` method returns 7. The `toString()` method returns nothing more than the word `"Stream"`.

[2] https://leanpub.com/itext_pdfcreate

We can store the long `String` we used in code sample 1.7 "as is" inside the stream. In this case, invoking the `getBytes()` method will return the bytes you used in the constructor.

If a stream is compressed, for instance by using the `flateCompress()` method, the `getBytes()` method will return `null`. In this case, the bytes are stored inside a `ByteArrayOutputStream` and you can write these bytes to an `OutputStream` using the `writeContent()` method. We didn't do that because it doesn't make much sense for humans to read a compressed stream.

The `PdfStream` instance remembers the original length aka the raw length. The length of the compressed stream is stored in the dictionary.

```
com.itextpdf.text.pdf.PdfStream:
-> stream? true
-> type: 7
-> toString: Stream
-> raw length: 68
-> size: 2
 /Filter: /FlateDecode
 /Length: 67
```

In this case, compression didn't make much sense: 68 bytes were compressed into 67 bytes. In theory, you could choose a different compression level. The `PdfStream` class has different constants such as `NO_COMPRESSION` (0), `BEST_SPEED` (1) and `BEST_COMPRESSION` (9). In practice, we'll always use `DEFAULT_COMPRESSION` (-1).

1.2.8 PdfNull

We're using the `PdfNull` class internally in some very specific cases, but there's very little chance you'll ever need to use this class in your own code. For instance: it's better to remove an entry from a dictionary than to set its value to `null`; it saves the PDF consumer processing time when parsing the files you've created.

Code sample 1.8: C0108_NullObject

```
1  public static void main(String[] args) {
2      showObject(PdfNull.PDFNULL);
3  }
4  public static void showObject(PdfNull obj) {
5      System.out.println(obj.getClass().getName() + ":");
6      System.out.println("-> type: " + obj.type());
7      System.out.println("-> bytes: " + new String(obj.getBytes()));
8      System.out.println("-> toString: " + obj.toString());
9  }
```

The output of code sample 1.8 is pretty straight-forward: the fully qualified name of the class, its type (8) and the output of the `getBytes()` and `toString()` methods.

```
com.itextpdf.text.pdf.PdfNull:
-> type: 8
-> bytes: null
-> toString: null
```

These were the eight basic types, numbered from 1 to 8. Two more numbers are reserved for specific `PdfObject` classes: 0 and 10. Let's start with the class that returns 0 when you call the `type()` method.

1.2.9 PdfLiteral

The objects we've discussed so far were literally the first objects that were written when I started writing iText. Since 2000, they've been used to build billions of PDF documents. They form the foundation of iText's object-oriented approach to create PDF documents.

Working in an object-oriented way is best practice and it's great, but for some straight-forward objects, you wish you'd have a short-cut. That's why we created `PdfLiteral`. It's an iText object you won't find in the PDF specification or ISO-32000-1 or -2. It allows you to create any type of object with a minimum of overhead.

For instance: we often need an array that defines a specific matrix, called the identity matrix. It consists of six elements: 1, 0, 0, 1, 0 and 0. Should we really create a `PdfArray` object and add these objects one by one? Wouldn't it be easier if we just created the literal array: `[1 0 0 1 0 0]`?

That's what `PdfLiteral` is about. You create the object passing a `String` or a `byte[]`; you can even pass the object type to the constructor.

Code sample 1.9: C0109_LiteralObject

```
1  public static void main(String[] args) {
2      showObject(PdfFormXObject.MATRIX);
3      showObject(new PdfLiteral(
4          PdfObject.DICTIONARY, "<</Type/Custom/Contents [1 2 3]>>"));
5  }
6  public static void showObject(PdfObject obj) {
7      System.out.println(obj.getClass().getName() + ":");
8      System.out.println("-> type: " + obj.type());
9      System.out.println("-> bytes: " + new String(obj.getBytes()));
10     System.out.println("-> toString: " + obj.toString());
11 }
```

The `MATRIX` constant used in code sample 1.9 was created like this: `new PdfLiteral("[1 0 0 1 0 0]");` when we write this object to a PDF, it is treated in exactly the same way as if we'd had created a `PdfArray`, except that its type is 0 because `PdfLiteral` doesn't parse the `String` to check the type.

We also create a custom dictionary, telling the object its type is `PdfObject.DICTIONARY`. This doesn't have any impact on the fully qualified name. As the `String` passed to the constructor isn't being parsed, you can't ask the dictionary for its size nor get the key set of the entries.

The content is stored *literally*, as indicated in the name of the class: `PdfLiteral`.

```
com.itextpdf.text.pdf.PdfLiteral:
-> type: 0
-> bytes: [1 0 0 1 0 0]
-> toString: [1 0 0 1 0 0]
com.itextpdf.text.pdf.PdfLiteral:
-> type: 6
-> bytes: <</Type/Custom/Contents [1 2 3]>>
-> toString: <</Type/Custom/Contents [1 2 3]>>
```

It goes without saying that you should be very careful when using this object. As iText doesn't parse the content to see if its syntax is valid, you'll have to make sure you don't make any mistakes. We use this object internally as a short-cut, or when we encounter content that can't be recognized as being one of the basic types whilst reading an existing PDF file.

1.3 The difference between direct and indirect objects

To explain what the iText `PdfObject` with value `10` is about, we need to introduce the concept of indirect objects. So far, we've been working with direct objects. For instance: you create a dictionary and you add an entry that consists of a PDF name and a PDF string. The result looks like this:

```
<</Name (Bruno Lowagie)>>
```

The string value with my name is a *direct object*, but I could also create a PDF string and label it:

```
1 0 obj
(Bruno Lowagie)
endobj
```

This is an *indirect object* and we can refer to it from other objects, for instance like this:

```
<</Name 1 0 R>>
```

This dictionary is equivalent to the dictionary that used a direct object for the string. The `1 0 R` in the latter dictionary is called an *indirect reference*, and its iText implementation is called `PdfIndirectReference`. The type value is `10` and you can check if a `PdfObject` is in fact an indirect reference using the `isIndirect()` method.

A stream object may never be used as a direct object. For example, if the value of an entry in a dictionary is a stream, that value always has to be an indirect reference to an indirect object containing a stream. A stream dictionary can never be an indirect object. It always has to be a direct object.

An indirect reference can refer to an object of any type. We'll find out how to obtain the actual object referred to by an indirect reference in chapter 3.

1.4 Summary

In this chapter, we've had an overview of the building blocks of a PDF file:

- boolean,
- number,
- string,
- name,
- array,
- dictionary,
- stream, and
- null

Building blocks can be organized as numbered indirect objects that reference each other.

It's difficult to introduce code samples explaining how direct and indirect objects interact, without seeing the larger picture. So without further ado, let's take a look at the file structure of a PDF document.

2 PDF File Structure

Figure 2.1 shows a simple, single-page PDF document with the text “Hello World” opened in Adobe Reader.

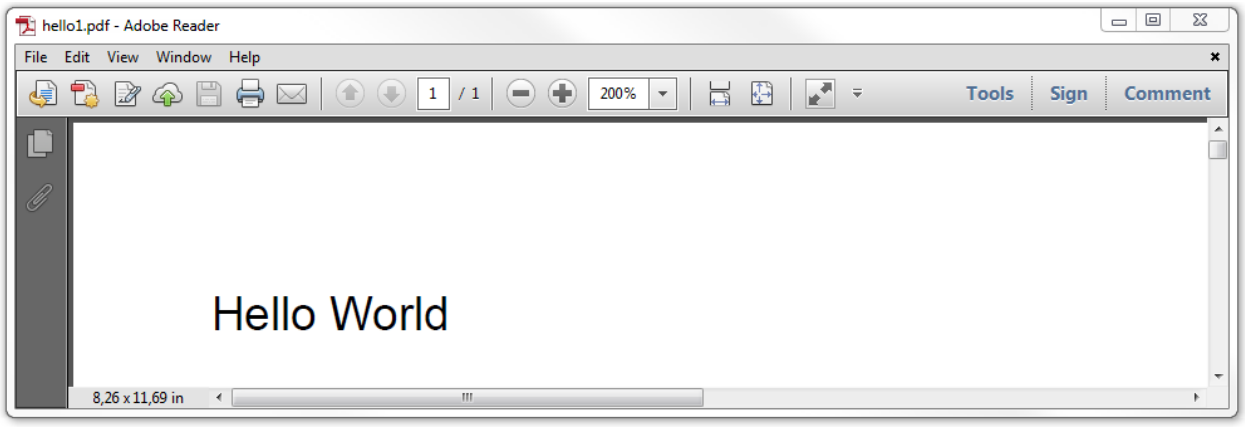


Figure 2.1: Hello World

Now let’s open the file in a text editor and examine its internal structure.

2.1 The internal structure of a PDF file

When we open the “Hello World” document in a plain text editor instead of in a PDF viewer, we soon discover that a PDF file consists of a sequence of indirect objects as described in the previous chapter.

Table 2.1 shows how to find the four different parts that define the “Hello World” document listed in code sample 2.1:

Table 2.1: Overview of the parts of a PDF file

Part	Name	Line numbers
1	The Header	Lines 1-2
2	The Body	Lines 3-24
3	The Cross-reference Table	Lines 25-33
4	The Trailer	Lines 34-40

Note that I’ve replaced a binary content stream by the words `*binary stuff*`. Lines that were too long to fit on the page were split; a \ character marks where the line was split.

Code sample 2.1: A PDF file inside-out

```
%PDF-1.4
%âãÏÓ
2 0 obj
<</Length 64/Filter/FlateDecode>>stream
*binary stuff*
endstream
endobj
4 0 obj
<</Parent 3 0 R/Contents 2 0 R/Type/Page/Resources<</ProcSet [/PDF /Text /ImageB /ImageC /\
ImageI]/Font<</F1 1 0 R>>>>/MediaBox[0 0 595 842]>>
endobj
1 0 obj
<</BaseFont/Helvetica/Type/Font/Encoding/WinAnsiEncoding/Subtype/Type1>>
endobj
3 0 obj
<</Type/Pages/Count 1/Kids[4 0 R]>>
endobj
5 0 obj
<</Type/Catalog/Pages 3 0 R>>
endobj
6 0 obj
<</Producer(iText® 5.4.2 ©2000-2012 1T3XT BVBA \(AGPL-version\))/ModDate(D:20130502165150+\
02'00')/CreationDate(D:20130502165150+02'00')>>
endobj
xref
0 7
0000000000 65535 f
0000000302 00000 n
0000000015 00000 n
0000000390 00000 n
0000000145 00000 n
0000000441 00000 n
0000000486 00000 n
trailer
<</Root 5 0 R/ID [<91bee3a87061eb2834fb6a3258bf817e><91bee3a87061eb2834fb6a3258bf817e>]/In\
fo 6 0 R/Size 7>>
%iText-5.4.2
startxref
639
%%EOF
```

Let's examine the four parts that are present in code sample 2.1 one by one.

2.1.1 The Header

Every PDF file starts with `%PDF-`. If it doesn't, a PDF consumer will throw an error and refuse to open the file because it isn't recognized as a valid PDF file. For instance: iText will throw an `InvalidPdfException` with the message *"PDF header signature not found."*

iText supports the most recent PDF specifications, but uses version 1.4 by default. That's why our "Hello World" example (that was created using iText) starts with `%PDF-1.4`.

Beginning with PDF 1.4, the PDF version can also be stored elsewhere in the PDF. More specifically in the *root* object of the document, aka the *catalog*. This implies that a file with header `%PDF-1.4` can be seen as a PDF 1.7 file if it's defined that way in the document root. This allows the version to be changed in an incremental update without changing the original header.

The second line in the header needs to be present if the PDF file contains binary data (which is usually the case). It consists of a percent sign, followed by at least four binary characters. That is: characters whose codes are 128 or greater. This ensures proper behavior of the file transfer applications that inspect data near the beginning of a file to determine whether to treat the file's contents as a text file, or as a binary file.

Line 1 and 2 start with a percent sign (%). Any occurence of this sign outside a string or stream introduces a comment. Such a comment consists of all characters after the percent sign up to (but not including) the End-of-Line marker. Except for the header lines discussed in this section and the End-of-File marker `%%EOF`, comments are ignored by PDF readers because they have no semantical meaning,

The Body of the document starts on the third line.

2.1.2 The Body

We recognize six indirect objects between line 3 and 24 in code sample 2.1. They aren't ordered sequentially:

1. Object 2 is a stream,
2. Object 4 is a dictionary of type `/Page`,
3. Object 1 is a dictionary of type `/Font`,
4. Object 3 is a dictionary of type `/Pages`,
5. Object 5 is a dictionary of type `/Catalog`, and
6. Object 6 is a dictionary for which no type was defined.

A PDF producer is free to add these objects in any order it desires. A PDF consumer will use the cross-reference table to find each object.

2.1.3 The Cross-reference Table

The cross-reference table starts with the keyword `xref` and contains information that allows access to the indirect objects in the body. For reasons of performance, a PDF consumer doesn't read the entire file.

Imagine a document with 10,000 pages. If you only want to see the last page, a PDF viewer doesn't need to read the content of the 9,999 previous pages. It can use the cross-reference table to retrieve only those objects needed as a resource for the requested page.

The keyword `xref` is followed by a sequence of lines that either consist of two numbers, or of exactly 20 bytes. In code sample 2.1, the cross-reference table starts with `0 7`. This means the next line is about object 0 in a series of seven consecutive objects: 0, 1, 2, 3, 4, 5, and 6.

There can be gaps in a cross-reference table. For instance, an additional line could be `10 3` followed by three lines about objects 10, 11, and 12.

The lines with exactly 20 bytes consist of three parts separated by a space character:

1. a 10-digit number representing the byte offset,
2. a 5-digit number indicates the generation of the object,
3. a keyword, either `n` if the object is *in use*, or `f` if the object is *free*.

Each of these lines ends with a 2-byte End-of-Line sequence.

The first entry in the cross-reference table representing object 0 at position 0 is always a free object with the highest possible generation number: 65,535. In code sample 2.1, it is followed by 6 objects that are in use: object 1 starts at byte position 302, object 2 at position 15, and so on.

Since PDF 1.5, there's another, more compact way to create a cross-reference table, but let's first take a look at the final part of the PDF file in code sample 2.1, the trailer.

2.1.4 The Trailer

The trailer starts with the keyword `trailer`, followed by the *trailer dictionary*. The trailer dictionary in line 35-36 of code sample 2.1 consists of four entries:

- The `/ID` entry is a file identifier consisting of an array of two byte sequences. It's only required for encrypted documents, but it's good practice to have them because some workflows depend on each document to be uniquely identified (this implies that no two files use the same identifier). For documents created from scratch, the two parts of the identifier should be identical.
- The `/Size` entry shows the total number of entries in the file's cross-reference table, in this case 7.
- The `/Root` entry refers to object 5. This is a dictionary of type `/Catalog`. This root object contains references to other objects defining the content. The Catalog dictionary is the starting point for PDF consumers that want to read the contents of a document.

- The `/Info` entry refers to object 6. This is the info dictionary. This dictionary can contain metadata such as the title of the document, its author, some keywords, the creation date, etc. This object will be deprecated in favor of XMP metadata in the next PDF version (PDF 2.0 defined in ISO-32000-2).

Other possible entries in the trailer dictionary are the `/Encrypt` key, which is required if the document is encrypted, and the `/Prev` key, which is present if the file has more than one cross-reference section. This will occur in the case of PDFs that are updated in append mode as will be explained in section 2.2.1.

Every PDF file ends with three lines consisting of the keyword `startxref`, a byte position, and the keyword `%%EOF`. In the case of code sample 2.1, the byte position points to the location of the `xref` keyword of the most recent cross-reference table.

Let's take a look at some variations on this file structure.

2.2 Variations on the file structure

Depending on the document requirements of your project, you'll expect a slightly different structure:

- When a document is updated and the bytes of the previous revision need to remain intact,
- When a document is postprocessed to allow fast web access, or
- When file size is important and therefore full compression is recommended.

Let's take a look at the possible impact of these requirements on the file structure.

2.2.1 PDFs with more than one cross-reference table

There are different ways to update the contents of a PDF document. One could take the objects of an existing PDF, apply some changes by adding and removing objects, and creating a new structure where the existing objects are reordered and renumbered. That's the default behavior of iText's `PdfStamper` class.

In some cases, this behavior isn't acceptable. If you want to add an extra signature to a document that was already signed, changing the structure of the existing document will break the original signature. You'll have to preserve the bytes of the original document and add new objects, a new cross-reference table and a new trailer. The same goes for *Reader enabled* files, which are files signed using Adobe's private key, adding specific usage rights to the file.

Code sample 2.2 shows three extra parts that can be added to code sample 2.1 (after line 40): an extra body, an extra cross-reference table and an extra trailer. This is only a simple example of a possible update to an existing PDF document; no extra visible content was added. We'll see a more complex example in the tutorial Sign your PDFs with iText[1].

[1] https://leanpub.com/itext_pdfsign

Code sample 2.2: A PDF file inside-out (part 2)

```
6 0 obj
<</Producer(iText® 5.4.2 ©2000-2012 1T3XT BVBA \(AGPL-version\))/ModDate(D:20130502165150+\
02'00')/CreationDate(D:20130502165150+02'00')>>
endobj
xref
0 1
0000000000 65535 f
6 1
0000000938 00000 n
trailer
<</Root 5 0 R/Prev 639/ID [<91bee3a87061eb2834fb6a3258bf817e><84c1b02d932693e4927235c277cc\
489e>]/Info 6 0 R/Size 7>>
%iText-5.4.2
startxref
1091
%%EOF
```

When we look at the new cross-reference table, we see that object 0 is again a free object, whereas object 6 is now updated.

Object 6 is reused and therefore the generation number doesn't need to be incremented. It remains `00000`. In practice, the generation number is only incremented if the status of an object changes from `n` to `f`.

Observe that the `/Prev` key in the trailer dictionary refers to the byte position where the previous cross-reference starts.

The first element of the `/ID` array generally remains the same for a given document. This helps Enterprise Content Management (ECM) systems to detect different versions of the same document. They shouldn't rely on it, though, as not all PDF processors support this feature. For instance: iText's `PdfStamper` will respect the first element of the ID array; `PdfCopy` typically won't because there's usually more than one document involved when using `PdfCopy`, in which case it doesn't make sense to prefer the identifier of one document over the identifier of another.

The file parts shown in code sample 2.2 are an incremental update. All changes are appended to the end of the file, leaving its original contents intact. One document can have many incremental updates.

The principle of having multiple cross-reference streams is also used in the context of linearization.

2.2.2 Linearized PDFs

A linearized PDF file is organized in a special way to enable efficient incremental access. Linearized PDF is sometimes referred to as PDF for "fast web view." Its primary goal is to enhance the viewing performance

whilst downloading a PDF file over a streaming communications channel such as the internet. When data for a page is delivered over the channel, you'd like to have the page content displayed incrementally as it arrives.

With the essential cross-reference at the end of the file, this isn't possible unless the file is linearized. All the content in the PDF file needs to be reorganized so that the first page can be displayed as quickly as possible without the need to read all of the rest of the file, or to start reading with the final cross-reference file at the very end of the file.

Such a reorganization of the PDF objects, creating a cross-reference for each page, can only be done after the PDF file is completed and after all resources are known. iText can read linearized PDFs, but it can't create a linearized PDF, nor can you (currently) linearize an existing PDF using iText.

2.2.3 PDFs with compressed object and cross-reference streams

Starting with PDF 1.5, the cross reference table can be stored as an indirect object in the body, more specifically as a stream object allowing big cross-reference tables to be compressed. Additionally, the file size can be reduced by putting different objects into one compressed object stream.

Code sample 2.3 has the same appearance as code sample 2.1 when opened in a PDF viewer, but the internal file structure is quite different:

Code sample 2.3: Compressed PDF file structure

```
%PDF-1.5
%âãÏÓ
2 0 obj
<</Length 64/Filter/FlateDecode>>stream
*binary stuff*
endstream
endobj
6 0 obj
<</Type/Catalog/Pages 3 0 R>>
endobj
7 0 obj
<</Producer(iText® 5.4.2 ©2000-2012 1T3XT BVBA \(AGPL-version\))/ModDate(D:20130502165150+\
02'00')/CreationDate(D:20130502165150+02'00')>>
endobj
5 0 obj
<</Type/ObjStm/N 3/Length 191/First 16/Filter/FlateDecode>>stream
*binary stuff*
endstream
endobj
8 0 obj
<</Type/XRef/W[1 2 2]/Root 6 0 R/Index[0 9]/ID [<21cb45acb652a807ba62d55f7b29d8be><21cb45a\
cb652a807ba62d55f7b29d8be>]/Length 41/Info 7 0 R/Size 9/Filter/FlateDecode>>stream
*binary stuff*
endstream
endobj
```

```
26 %iText-5.4.2
27 startxref
28 626
29 %%EOF
```

Note that the header now says `%PDF-1.5`. When I created this file, I've opted for full compression before opening the `Document` instance, and iText has automatically changed the version to 1.5.

The `startxref` keyword in line 28 no longer refers to the byte position of an `xref` keyword, but to the byte position of the stream object containing the cross-reference stream.

The stream dictionary of a cross-reference stream has a `/Length` and a `/Filter` entry just like all other streams, but also requires some extra entries as listed in table 2.2.

Table 2.2: Entries specific to a cross-reference stream dictionary

Key	Type	Value
Type	name	Required; always `/XRef`.
W	array	Required; an array of integers representing the size of the fields in a single cross reference entry.
Root	dictionary	Required; refers to the catalog dictionary; equivalent to the `/Root` entry in the trailer dictionary.
Index	array	An array containing a pair of integers for each subsection in the cross-reference table. The first integer shall be the first object number in the subsection; the second integer shall be the number of entries in the subsection.
ID	array	An array containing a pair of IDs equivalent to the `/ID` entry in the trailer dictionary.
Info	dictionary	An info dictionary, equivalent to the `/Info` entry in the trailer dictionary (deprecated in PDF 2.0).
Size	integer	Required; equivalent to the `/Size` entry in the trailer dictionary.
Prev	integer	Equivalent of the `/Prev` key in the trailer dictionary. Refers to the byte offset of the beginning of the previous cross-reference stream (if such a stream is present).

If we look at code sample 2.3, we see that the `/Size` of the cross-reference table is 9, and all entries are organized in one subsection `[0 9]`, which means the 9 entries are numbered from 0 to 8. The value of the `W` key, in our case `[1 2 2]`, tells us how to distinguish the different cross-reference entries in the stream, as well as the different parts of one entry.

Let's examine the stream by converting each byte to a hexadecimal number and by adding some extra white space so that we recognize the `[1 2 2]` pattern as defined in the `W` key:

```
00 0000 ffff
02 0005 0001
01 000f 0000
02 0005 0002
02 0005 0000
01 0157 0000
01 0091 0000
01 00be 0000
00 0000 ffff
```

We see 9 entries, representing objects 0 to 8. The first byte can be one out of three possible values:

- If the first byte is `00`, the entry refers to a free entry. We see that object 0 is free (as was to be expected), as well as object 8, which is the object that stores the cross-reference stream itself.
- If the first byte is `01`, the entry refers to an object that is present in the body as an uncompressed indirect object. This is the case for objects 2, 5, 6, and 7. The second part of the entry defines the byte offset of these objects: 15 (`000f`), 343 (`0157`), 145 (`0091`) and 190 (`00be`). The third part is the generation number.
- If the first byte is `02`, the entry refers to a compressed object. This is the case with objects 1, 3, and 4. The second part gives you the number of the object stream in which the object is stored (in this case object 5). The third part is the index of the object within the object stream.

Objects 1, 3, and 4 are stored in object 5. This object is an object stream, and its stream dictionary requires some extra keys as listed in table 2.3.

Table 2.3: Entries specific to an object stream dictionary

Key	Type	Value
Type	name	Required; always `/ObjStm`.
N	integer	Required; the number of indirect objects stored in the stream.
First	integer	Required; the byte offset in the decoded stream of the first compressed object
Extends	stream	A reference to another object stream, of which the current object shall be considered an extension.

The `N` value of the stream dictionary in code sample 2.3 tells us that there are three indirect objects stored in the object stream. The entries in the cross-reference stream tell us that these objects are numbered and ordered as 4, 1, and 3. The `First` value tells us that object 4 starts at byte position 16.

We'll find three pairs of integers, followed by three objects starting at byte position 16 when we uncompress the object stream stored in object 5. I've added some extra newlines to the uncompressed stream so that we can distinguish the different parts:

```
4 0
1 142
3 215
<</Parent 3 0 R/Contents 2 0 R/Type/Page/Resources<</ProcSet [/PDF /Text /ImageB /ImageC /\
ImageI]/Font<</F1 1 0 R>>>>/MediaBox[0 0 595 842]>>
<</BaseFont/Helvetica/Type/Font/Encoding/WinAnsiEncoding/Subtype/Type1>>
<</Type/Pages/Count 1/Kids[4 0 R]>>
```

The three pairs of integers consist of the numbers of the objects (4, 1, and 3), followed by their offset relative to the first object stored in the stream. We recognize a dictionary of type `/Page` (object 4), a dictionary of type `/Font` (object 1), and a dictionary of type `/Pages` (object 3).

You can never store the following objects in an object stream:

- stream objects,
- objects with a generation number different from zero,
- a document's encryption dictionary,
- an object representing the value of the `/Length` entry in an object stream dictionary,
- the document catalog dictionary,
- the linearization dictionary, and
- page objects of a linearized file.

Now that we know how a cross-reference is organized and how indirect objects are stored either in the body or inside a stream, we can retrieve all the relevant PDF objects stored in a PDF file.

2.3 Summary

In this chapter, we've examined the four parts of a PDF file: the header, the body, the cross-reference table and the trailer. We've learned that some PDFs have incremental updates, that the cross-reference table can be compressed into an object, and that objects can be stored inside an object stream. We can now start exploring the file structure of every PDF file that can be found in the wild.

While looking under the hood of some simple PDF documents, we've encountered objects such as the Catalog dictionary, Pages dictionaries, Page dictionaries, and so on. It's high time we discover how these objects relate to each other and how they form a document.

3 PDF Document Structure

In chapter 1, we've learned about the different types of objects available in the Portable Document Format, and we discovered that one object can refer to another using an indirect reference. In chapter 2, we've learned how the objects are stored in a file, as well as where to find indirect objects based on their object number. In this chapter, we're going to combine this knowledge to find out how these objects are structured into a hierarchy that defines a document.

3.1 Viewing a document as a tree structure using RUPS

The seemingly linear sequence of PDF objects we see when we open a PDF file in a text editor, isn't as linear as one might think at first sight.

Figure 3.1 shows the "Hello World" document we examined in code sample 2.1, opened in iText RUPS.

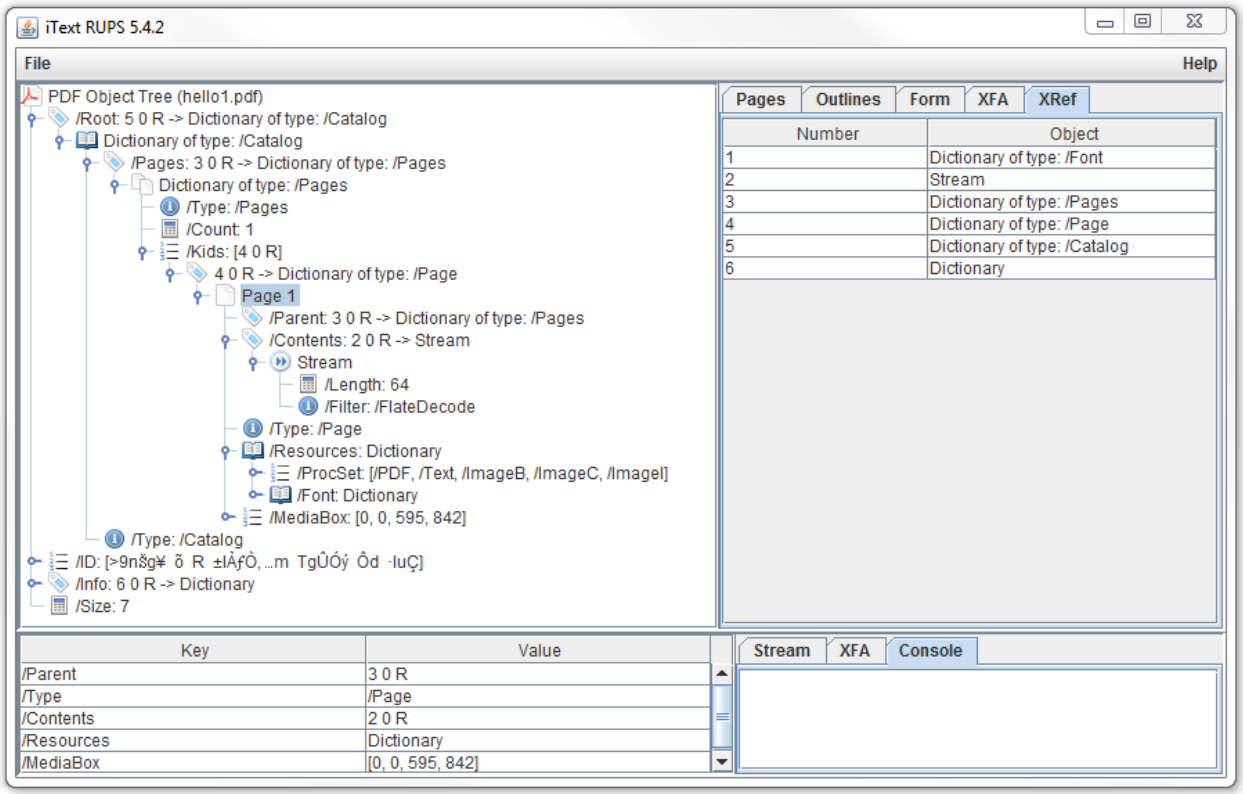


Figure 3.1: Hello World opened in iText RUPS

RUPS offers a Graphical User Interface that allows you to look inside a PDF. It's written in Java and compiled to a Windows executable. You can download the source code and the binary from SourceForge[1].

[1] http://sourceforge.net/projects/itextrups/

To the left, you recognize the entries of the trailer dictionary (see section 2.1.4). These entries are visualized in a Tree-view panel as the branches of a tree. The most prominent branch is the `/Root` dictionary. In figure 3.1, we've opened the `/Pages` dictionary, and we've unfolded the leaves of the `/Page` dictionary representing "Page 1" of the document.

To the right, there's a panel with different tabs. We see the **XRef** tab, listing the entries of the cross-reference table. It contains all the objects we discussed in section 2.1.3, organized in a table with rows numbered from 1 to 6. Clicking a row opens the corresponding object in the Tree-view panel. We'll take a look at the other tabs later on.

At the bottom, we can find info about the object that was selected. In this case, RUPS shows a tabular structure listing the keys and values of the `/Page` dictionary that was opened in the tree view panel.

To the right, we see another panel with different tabs. The **Console** tab shows whatever output is written to the `System.out` or `System.err` while using RUPS. Here's where you'll find the stack trace when you try reading a file that can't be parsed by iText because it contains invalid PDF syntax. We'll have a closer look at the **Stream** panel in part 2 and at the **XFA** panel in part 3 of this book.

Figure 3.2 shows the "Hello World" document we examined in code sample 2.3.

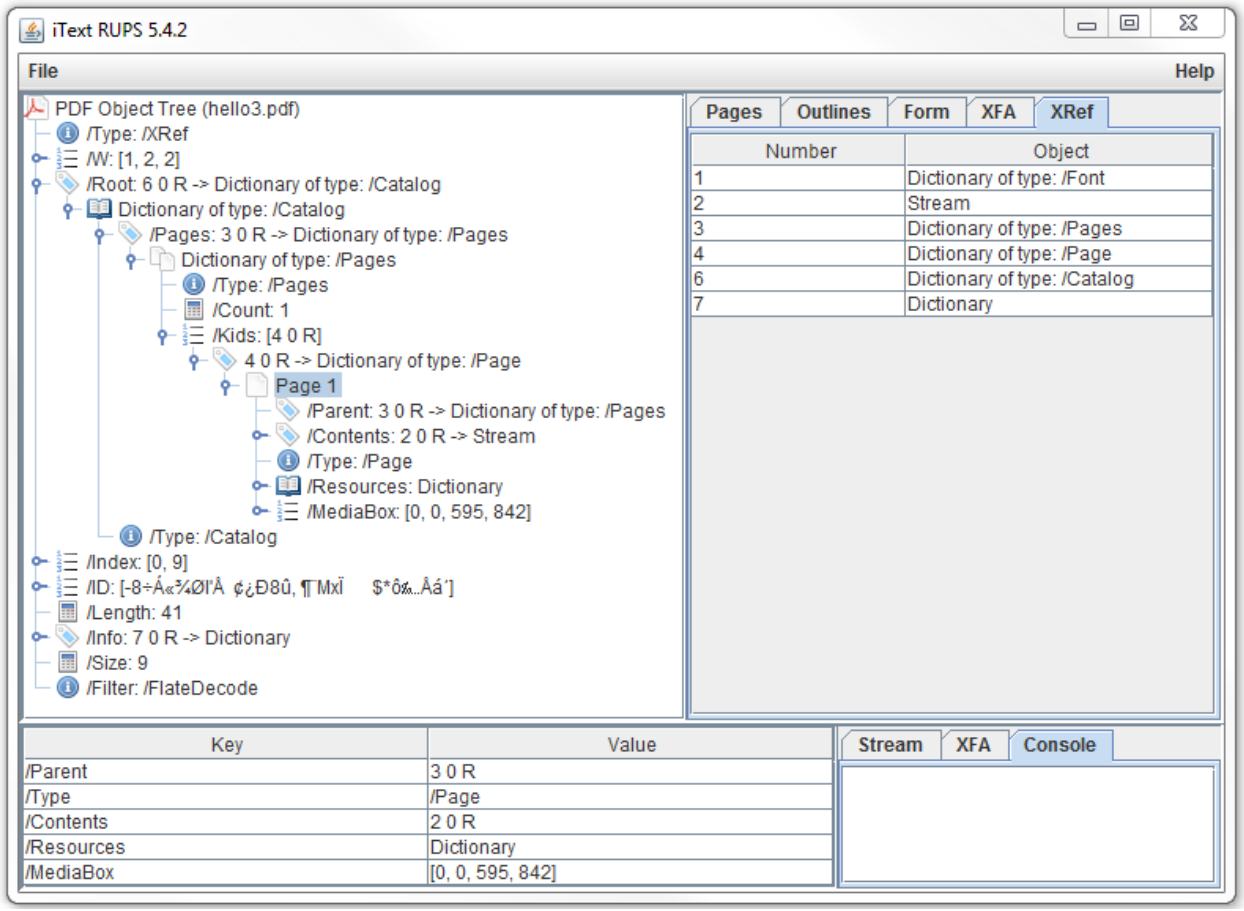


Figure 3.2: Hello World opened in iText RUPS

When you open a file with a compressed cross-reference stream, RUPS shows the `/XRef` dictionary instead of the trailer dictionary (because there is no trailer dictionary).

The **XRef** table on the right is also slightly different. Based on what we know from section 2.2.3 about this "Hello World" file, we notice that two objects are missing:

- *object 5* — a compressed object stream. Instead of showing the original stream, RUPS shows the objects that were compressed into this stream: 1, 4 and 5.
- *object 8* — the compressed cross-reference stream. This stream isn't shown either; instead its content is interpreted and visualized in the **XRef** tab.

When you open a document that was incrementally updated in RUPS, you'll only see the most recent objects. RUPS doesn't show any unused objects.

The history behind RUPS

I wrote RUPS out of frustration, at a time iText wasn't generating any revenue. When I needed to debug a PDF file, I used to open that PDF in a text editor. I then had to search through that text file looking for specific object numbers and references. When I needed to examine streams, I used the iText toolbox, a predecessor of RUPS, to decompress the binary data.

All of this was very time-consuming and almost unaffordable as long as I didn't get paid for debugging other people's documents. So I've spent the Christmas holidays of 2007 writing a GUI to "Read and Update PDF Syntax" aka "RUPS". Rups is the Dutch word for caterpillar, and I imagined the GUI as a tool to penetrate into the heart of a document, the way a caterpillar eats its way through the leaves of a plant.

My initial idea was to also allow people to change objects at their core and by doing so, to update their PDFs manually. We've only recently started implementing functionality that allows updating keys in dictionaries and applying other minor changes. Such functionality makes it very easy for people who aren't fluent in PDF to cause serious damage to a PDF file. We still aren't sure if it's a good idea to allow this kind of PDF updating.

Now that we have a means to look at the document structure using a tool with a GUI, let's find out how we can obtain the different objects that compose a PDF document programmatically, using code.

3.2 Obtaining objects from a PDF using PdfReader

When you open a document with RUPS, RUPS uses iText's `PdfReader` class under the hood. This class allows you to inspect a PDF file at the lowest level. Code sample 3.1 shows how we can create such a `PdfReader` instance and fetch different objects.

Code sample 3.1: C0301_TrailerInfo

```
1  public static void main(String[] args) throws IOException {
2      PdfReader reader =
3          new PdfReader("src/main/resources/primes.pdf");
4      PdfDictionary trailer = reader.getTrailer();
5      showEntries(trailer);
6      PdfNumber size = (PdfNumber)trailer.get(PdfName.SIZE);
7      showObject(size);
8      size = trailer.getAsNumber(PdfName.SIZE);
9      showObject(size);
10     PdfArray ids = trailer.getAsArray(PdfName.ID);
11     PdfString id1 = ids.getAsString(0);
12     showObject(id1);
13     PdfString id2 = ids.getAsString(1);
14     showObject(id2);
15     PdfObject object = trailer.get(PdfName.INFO);
16     showObject(object);
17     showObject(trailer.getAsDict(PdfName.INFO));
18     PdfIndirectReference ref = trailer.getAsIndirectObject(PdfName.INFO);
19     showObject(ref);
20     object = reader.getPdfObject(ref.getNumber());
21     showObject(object);
22     object = PdfReader.getPdfObject(trailer.get(PdfName.INFO));
23     showObject(object);
24     reader.close();
25 }
26 public static void showEntries(PdfDictionary dict) {
27     for (PdfName key : dict.getKeys()) {
28         System.out.print(key + ": ");
29         System.out.println(dict.get(key));
30     }
31 }
32 public static void showObject(PdfObject obj) {
33     System.out.println(obj.getClass().getName() + ":");
34     System.out.println("-> type: " + obj.type());
35     System.out.println("-> toString: " + obj.toString());
36 }
```

In this code sample, we create a `PdfReader` object that is able to read and interpret the PDF syntax stored in the file `primes.pdf`. This `reader` object will allow us to obtain any indirect object as an iText PDF object from the body of the PDF document. But let's start by fetching the trailer dictionary.

In line 4, we get the trailer dictionary using the `getTrailer()` method. We take a look at its entries the same way we looked a the entries of other dictionaries in section 1.2.6.

The `showEntries()` method produces the following output:

```
/Root: 762 0 R
/ID: [8Ã¯?□2õg©¬?Ô , 8Ã¯?□2õg©¬?Ô ]
/Size: 764
/Info: 763 0 R
```

In line 6 of code sample 3.1, we use the same `get()` as in the `showEntries()` method to obtain the value of the `/Size` entry. As we expect a number, we cast the `PdfObject` to a `PdfNumber` instance. We'll get a `ClassCastException` if the value of the entry is of a different type. The same exception will be thrown if the entry is missing in the dictionary, in which case the `get()` method will return `null`.

One way to avoid `ClassCastException` problems, is to get the value as a `PdfObject` instance first and to check whether or not it's `null`. If it's not, we can check the type before casting the `PdfObject` to one of its subclasses. An alternative to this convoluted method sequence would be to use one of the `getAsX()` methods listed in table 3.1.

Table 3.1: Overview of the getters available in PdfArray and PdfDictionary

Method name	Return type
`get()` / `getPdfObject()`	a `PdfObject` instance (could even be an indirect reference). The `get()` method is to be used for entries in a `PdfDictionary`; the `getPdfObject()` for elements in a `PdfArray`.
`getDirectObject()`	a `PdfObject` instance. Indirect references will be resolved. In case the value of an entry is referenced, `PdfReader` will go and fetch the `PdfObject` using that reference. You'll get a direct object, or `null` if the object can't be found.
`getAsBoolean()`	a `PdfBoolean` instance.
`getAsNumber()`	a `PdfNumber` instance.
`getAsString()`	a `PdfString` instance.
`getAsName()`	a `PdfName` instance.
`getAsArray()`	a `PdfArray` instance.
`getAsDict()`	a `PdfDictionary` instance.
`getAsStream()`	a `PdfStream` instance, that can be cast to a `PRStream` object.
`getAsIndirectObject()`	a `PdfIndirectReference` instance, that can be cast to a `PRIndirectReference` object.

These methods either return a specific subclass of `PdfObject`, or they return `null` if the object was of a different type or missing. In line 8 of code sample 3.1, we get a `PdfNumber` by using `trailer.getAsNumber(PdfName.SIZE);`

Suppose that we had used the `getAsString()` method instead of the `getAsNumber()` method. This would have returned `null` because the size isn't expressed as a `PdfString` value. This behavior is useful in case you don't know the type of the value for a specific entry in advance. For instance, when we'll talk about *named destinations* in section 35.2.1.1, we'll see that a named destination can be defined using either a `PdfString` or a `PdfName`. We could use the `getAsName()` method as well as the `getAsString()` method and check which method doesn't return `null` to determine which flavor of named destination we're dealing with.

When invoked on a PdfDictionary, the methods listed in table 3.1 require a PdfName —the key— as parameter; when invoked on a PdfArray, they require an int —the index. In line 10 of code sample 3.1, we get the /ID entry as a PdfArray, and we get the two elements of the array using the getAsString() method and the indexes 0 and 1.

In line 15, we ask for the /Info entry, but the info dictionary isn't stored in the trailer dictionary as a direct object. The entry in the trailer dictionary refers to an indirect object with number 763. If we want the actual dictionary, we need to use the getAsDict() method. This method will look at the object number of the indirect reference and fetch the corresponding indirect object from the PdfReader instance.

Take a look at the output of the showObject() methods in line 16 and line 17 to see the difference:

```
com.itextpdf.text.pdf.PRIndirectReference:
-> type: 10
-> toString: 763 0 R
com.itextpdf.text.pdf.PdfDictionary:
-> type: 6
-> toString: Dictionary
```

The get() method returns the reference, the getAsDict() method returns the actual object by fetching the content of object 763. Note that the reference instance is of type PRIndirectReference.

The PdfStream and PdfIndirectReference objects have PRStream and PRIndirectReference subclasses. The prefix PR refers to PdfReader and the object instances contain more information than the object instances we've discussed in chapter 1. For instance: if you want to extract the content of a stream, you'll need the PRStream instance instead of the PdfStream object.

On line 18, we try a slightly different approach. First, we get the indirect reference value of the /Info dictionary using the getAsIndirectReferenceObject() method. Then we get the actual object from the PdfReader by using the reference number. PdfReader's getPdfObject() method can give you every object stored in the body of a PDF file by its number. PdfReader will fetch the byte position of the indirect object from the cross-reference table and parse the object found at that specific byte offset.

As an alternative, you can also use PdfReader's static getPdfObject() method that accepts a PdfObject instance as parameter. If this parameter is an indirect reference, the reference will be resolved. If it's a direct object, that object will be returned as-is.

Now that we've played with different objects obtained from a PdfReader instance, let's explore the document structure using code. Looking at what RUPS shows us, the /Root dictionary aka the Document Catalog dictionary is where we should start. This dictionary has two required entries. One is the /Type which must be /Catalog. The other is the /Pages entry which refers to the root of the page tree.

We'll look at the optional entries in a moment, but let's begin by looking at the page tree.

3.3 Examining the page tree

Every page in a PDF document is defined using a /Page dictionary. These dictionaries are stored in a structure known as the page tree. Each /Page dictionary is the child of a page tree node, which is a dictionary of type

/Pages. One could work with a single page tree node, the one that is referred to from the catalog, but that would be bad practice. The performance of PDF consumers can be optimised by contructing a balanced tree.

If you create a PDF using iText, you won't have more than 10 /Page leaves or /Pages branches attached to every /Pages node. By design, a new intermediary page tree node is introduced by iText every 10 pages.

Before we start coding, let's take a look at figure 3.3. It shows part of the page tree of the primes.pdf document using RUPS, starting with the root node.

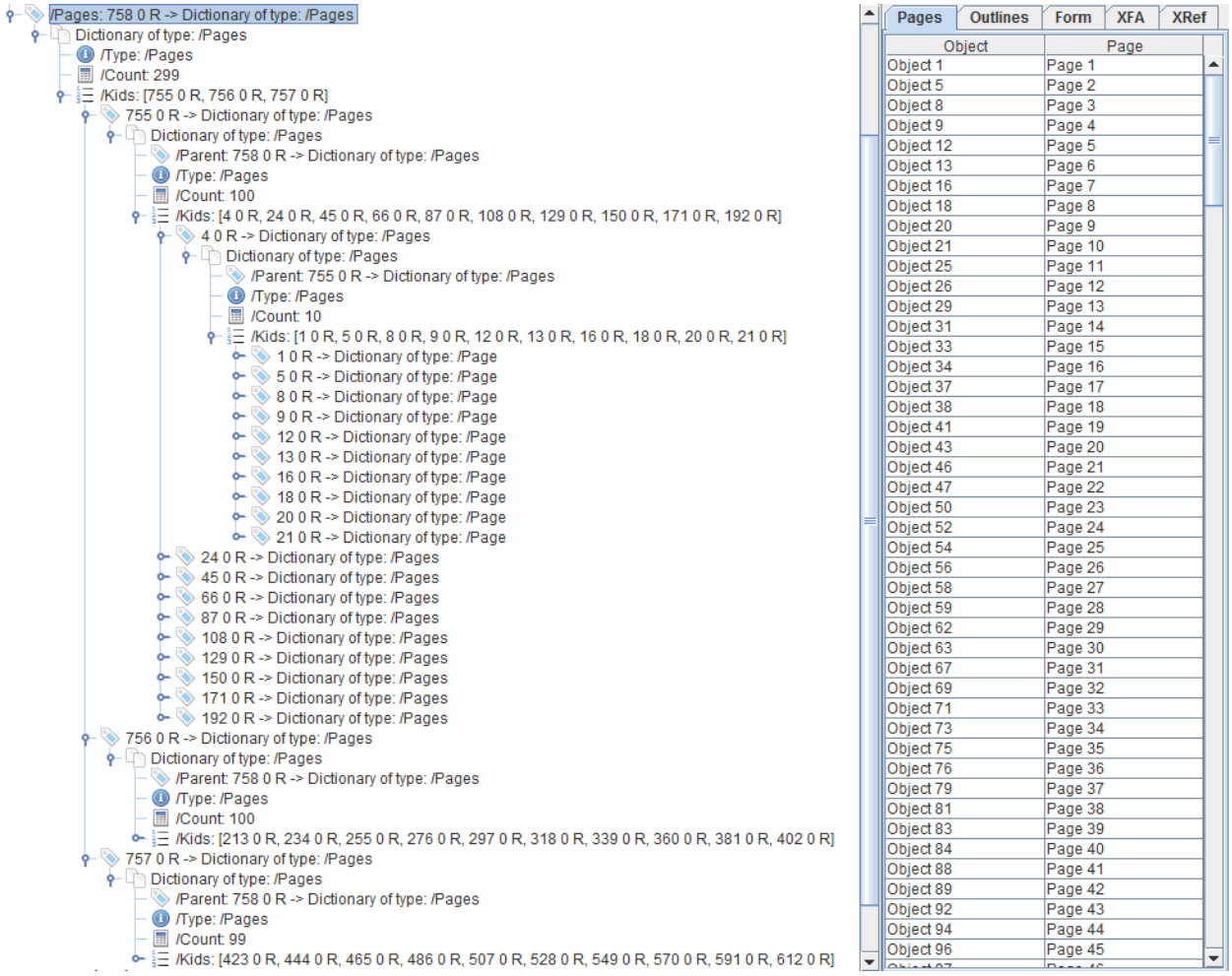


Figure 3.3: The page tree of primes.pdf opened in RUPS

The/Count entry of a page tree node shows the total number of leaf nodes attached directly or indirectly to this branch. The root of the page tree (object 758) shows that the document has 299 pages. The /Kids entry is an array with references to three other page tree nodes (objects 755, 756 and 757). The 299 leaves are nicely distributed over these three branches: 100, 100 and 99 pages. Each branch or leaf requires a /Parent entry referring to its parent; for the root node, the /Parent entry is forbidden.

When we expand the first page tree node, we discover that this tree node has ten branches. The first of these ten page tree nodes (object 4) has ten leaves, each leaf being a dictionary of type `/Page`. If you look to the panel at the right, you see that we've selected the **Pages** tab. This tab shows a table in which every row represents a page in the document. In the first column, you'll find the object number of a `/Page` dictionary; in the second column, you'll find its page number.

3.3.1 Page Labels

A page object on itself doesn't know anything about its page number. The page number of a page is calculated based on the occurence of the page dictionary in the page tree. In figure 3.3, RUPS has examined the page tree, and attributed numbers going from 1 to 299.

If the Catalog has a `/PageLabels` entry, viewers can present a different numbering, for instance using latin numbering, such as *i, ii, iii, iv*, etc... It's important to understand that page labels and even page numbers are completely independent from the number that may or may not be visible on the actual page. Both the page number and its label only serve as an extra info when browsing the document in a viewer. You won't see any of these page labels on the printed document.

Figure 3.4 shows an example of a PDF file with a `/PageLabels` entry.

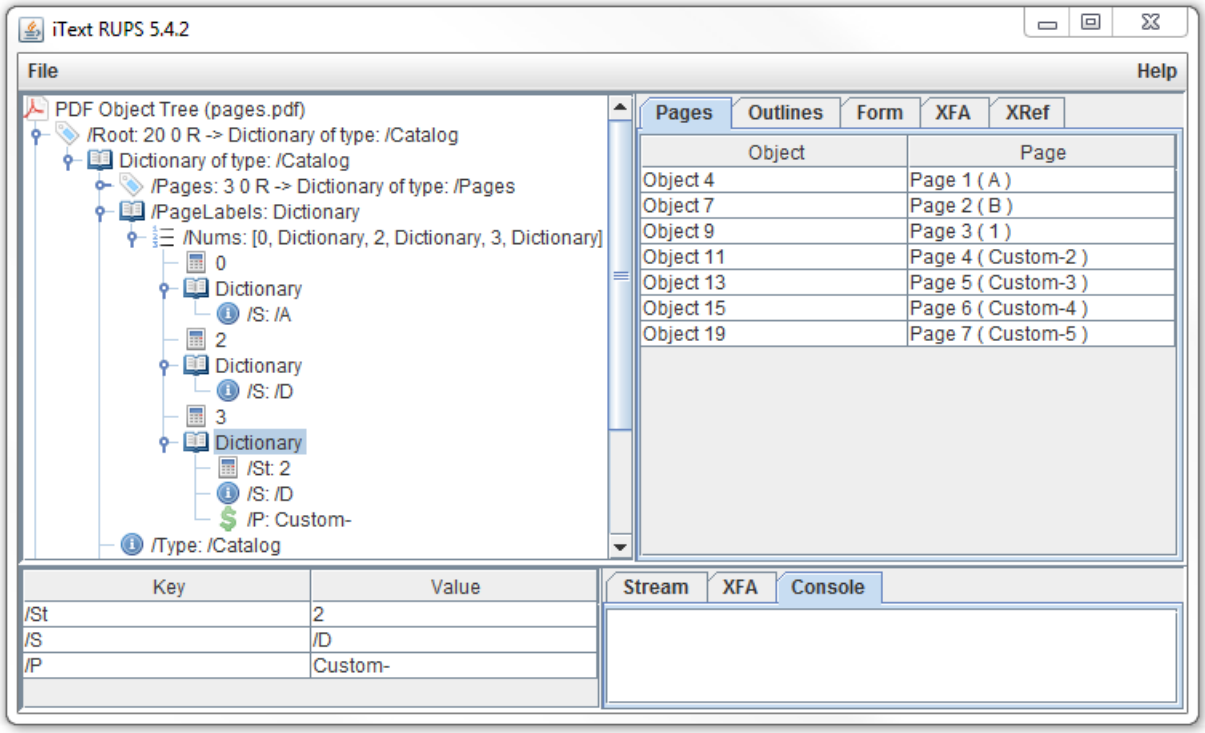


Figure 3.4: Using page labels

The value of the `/PageLabels` entry is a number tree.

What is a number tree?

A number tree serves a similar purpose as a dictionary, associating keys and values, but the keys are numbers, they are ordered, and a structure similar to the page tree (involving branches and leaves) can be used. The leaves are stored in an array that looks like this `[key1 value1 key2 value2 ... keyN valueN]` where the keys are numbers sorted in numerical order and the values are either references to a string, array, dictionary or stream, or direct objects in case of null, boolean, number or name values. See also section 3.5.1 for the definition of a name tree.

In the case of a number tree defining page labels, you always need a `0` key for the first page. The value of each entry will be a page label dictionary. Table 3.2 lists the possible entries of such a dictionary.

Table 3.2: Entries in a page label dictionary

Key	Type	Value
Type	name	Optional value: `/PageLabel`
S	name	The numbering style: - `/D` for decimal, - `/R` for uppercase roman numerals, - `/r` for lowercase roman numerals, - `/A` for uppercase letters, - `/a` for lowercase letters. In case of letters, the pages go from A to Z, then continue from AA to ZZ. If the `/S` entry is missing, page numbers will be omitted.
P	string	A prefix for page labels.
St	number	The first page number —or its equivalent— for the current page label range.

Looking at figure 3.4, we see three page label ranges:

1. *Index `0` (page 1)*— the page labels consist of uppercase letters,
2. *Index 2 (page 3)*— the page labels consist of decimals. As we've started a new range, the numbering restarts at 1. This means that page 3 will get "1" as page label.
3. *Index 3 (page 4)*— the page labels consist of decimals, but starts with label "2" (as defined in the `/St` entry). It also introduces a prefix (`/P`): "Custom-".

When opened in a PDF viewer, the pages of this document will be numbered *A*, *B*, *1*, *Custom-2*, *Custom-3*, *Custom-4*, and *Custom-5*. Talking about page labels was fun, but now let's find out how to obtain a page dictionary based on its sequence in the page tree.

3.3.2 Walking through the page tree

Code sample 3.2 shows how we could walk through the page tree to find all the pages in a document. This time we get the Catalog straight from the `reader` instance using the `getCatalog()` method instead of using `trailer.getAsDict(PdfName.ROOT)`. Once we have the Catalog, we get the `/Pages` entry, and pass it to the `expand()` method.

Code sample 3.2: C0302_PageTree

```
public static void main(String[] args) throws IOException {
    PdfReader reader
        = new PdfReader("src/main/resources/primes.pdf");
    PdfDictionary dict = reader.getCatalog();
    PdfDictionary pageroot = dict.getAsDict(PdfName.PAGES);
    new C0302_PageTree().expand(pageroot);
}

private int page = 1;
public void expand(PdfDictionary dict) {
    if (dict == null)
        return;
    PdfIndirectReference ref = dict.getAsIndirectObject(PdfName.PARENT);
    if (dict.isPage()) {
        System.out.println("Child of " + ref + ": PAGE " + (page++));
    }
    else if (dict.isPages()) {
        if (ref ==  null)
            System.out.println("PAGES ROOT");
        else
            System.out.println("Child of " + ref + ": PAGES");
        PdfArray kids = dict.getAsArray(PdfName.KIDS);
        System.out.println(kids);
        if (kids != null) {
            for (int i = 0; i < kids.size(); i++) {
                expand(kids.getAsDict(i));
            }
        }
    }
}
```

The `C0302_PageTree` example has a single private member variable `page` that is initialized at `1`. This variable is used in the recursive `expand()` method:

- If the dictionary passed to the method is of type `/Page`, the `isPage()` method will return `true`, and we'll increment the page number, writing it to the `System.out` along with info about the parent.
- If the dictionary passed to the method is of type `/Pages`, the `isPages()` method will return `true`, and we'll loop over all the `/Kids` array, calling the `expand()` method recursively for every branch or leaf.

The output of code sample 3.2 is consistent with what we saw in figure 3.3:

```
PAGES ROOT
[755 0 R, 756 0 R, 757 0 R]
Child of 758 0 R: PAGES
[4 0 R,24 0 R,45 0 R,66 0 R,87 0 R,108 0 R,129 0 R,150 0 R,171 0 R,192 0 R]
Child of 755 0 R: PAGES
[1 0 R,5 0 R,8 0 R,9 0 R,12 0 R,13 0 R,16 0 R,18 0 R,20 0 R,21 0 R]
Child of 4 0 R: PAGE 1
Child of 4 0 R: PAGE 2
Child of 4 0 R: PAGE 3
Child of 4 0 R: PAGE 4
Child of 4 0 R: PAGE 5
Child of 4 0 R: PAGE 6
Child of 4 0 R: PAGE 7
Child of 4 0 R: PAGE 8
Child of 4 0 R: PAGE 9
Child of 4 0 R: PAGE 10
Child of 755 0 R: PAGES
[25 0 R,26 0 R,29 0 R,31 0 R,33 0 R,34 0 R,37 0 R,38 0 R,41 0 R,43 0 R]
Child of 24 0 R: PAGE 11
Child of 24 0 R: PAGE 12
...
```

This is one way to obtain the `/Page` dictionary of a certain page. Fortunately, there's a more straight-forward method. In code sample 3.3, the `getNumberOfPages()` method provides us with the total number of pages. We loop from 1 to that number and use the `getPageN()` method to get the `/Page` dictionary for each separate page.

Code sample 3.3: C0303_PageTree

```
1  int n = reader.getNumberOfPages();
2  PdfDictionary page;
3  for (int i = 1; i <= n; i++) {
4      page = reader.getPageN(i);
5      System.out.println("The parent of page " + i + " is " + page.get(PdfName.PARENT));
6  }
```

The output of this code snippet corresponds with what we had before:

```
The parent of page 1 is 4 0 R
The parent of page 2 is 4 0 R
The parent of page 3 is 4 0 R
The parent of page 4 is 4 0 R
The parent of page 5 is 4 0 R
The parent of page 6 is 4 0 R
The parent of page 7 is 4 0 R
The parent of page 8 is 4 0 R
The parent of page 9 is 4 0 R
The parent of page 10 is 4 0 R
The parent of page 11 is 24 0 R
The parent of page 12 is 24 0 R
...
```

ISO-32000-1 and -2 define many possible entries for the `/Page` dictionary. It would lead us too far to discuss them all, but let's take a look at the most important ones.

3.4 Examining a page dictionary

Every `/Page` dictionary specifies the attributes of a single page. Table 3.3 lists the *required* entries in the `/Page` dictionary.

Table 3.3: Required entries in a page dictionary

Key	Type	Value
Type	name	Must be `/Page`
Parent	name	The page tree node that is the immediate parent of the page.
Resources	string	A dictionary containing any resources needed for the page. If the page doesn't require resources,an empty dictionary must be present. We'll discuss the possible entries in section 3.4.1.2.
MediaBox	rectangle	A rectangle defining the page size: the physical boundaries on which the page shall be displayed or printed. Other (optional) boundaries will be discussed in section 3.4.2.2.

The actual content of the page is stored in the `/Content` entry. This entry isn't listed in table 3.3 because it isn't required. If it's missing, the page is blank.

The value of the `/Contents` entry can either be a reference to a stream or an array. If it's an array, the elements consist of references to streams that need to be concatenated when rendering the page content.

3.4.1 The content stream and its resources

We'll discuss the syntax needed to describe the content in Part 2, but let's already peek at the content of the `/Contents` entry.

3.4.1.1 The content stream of a page

In code sample 3.4, we get the value of the `/Contents` entry as a `PRStream`; a `PdfStream` wouldn't be sufficient to get the stream bytes. `PdfReader` has two types of static methods to extract the contents of a stream: `getStreamBytesRaw()` gets the original bytes; `getStreamBytes()` returns the uncompressed bytes.

Code sample 3.4: C0304_PageContent

```
1 PdfDictionary page = reader.getPageN(1);
2 PRStream contents = (PRStream)page.getAsStream(PdfName.CONTENTS);
3 byte[] bytes = PdfReader.getStreamBytes(contents);
4 System.out.println(new String(bytes));
```

The first page of the document we're parsing has two paragraphs: *Hello World* en *Hello People*. You can easily recognize these sentences in the output that is produced by code sample 3.4:

```
q
BT
36 806 Td
0 -18 Td
/F1 12 Tf
(Hello World )Tj
0 0 Td
0 -18 Td
(Hello People )Tj
0 0 Td
ET
Q
```

Don't worry about the syntax. Every operator and operand will be explained in chapter 4, entitled "Graphics state" —for example: `q` and the `Q` are graphics state operators— and chapter 5, entitled "Text State" —`BT` and `ET` are text state operators.

The second page looks identical to the first page when opening the document in a PDF viewer. Internally there's a huge difference.

Code sample 3.5 shows a short-cut method to get the content stream of a page.

Code sample 3.5: C0304_PageContent

```
1 bytes = reader.getPageContent(2);
2 System.out.println(new String(bytes));
```

The resulting stream looks like this:

```
q
BT
36 806 Td
ET
Q
BT
/F1 12 Tf
88.66 788 Td
(ld)Tj
-22 0 Td
(Wor)Tj
-15.33 0 Td
(llo)Tj
-15.33 0 Td
(He)Tj
ET
q 1 0 0 1 36 763 cm /Xf1 Do Q
```

We still recognize the text *Hello World*, but it's mangled into *ld*, *Wor*, *llo* and *He*. Many PDFs aren't created in an ideal way. You're bound to encounter documents of which the content stream doesn't contain the exact words you expect. You shouldn't expect the content stream of a PDF to be human-readable.

Looking at the output of code sample 3.5, you notice that the worlds *Hello People* seem to be missing from this content stream. This text snippet is added as an *external object* or *XObject* marked as `/Xf1`. We'll find this object in the `/XObject` entry of the page resources.

3.4.1.2 The Resources dictionary

In code snippet 3.6, we get the resources from the page dictionary, and we list all of its entries.

Code sample 3.6: C0304_PageContent

```
page = reader.getPageN(2);
PdfDictionary resources = page.getAsDict(PdfName.RESOURCES);
for (PdfName key : resources.getKeys()) {
    System.out.print(key);
    System.out.print(": ");
    System.out.println(resources.getDirectObject(key));
}
```

The output of this code snippet looks like this:

```
/XObject: Dictionary
/Font: Dictionary
```

If we'd examine these dictionaries, we'd find the key `/Xf1` refering to a Form XObject in the former, and the key `/F1` refering to a font in the latter. We recognize references to these keys in the content stream resulting from code snippet 3.5.

Code sample 3.7 lists the entries and the content stream of the XObject with name `/Xf1` (line 2) and the entries of the font dictionary with name `/F1` (line 9):

Code sample 3.7: C0304_PageContent

```
1  PdfDictionary xObjects = resources.getAsDict(PdfName.XOBJECT);
2  PRStream xObject = (PRStream)xObjects.getAsStream(new PdfName("Xf1"));
3  for (PdfName key : xObject.getKeys()) {
4      System.out.println(key + ": " + xObject.getDirectObject(key));
5  }
6  bytes = PdfReader.getStreamBytes(xObject);
7  System.out.println(new String(bytes));
8  PdfDictionary fonts = resources.getAsDict(PdfName.FONT);
9  PdfDictionary font = fonts.getAsDict(new PdfName("F1"));
10 for (PdfName key : font.getKeys()) {
11     System.out.println(key + ": " + font.getDirectObject(key));
12 }
```

The resulting output for the XObject looks like this:

```
/Matrix: [1, 0, 0, 1, 0, 0]
/Filter: /FlateDecode
/Type: /XObject
/FormType: 1
/Length: 48
/Resources: Dictionary
/Subtype: /Form
/BBox: [0, 0, 250, 25]
BT
/F1 12 Tf
0 7 Td
(Hello People )Tj
ET
```

Now we clearly see *Hello People* in the content stream of the external object. This form XObject also contains a resources dictionary. We'll discuss the entries of the XObject's stream dictionary in more detail in part 2.

The resulting output for the font looks like this:

```
/Type: /Font
/BaseFont: /Helvetica
/Subtype: /Type1
/Encoding: /WinAnsiEncoding
```

A trained eye immediately notices that this dictionary represents the standard Type1 font Helvetica, a font that doesn't need to be embedded as it's one of the 14 standard Type 1 fonts. In part 2, we'll find out more about the different fonts in a PDF.

Table 3.4 offers the complete list of entries one can encounter in a `/Resources` dictionary.

Table 3.4: Possible entries in a resources dictionary

Key	Type	Value
Font	dictionary	A dictionary that maps resource names to font dictionaries.
XObject	dictionary	A dictionary that maps resource names to external objects.
ExtGState	dictionary	A dictionary that maps resource names to graphics state parameter dictionaries.
ColorSpace	dictionary	A dictionary that maps each resource name to either the name of a device-dependent colorspace or an array describing a color space.
Pattern	dictionary	A dictionary that maps resource names to pattern objects.
Shading	dictionary	A dictionary that maps resource names to shading dictionaries.
Properties	dictionary	A dictionary that maps resource names to property list dictionaries for marked content.
ProcSet	array	A feature that became obsolete in PDF 1.4 and that can safely be ignored.

Table 3.4 contains plenty of concepts that need further explaining, but we'll have to wait until part 2 before we can discuss them. Let's move on for now and take a look at page boundaries.

3.4.2 Page boundaries and sizes

The size of a page is stored in the `/MediaBox`, which was an entry listed as required in table 3.3. You can get this value as a `PdfArray` using `pageDict.getAsArray(PdfName.MEDIABOX);` but a more programmer-friendly way is shown in code sample 3.8.

Code sample 3.8: C0305_PageBoundaries

```
1  PdfReader reader =
2      new PdfReader("src/main/resources/pages.pdf");
3  show(reader.getPageSize(1));
4  show(reader.getPageSize(3));
5  show(reader.getPageSizeWithRotation(3));
6  show(reader.getPageSize(4));
7  show(reader.getPageSizeWithRotation(4));
```

We see two convenience methods, named `getPageSize()` and `getPageSizeWithRotation()`. These methods return a `Rectangle` object that is passed to the `show()` method (see code sample 3.9).

Code sample 3.9: C0305_PageBoundaries

```
1  public static void show(Rectangle rect) {
2      System.out.print("llx: ");
3      System.out.print(rect.getLeft());
4      System.out.print(", lly: ");
5      System.out.print(rect.getBottom());
6      System.out.print(", urx: ");
7      System.out.print(rect.getRight());
8      System.out.print(", lly: ");
9      System.out.print(rect.getTop());
10     System.out.print(", rotation: ");
11     System.out.println(rect.getRotation());
12 }
```

Let's discuss the difference between `getPageSize()` and `getPageSizeWithRotation()` by examining the output of code sample 3.8:

```
llx: 0.0, lly: 0.0, urx: 595.0, lly: 842.0, rotation: 0
llx: 0.0, lly: 0.0, urx: 595.0, lly: 842.0, rotation: 0
llx: 0.0, lly: 0.0, urx: 842.0, lly: 595.0, rotation: 90
llx: 0.0, lly: 0.0, urx: 842.0, lly: 595.0, rotation: 0
llx: 0.0, lly: 0.0, urx: 842.0, lly: 595.0, rotation: 0
```

The first page in the `pages.pdf` document is an A4 page in portrait orientation. If you'd extract the `/MediaBox` array, you'd get `[0 0 595 842]`.

3.4.2.1 Pages in landscape

Page 3 (see line 4 in code sample 3.8) is also an A4 page, but it's oriented in landscape. The `/MediaBox` entry is identical to the one used for the first page `[0 0 595 842]`, and that's why `getPageSize()` returns

the same result. The page is in landscape, because the `\Rotate` entry in the page dictionary is set to `90`. Possible values for this entry are `0` (which is the default value if the entry is missing), `90`, `180` and `270`. The `getPageSizeWithRotation()` method takes this value into account. It swaps the width and the height so that you're aware of the difference. It also gives you the value of the `/Rotate` entry.

Page 4 also has a landscape orientation, but in this case, the rotation is mimicked by adapting the `/MediaBox` entry. In this case the value of the `/MediaBox` is `[0 0 842 595]` and if there's a `/Rotate` entry, its value is `0`. That explains why the output of the `getPageSizeWithRotation()` method is identical to the output of the `getPageSize()` method.

3.4.2.2 The CropBox and other page boundaries

Looking at figure 3.5, we see the page labels we've discussed in section 3.3.1, we see the different orientations discussed in section 3.4.2.1, and we see something strange happening to the *Hello World* text on page 5.

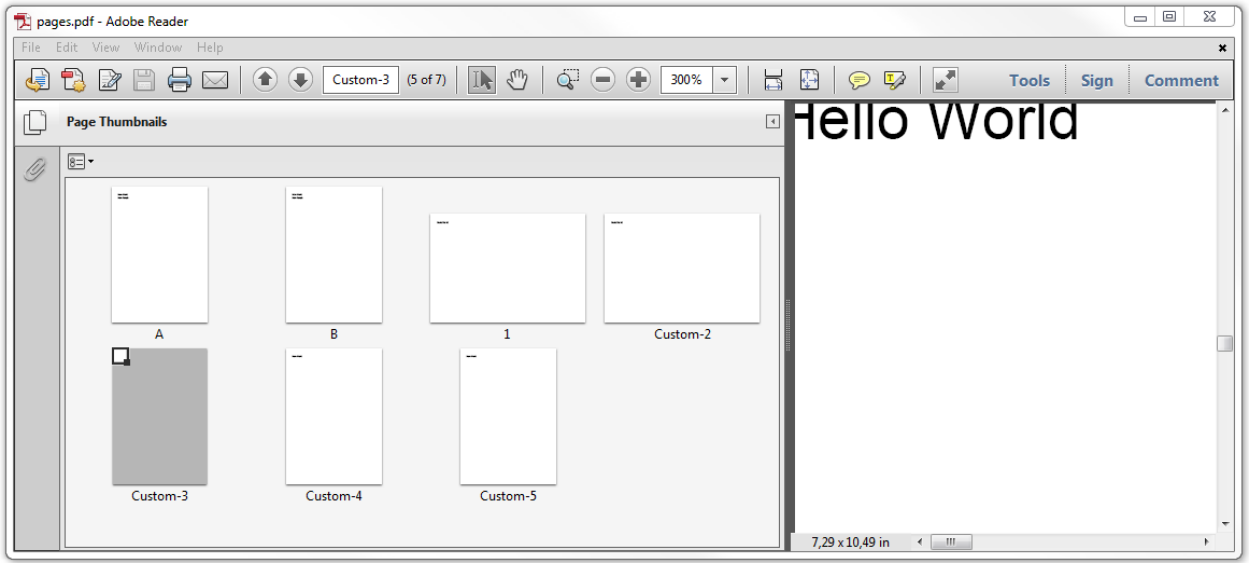


Figure 3.5: Pages and page boundaries

The text is cropped because there's a `/CropBox` entry in the page dictionary. As shown in code sample 3.10, you can get the cropbox using the `getCropBox()` method.

Code sample 3.10: C0305_PageBoundaries

```
1 show(reader.getPageSize(5));
2 show(reader.getCropBox(5));
```

These two lines result in the following output:

```
llx: 0.0, lly: 0.0, urx: 595.0, lly: 842.0, rotation: 0
llx: 40.0, lly: 40.0, urx: 565.0, lly: 795.0, rotation: 0
```

Let's consult ISO-32000-1 or -2 to find out the difference between the `/MediaBox` returned by the `getPageSize()` method and the `/CropBox` returned by the `getCropBox()` method.

- *The media box*— defines the boundaries of the physical medium on which the page is to be printed. It may include any extended area surrounding the finished page for bleed, printing marks, or other such purposes. It may also include areas close to the edges of the medium that cannot be marked because of physical limitations of the output device. Content falling outside this boundary may safely be discarded without affecting the meaning of the PDF file.
- *The crop box*— defines the region to which the contents of the page shall be clipped (cropped) when displayed or printed. Unlike the other boxes, the crop box has no defined meaning in terms of physical page geometry or intended use; it merely imposes clipping on the page contents. However, in the absence of additional information, the crop box determines how the page's contents shall be positioned on the output medium.

Summarized: the media box is the working area on your page, but only the content inside the crop box will be visible.

FAQ: I've added content to a page and it isn't visible

Maybe you're adding the content outside the visible area. The lower-left corner of the rectangle defining the media box and the crop box (if present) doesn't necessarily correspond with the coordinate `x = 0; y = 0`. You could easily have a media box defined like this: `[595 842 1190 1684]`. This is still an A4 page, but if you add a watermark at the coordinate `x = 397.5; y = 421`, that watermark won't be visible as it's outside the visible area of the page. If `rect` is the visible area of a page, you could center the watermark using these formulas for `x` and `y`:

```
float x = rect.getLeft() + rect.getWidth() / 2f;
float y = rect.getBottom() + rect.getHeight() / 2f;
```

These formulas calculate the coordinate of the middle of the page.

The definition of the media box and the crop box mention that a page can have some other boxes too:

- *The bleed box*— defines the region to which the contents of the page shall be clipped when output in a production environment. This may include any extra bleed area needed to accommodate the physical limitations of cutting, folding, and trimming equipment. The actual printed page may include printing marks that fall outside the bleed box.
- *The trim box*— defines the intended dimensions of the finished page after trimming. It may be smaller than the media box to allow for production-related content, such as printing instructions, cut marks, or color bars.
- *The art box*— defines the extent of the page's meaningful content (including potential white space) as intended by the page's creator.

If present, you'll find these boxes in the page dictionary by the following names: `/Bleedbox`, `/TrimBox` and `/ArtBox`. Code sample 3.11 shows how to obtain the media box and art box of page 7:

Code sample 3.11: C0305_PageBoundaries

```
show(reader.getBoxSize(7, "media"));
show(reader.getBoxSize(7, "art"));
```

The resulting output looks like this:

```
llx: 0.0, lly: 0.0, urx: 595.0, lly: 842.0, rotation: 0
llx: 36.0, lly: 36.0, urx: 559.0, lly: 806.0, rotation: 0
```

The `getBoxSize()` method accepts the following values for the `boxName` parameter: `media`, `crop`, `bleed`, `trim` and `art`. As you probably noticed, it can be used as an alternative for the `getPageSize()` and `getCropBox()` method.

All boxes, except for the media box can be visualized in an interactive PDF processor. This is done using a box color information dictionaries for each boundary. These box color information dictionaries are stored in the `/BoxColorInfo` entry of the page dictionary.

3.4.2.3 The measurement unit

Let's consult the PDF specification to find out what it says about the measurement unit:

ISO-32000-1 states: *The default for the size of the unit in default user space (1 / 72 inch) is approximately the same as a point, a unit widely used in the printing industry. It is not exactly the same, however; there is no universal definition of a point.* In short: 1 in. = 25.4 mm = 72 user units (which roughly corresponds to 72 pt).

The media box of the pages we've discussed so far was `[0 0 595 842]`. By default, this corresponds with the standard page size A4, measuring 210 by 297 mm (or 8.27 by 11.69 in), but that's not always the case. Figure 3.6 shows pages 5 and 6 of the `pages.pdf` document.

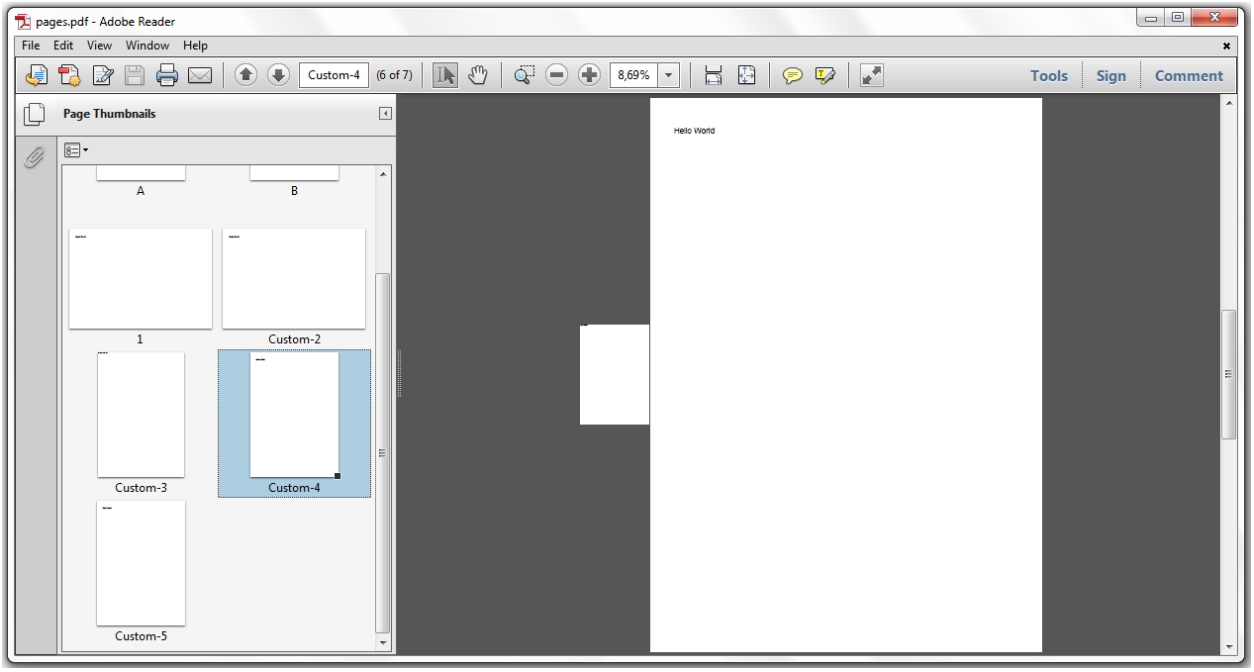


Figure 3.6: Pages with a different user unit

If we'd examine page 6, we'd discover that the media box is the only page boundary. It's defined as`[0 0 595 842]`, but when we look at the document properties, we see that the page size is 41.32 by 58.47 in. which is about 5 times bigger than the A4 page we expected. This difference is caused by the fact that a different user unit was used. In code sample 3.12, we get the `/UserUnit` value from the page dictionary of page 6. The output of this code sample is indeed 5.

Code sample 3.12: C0305_PageBoundaries

```
1 PdfDictionary page6 = reader.getPageN(6);
2 System.out.println(page6.getAsNumber(PdfName.USERUNIT));
```

Theoretically, you could create pages of any size, but the PDF specification imposes limits depending on the PDF version of the document. Table 3.5 shows how the implementation limits changed throughout the history of PDF.

Table 3.5: Possible entries in a resources dictionary

PDF version	Minimum page size	Maximum page size
PDF 1.3 and earlier	72 x 72 units (1 x 1 in.)	3240 x 3240 units (45 x 45 in.)
PDF 1.4 and later	3 x 3 units (approx. 0.04 x 0.04 in.)	14,400 x 14,400 units (200 x 200 in.)

Changing the user unit has been possible since PDF 1.6. The minimum value of the user unit is 1 (this is the default; 1 unit = 1/72 in.); the maximum value as defined in PDF 1.7 is 75,000 points (1 unit = 1042 in.).

The PDF ISO standards don't restrict the the range of supported values for the user unit, but says that the value is implementation-dependent.

Let's conclude that the biggest PDF page you can create without hitting any viewer implementations measures 15,000,000 x 15,000,000 in or 381 x 381 km. That's almost 5 times the size of Belgium (the country where iText was born).

3.4.3 Annotations

Figure 3.7 shows the last page in the `pages.pdf` document. If you hover over the words *Hello World*, a link to *http://maps.google.com* appears. If you click the small page icon, a post-it saying *This is a post-it annotation* pops up. These are annotations.

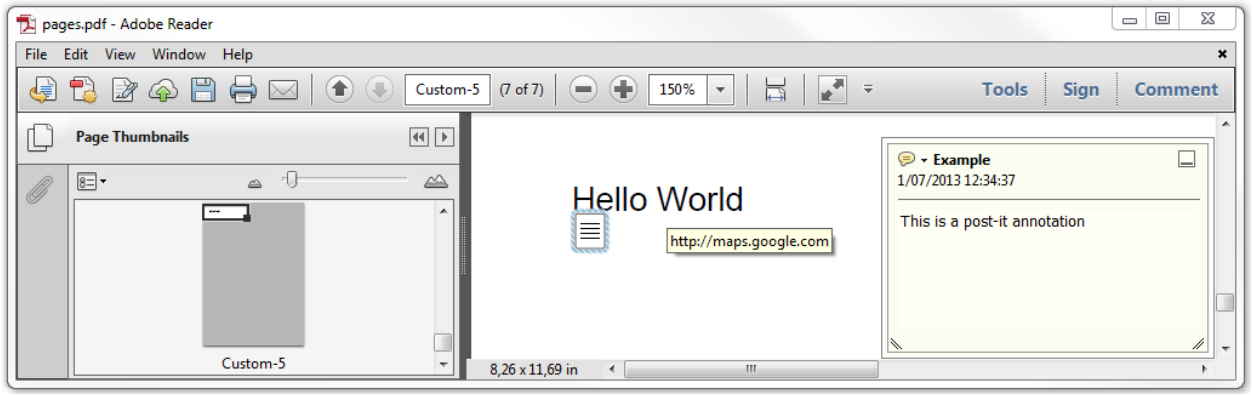


Figure 3.7: Annotations

If a page contains annotations, you'll find them in the `/Annots` entry of the page dictionary. This is an array of annotation dictionaries.

In figure 3.7, we're dealing with a `/Text` and a `/Link` annotation. Other types include *line*, *stamp*, *rich media*, *watermark*, and many other annotations. We'll get a closer look at all of these types in chapter 7. For now, it's sufficient to know that an annotation is an interactive object associated with a page.

Annotations aren't part of the content stream of a page. PDF viewers will always always render visible annotations on top of all the other content of a page.

Code sample 3.13 shows how to obtain the annotation dictionaries from page 7's page dictionary.

Code sample 3.13: C0306_PageAnnotations

```
PdfArray annots = page.getAsArray(PdfName.ANNOTS);
for (int i = 0; i < annots.size(); i++) {
    System.out.println("Annotation " + (i + 1));
    showEntries(annots.getAsDict(i));
}
```

We reuse the `showEntries` method from code sample 3.1, and this gives us the following result:

```
Annotation 1
/Contents: This is a post-it annotation
/Subtype: /Text
/Rect: [36, 768, 56, 788]
/T: Example
Annotation 2
/C: [0, 0, 1]
/Border: [0, 0, 0]
/A: Dictionary
/Subtype: /Link
/Rect: [66.67, 785.52, 98, 796.62]
```

The first annotation is a simple text annotation with title *Example* and contents *This is a post-it annotation*. The `/Rect` entry defines the coordinates of the clickable icon.

The second one is a link annotation. The `/C` entry defines the color of the border, but as the third element of the `/Border` array is `0`, no border is shown. You'll learn all about the different properties available for annotations in chapter 7.

Code sample 3.14 allows us to obtain the keys of the value of the `/A` entry, an action dictionary.

Code sample 3.14: C0306_PageAnnotations

```
PdfDictionary link = annots.getAsDict(1);
showEntries(link.getAsDict(PdfName.A));
```

The resulting output looks like this:

```
/URI: http://maps.google.com
/S: /URI
```

The action is of type `/URI`. You can trigger it by clicking the rectangle defined by `/Rect` in the link annotation.

Let's take a look at some other possible entries of the page dictionary before we return to the document catalog.

3.4.4 Other entries of the page dictionary

Table 3.6 lists entries of the page dictionary we haven't discussed so far.

Table 3.6: Optional entries of the page dictionary

Key	Type	Description
Metadata	stream	A stream containing XML metadata for the page in XMP format. This entry is also available in the document catalog. We'll discuss it in the next section.
AF	array	References to embedded files associated with this page. This entry is also available in the document catalog. We'll discuss this in the next section.
AA	dictionary	Additional actions to be performed when the page is opened or closed. This entry is also available in the document catalog. We'll discuss it in the next section.
StructParents	integer	Required if the page contains structural items. We'll discuss this in chapter 6.
Tabs	name	A name containing the tab order used for the annotations on the page. Possible values are: - `/R` for row order, - `/C` for column order, - `/S` for structure order, - `/A` for annotations array order (PDF 2.0) and - `/W` for widget order (PDF 2.0).
Thumb	stream	A stream object that defines the page's thumbnail image.
VP	array	An array of viewport dictionaries. This is outside the scope of this book.
Dur	number	The page's display duration in seconds during presentations.
Trans	dictionary	A transition dictionary describing the effect when opening the page in a viewer.
PZ	number	The preferred zoom factor for the page. For more info, see next section.
PresSteps	dictionary	Used in the context of sub-page navigation while presenting a PDF.
PieceInfo	dictionary	A page piece dictionary for the page. This entry is also available in the document catalog. We'll briefly discuss it in the next section.
LastModified	date	Required if **PieceInfo** is present. Contains the date and time when the page's contents were recently modified.
B	array	An array containing indirect references to article beads.
DPart	dictionary	An indirect reference to a DPart dictionary.
OutputIntents	array	This entry is also available in the document catalog.
SeparationInfo	dictionary	Information needed to generate color separations for the page.

Table 3.6: Optional entries of the page dictionary

Key	Type	Description
Group	dictionary	Used in the context of transparency group XObject.
ID	byte string	A digital identifier used in the context of Web Capture.
TemplateInstantiated	name	Required if this page was created from a named page object, in which case it's the name of the originating page obejct.

Many of these entries are outside the scope of this book, or can only described only briefly here, please consult ISO-32000-1 or -2 for more info. Now it's high time to take a closer look at the document catalog.

3.5 Optional entries of the Document Catalog Dictionary

We've already used the `getCatalog()` method in section 3.3.2 to get the root of the page tree. We learned that the document catalog has two required entries, `/Type` and `/Pages`. In this section, we'll discuss the optional entries of the root dictionary.

3.5.1 The names dictionary

We were already introduced to the concept of a number tree when we discussed page labels in section 3.3.1. When the keys or such a tree structure consist of strings instead of numbers, we talk about a name tree.

What is a name tree?

A name tree serves a similar purpose to a dictionary, associating keys and values. Based on the fact that we call this structure a name tree, you may expect that the keys are names, but they aren't. The keys are strings, they are ordered, and they are structured as a tree (involving branches and leaves). The leaves are stored in an array that looks like this `[key1 value1 key2 value2 ... keyN valueN]` where the keys are strings sorted in alphabetical order and the values are either references to a string, array, dictionary or stream, or direct objects in case of null, boolean, number or name values.

A name tree —or a number tree for that matter— usually consists of the following elements:

- The *root node* is a dictionary that refers to intermediate or leaf nodes in its `/Kids` entry. Alternatively, it can have a `/Names` entry in the case of a name tree, and a `/Nums` entry in the case of a number tree.
- An *intermediate node* is a dictionary with a `/Kids` entry referring to an array of intermediate nodes or leave nodes.
- A *leaf node* is a dictionary that has a `/Names` entry in the case of a name tree, and a `/Nums` entry in the case of a number tree.

Each intermediate and leaf node also has a `/Limits` entry, which is an array of two strings in case of a name tree, and an array of two integers in the case of a number tree. The elements of the array specify the least and greatest keys present in the node or its subnodes.

The document's name dictionary, which is the `/Names` entry of the document catalog, refers to one or more name trees. Each entry in the name dictionary is a name tree for a specific category of objects that can be referred to by name rather than by object reference. These are the most important categories:

- `/Dests`— maps name strings to destinations. We'll learn more about this in the next section.
- `/AP`— maps name strings to annotation appearance streams. We've already seen some simple annotations in section 3.4.3. In chapter 7, we'll create a custom appearance for some more complex annotations. Such an appearance could be referred to by name.
- `/JavaScript`— maps name strings to document-level JavaScript actions. We've learned about a simple URI action, but soon we'll find out that we can also use JavaScript to program custom actions.
- `/EmbeddedFiles`— maps name strings to file specifications for embedded file streams.

If you study ISO-32000-1 or -2, you'll discover that there are more categories, but they are out of scope of this book. In the next section, we'll take a look at a name tree containing named destinations.

3.5.2 Document navigation and actions

In section 3.3, we've examined the page tree, and we've learned how to navigate through a document programmatically. Now we're going to take a look at some ways we can help the end user navigate through the document using a PDF viewer.

3.5.2.1 Destinations

There are different ways to define destinations. You can associate names with destinations, or you can use explicit destinations.

3.5.2.1.1 Named destinations When we discussed the name dictionary, we listed `/Dests` as one possible category of named objects. This name tree defines *named destinations* with string values as keys.

The concept of using a name tree for named destinations was introduced in PDF 1.2. In PDF 1.1, named destinations were defined using names instead of string. They were stored in a `/Dests` entry of the document catalog. It referred to a dictionary with the names as keys and their destination dictionary as values. Most of the PDF producers have abandoned using this Catalog entry in favor of the `/Dests` entry in the name tree.

Code sample 3.15 shows how we get the first named destination from the name tree in the name dictionary.

Code sample 3.15: C0306_DestinationsOutlines

```
PdfReader reader = new PdfReader("src/main/resources/primes.pdf");
PdfDictionary catalog = reader.getCatalog();
PdfDictionary names = catalog.getAsDict(PdfName.NAMES);
PdfDictionary dests = names.getAsDict(PdfName.DESTS);
PdfArray array = dests.getAsArray(PdfName.NAMES);
System.out.println(array.getAsString(0));
System.out.println(array.getAsArray(1));
```

This is the resulting output:

```
Prime101
[210 0 R, /XYZ, 36, 782, 0]
```

This means that we can now use the string `Prime101` as a name to refer to a destination on the page described by the indirect object with number 210. Code sample 3.16, shows a short-cut to get the same information using much less code:

Code sample 3.16: C0306_DestinationsOutlines

```
Map<String, String> map =
    SimpleNamedDestination.getNamedDestination(reader, false);
System.out.println(map.get("Prime101"));
```

The `SimpleNamedDestination` class can be used to obtain a map of all named destinations. If you use `false` as the second parameter of the `getNamedDestination()` method, you get the named destinations stored in the `/Dests` entry of the names dictionary —destinations referred to by strings (since PDF 1.2). If you use `true`, you get the named destinations stored in the `/Dests` entry —destinations referred to by names (PDF 1.1).

The `SimpleNamedDestination` class also has an `exportToXML()` method that allows you to export the named destinations to an XML file.

In code sample 3.14 and 3.15, the destination is defined using four parameters: the name `/XYZ` and three numbers, the x-position 36, the y-position 782, and the zoom factor 0 (no zoom). This is an explicit destination.

3.5.2.1.2 Explicit destinations Destinations are defined as an array consisting of a reference to a page dictionary and a location on a page, optionally including the zoom factor. Table 3.7 lists the different names and parameters that can be used to define the location and zoom factor.

Table 3.7: Destination syntax

Type	Extra parameters	Description
`/Fit`	-	The page is displayed with its contents magnified just enough to fit the document window, both horizontally and vertically.
`/FitB`	-	The page is displayed magnified just enough to fit the bounding box of the contents (the smallest rectangle enclosing all of its contents).
`/FitH`	*top*	The page is displayed so that the page fits within the document window horizontally (the entire width of the page is visible). The extra parameter specifies the vertical coordinate of the top edge of the page.
`/FitBH`	*top*	This option is almost identical to `/FitH`, but the with of the bounding box of the page is visible. This isn't necessarily the entire width of the page.
`/FitV`	*left*	The page is displayed so that the page fits within the document window vertically (the entire height of the page is visible). The extra parameter specifies the horizontal coordinate of the left edge of the page.
`/FitBV`	*left*	This option is almost identical to `/FitV`, but the height of the bounding box of the page is visible. This isn't necessarily the entire height of the page.
`/XYZ`	*left, top, zoom*	The *left* parameter defines an `x` coordinate, *top* defines a `y` coordinate, and *zoom* defines a zoom factor. If you want to keep the current `x` coordinate, the current `y` coordinate, or zoom factor, you can pass negative values or `0` for the corresponding parameter.
`/FitR`	*left, bottom, right, top*	The parameters define a rectangle. The page is displayed with its contents magnified just enough to fit this rectangle. If the required zoom factors for the horizontal and the vertical magnification are different, the smaller of the two is used.

Destinations may be associated with outline items, link annotations, or actions. Let's start with outlines.

3.5.2.2 The outline tree

Figure 3.8 shows a PDF with bookmarks for every prime number.

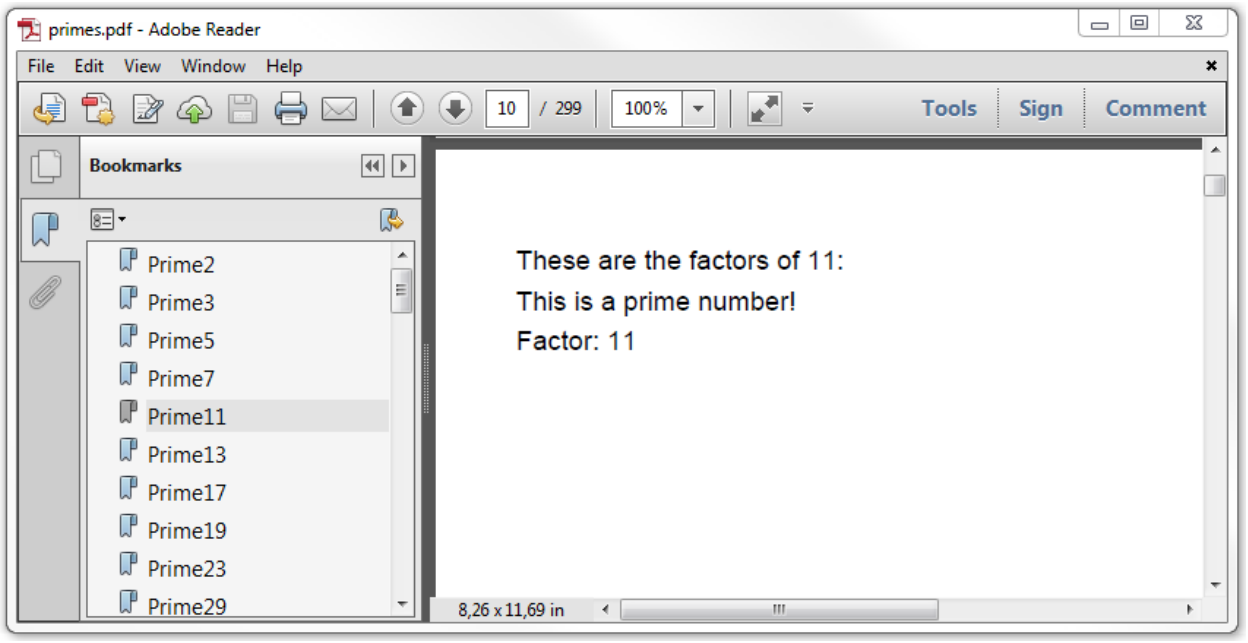


Figure 3.8: A PDF with bookmarks

Bookmarks can have a tree-structured hierarchy, but in this case, we only have a simple list of items. If we click an item, you either jump to a destination, which is the case if you click *Prime11*, or you trigger an action.

The root of this bookmarks tree is specified by the `/Outlines` entry of the document catalog. In code sample 3.17, we extract the root, its first leaf and its last leaf as a dictionary:

Code sample 3.17: C0306_DestinationsOutlines

```
1 PdfDictionary outlines = catalog.getAsDict(PdfName.OUTLINES);
2 System.out.println("Root:");
3 showEntries(outlines);
4 System.out.println("First:");
5 showEntries(outlines.getAsDict(PdfName.FIRST));
6 System.out.println("Last:");
7 showEntries(outlines.getAsDict(PdfName.LAST));
```

The output of this code sample looks like this:

```
Root:
/Type: /Outlines
/Count: 62
/Last: 754 0 R
/First: 693 0 R
First:
/Next: 694 0 R
/Parent: 692 0 R
/Title: Prime2
/Dest: [1 0 R, /Fit]
Last:
/Parent: 692 0 R
/Title: Prime293
/Dest: [614 0 R, /Fit]
/Prev: 753 0 R
```

This corresponds with what we see when we look at the outline tree using RUPS. See figure 3.9.

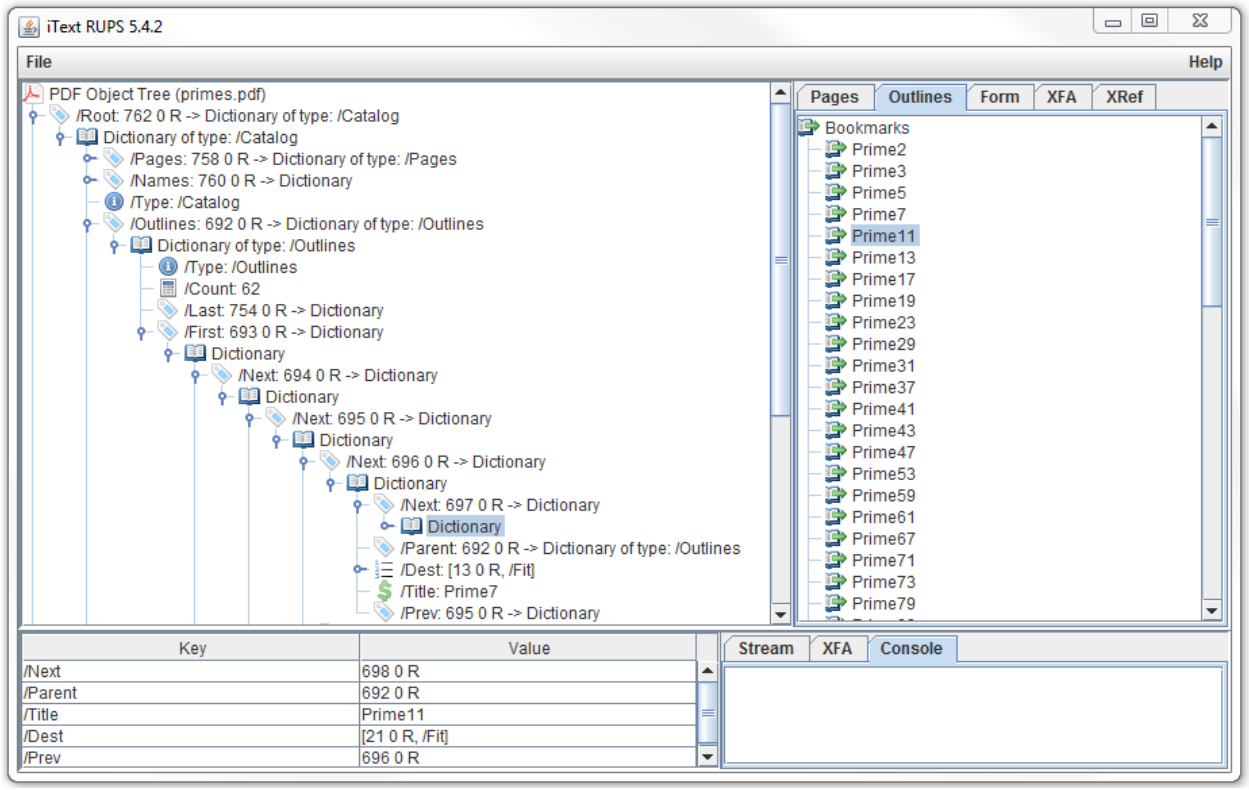


Figure 3.9: The outline tree

Note that all the entries at the same level are chained to eachother as a linked list, with the `/Next` entry referring to the next item and the `/Prev` entry referring to the previous item.

The root of the outline tree is an outline dictionary. Table 3.8 explains the different entries.

Table 3.8: Entries in the outline dictionary

Key	Type	Value
Type	name	If present, the value must be `/Outlines`.
First	dictionary	An indirect reference to the first top-level outline item.
Last	dictionary	An indirect reference to the last top-level outline item.
Count	integer	The total number of open outline items. The value can't be negative. It's to be omitted if there are no open items.

Table 3.9 shows the possible entries for outline item dictionaries.

Table 3.9: Entries in the outline dictionary

Key	Type	Value
Title	string	The text for the outline as displayed in the bookmarks panel.
Parent	dictionary	The parent of this outline, this is the outline dictionary for top-level outline items, otherwise it's another outline item.
Prev	dictionary	The previous item at the current outline level. Not present for the first one.
Next	dictionary	The next item at the current outline level. Not present for the last one.
First	dictionary	An indirect reference to the first descendant of which this item is the parent.
Last	dictionary	An indirect reference to the last descendant of which this item is the parent.
Count	integer	The total number of open outline items. If the outline item is closed, a negative value is used with as absolute value the number of descendants that would be opened if the item was open.
Dest	name, string or array	In the case of named destinations a name or string will be used. In the case of explicit destinations an array is used. See section 3.5.2.1.
A	dictionary	The action that needs to be performed. See next section.
SE	dictionary	The structure item to which the outline refers; see chapter 6.
C	array	An array of three numbers ranging from `0.0` to `1.0` representing an RGB color.
F	integer	The style used for the text: by default `0` for regular text, `1` for italic, `2` for bold, `3` for bold and italic.

Just like with named destinations where we had the `SimpleNamedDestination` convenience class, there's also a `SimpleBookmark` class that allows you to get more information about bookmarks, without having to fetch

all the outline item dictionaries. Code sample 3.18 gives you an example of how to create a `List` containing all the bookmarks.

Code sample 3.18: C0306_DestinationsOutlines

```
List<HashMap<String, Object>> bookmarks = SimpleBookmark.getBookmark(reader);
for (HashMap<String, Object> item : bookmarks) {
    System.out.println(item);
}
```

The output for the `primes.pdf` starts like this:

```
{Action=GoTo, Page=1 Fit, Title=Prime2}
{Action=GoTo, Page=2 Fit, Title=Prime3}
{Action=GoTo, Page=4 Fit, Title=Prime5}
{Action=GoTo, Page=6 Fit, Title=Prime7}
{Action=GoTo, Page=10 Fit, Title=Prime11}
```

The map consists of strings such as `Open`, `Title`, `Page`, `Color`, `Style`, `Kids`, and so on. In the case of the `Kids` entry, the object is a `List` with more `HashMap` elements.

The `SimpleBookmark` class also has an `exportToXML()` method that allows you to export the bookmarks to an XML file.

Note that iText uses the keyword `Action` to indicate that clicking a bookmark item is the equivalent of a `GoTo` action.

Some outline actions, such as JavaScript actions, aren't picked up by `SimpleBookmark`.

Let's take a closer look at actions in general.

3.5.2.3 Actions

An action in a PDF document is defined using an action dictionary that specifies what the viewer should do when the action is triggered. Table 3.10 shows the entries common to all action dictionaries.

Table 3.10: Entries common to all action entries

Key	Type	Value
Type	name	If present, the value must be `/Action`.
S	name	The type of action, see table 3.11.
Next	dictionary or array	The next action or sequence of actions that must be performed after this action. This allows actions to be *chained*.

Table 3.11 shows the pssible values for the /S entry.

Table 3.11: Action types

Action Type	Context	Description
Named	Nav.	Execute a predefined action.
GoTo	Nav.	Go to a destination in the current document.
GoToR	Nav.	Go to a destination in a remote document
URI	Nav.	Resolve a uniform resource identifier (URI).
GoToE	Nav.	Go to a destination in an embedded PDF file.
GoToDp	Nav.	Go to a document part.
Hide	Ann.	Set an annotation's Hidden flag.
Sound	Ann.	Play a sound.
Movie	Ann.	Play a movie.
Rendition	Ann.	Controls playing of multimedia content.
GoTo3DView	Ann.	Sets the current view of a 3D annotation.
RichMediaExecute	Ann.	Specifies a command to be sent to a rich media annotation's handler.
SubmitForm	Form	Send form data to a uniform resource locator (URL).
ResetForm	Form	Reset the fields in a form to their default values.
ImportData	Form	Import field values from a file.
SetOCGState	OCG	Sets the states of optional content groups.
JavaScript	Misc.	Execute a JavaScript script.
Launch	Misc.	Launch an application, usually to open a file.
Trans	Misc.	Updates the display of a document using a transition dictionary.
Thread	Misc.	Begin reading an article thread

We won't discuss all these types of action in detail. We'll look at some of the actions marked with *Nav.* which allow us to navigate through a document. We'll discuss some of the actions marked with *Ann.* triggering events related to annotations in chapter 7, and actions marked with *Form* in chapter 7. The OCG action will be discussed in chapter 6.

Actions can be triggered in many ways. For now we'll only look at two places where we can find actions.

- The /OpenAction entry in the catalog. Its value can be an array defining an explicit destination or an action dictionary. The action is triggered upon opening the document.
- The /AA entry that can occur in root, page, annotation and form field dictionaries. Its value is an *additional actions* dictionary defining actions that need to be executed in response to various trigger events.

We'll discuss the events that can be triggered from an annotation in chapter 7 and those that can be triggered from a form field in chapter 8. Table 3.12 shows the possible entries in the additional actions dictionary of a page dictionary.

Table 3.12: Entries in a page object's additional actions dictionary

Key	Type	Value
O	dictionary	An action that will be executed when the page is opened, for instance when the user navigates to it from the next or previous page.
C	dictionary	An action that will be executed when the page is closed, for instance when the user navigates away from it by clicking a link to another page.

Table 3.13 shows the possible entries in the additional actions dictionary of the document catalog.

Table 3.13: Entries in the document catalog's additional actions dictionary

Key	Type	Value
WC	dictionary	A JavaScript action to be executed before closing the document ("will close").
WS	dictionary	A JavaScript action to be executed before saving the document ("will save")
DS	dictionary	A JavaScript action to be executed after saving the document ("did save")
WP	dictionary	A JavaScript action to be executed before printing the document ("will print")
DP	dictionary	A JavaScript action to be executed after printing the document ("did print")

Just like an HTML file, a PDF document can contain JavaScript. There are some differences in the sense that you get extra objects that allow you to use functionality that is specific to a PDF viewer. We'll take a look at some JavaScript examples in chapter 7.

The "will save" action isn't triggered by a "Save As" operation. In practice this often means that it's only triggered in Adobe Acrobat, not in Adobe Reader.

Let's conclude this section about actions by looking at some of the actions marked with *Nav.* in table 3.11.

A *named action* is an action of which the value of the `/S` entry is `/Named`. It has an additional `/N` entry of which the value is one of the names listed in table 3.14.

Table 3.14: Named actions

Name	Action
NextPage	Go to the next page of the document.
PrevPage	Go to the previous page of the document.
FirstPage	Go to the first page of the document.
LastPage	Go to the last page of the document.

A *go to* action is an action of which the value of the /S entry is /GoTo. It must have a /D entry that can be a name or a string in case of named destinations, or an array in case of an explicit destination. Starting with PDF 2.0, there can also be an /SD entry to jump to a specific structure destination (see chapter 6).

The *remote go to* action is very similar. It's an action of which the value of the /S entry is /GoToR. It must have an /F entry of which the value is a file specification dictionary, as well as a /D entry. The value of the /D entry can be a name or a string in case of named destinations. When an explicit destination is defined, an array is used of which the first element isn't an indirect reference to a page dictionary, but instead the page number of the target destination.

There's an optional NewWindow entry of which the value is a Boolean. If true, the document will be opened in a new window, provided that the referring document is opened in a standalone PDF viewer. For instance: setting this entry to true won't open a new browser window if the document is opened in Adobe Reader's browser plug-in.

An *URI* action is an action of which the value of the /S entry is URI. It must have an URI entry defining the uniform resource identifier to resolve, for instance: a link to open a hyperlink in a browser.

When the Catalog of a document contains a /URI entry, it refers to a dictionary with a single entry, /Base, of which the value is the base URI that shall be used when resolving relative URI references throughout the document.

An URI action dictionary can have an /IsMap entry of which the value is a Boolean. If true, a mouse position will be added to the URI in the form of a query string containing the values of an x and a y coordinate.

Another way to send data to a server involves interactive forms.

3.5.3 Interactive forms

When you find an /AcroForm key in the root dictionary of a PDF file, your document is an interactive form. There are different flavors of forms in PDF.

1. One is based on *AcroForm* technology. These forms are defined using the PDF objects described in chapter 1, for instance an array of fields where each field is defined using a dictionary.
2. Another type of form uses the *XML Forms Architecture* (XFA). In this case, the PDF file acts as a container for a set of streams that form a single XML file defining the appearance as well as the data of the form.

3. Finally, there are also hybrid forms that contain both an AcroForm description and an XFA stream defining the form.

Forms are used for different purposes. One purpose can be to allow users to fill out a form manually and to submit the filled out data to a server. Another purpose could be to create a template that can be filled out in an automated process. For instance: one could create a PDF containing an AcroForm form that represents a voucher for an event. Such a PDF could have fields that act as placeholders for the name of an attendee, a date, a bar code, etc... These static placeholders can then be filled out automatically using data from a database. If you need dynamic fields, XFA is your only option. You could create an XFA template based on your own XML schema, and then inject different sets of XML data into this template.

We'll discuss these types of forms in more detail in chapter 8. For now, we'll just take a look at an example of each flavor and find out how to tell the different flavors apart.

3.5.3.1 AcroForm technology

Figure 3.110 shows an interactive form based on AcroForm technology. It contains text fields that can be filled out with a title of a movie, a director, etc. It also contains some check boxes and radio buttons. To the left, we see the same document opened in RUPS. We see that the Catalog has an ´/AcroForm´ entry. The value of the `/Fields` entry of this dictionary is an array that isn't empty. Its elements are AcroForm fields.

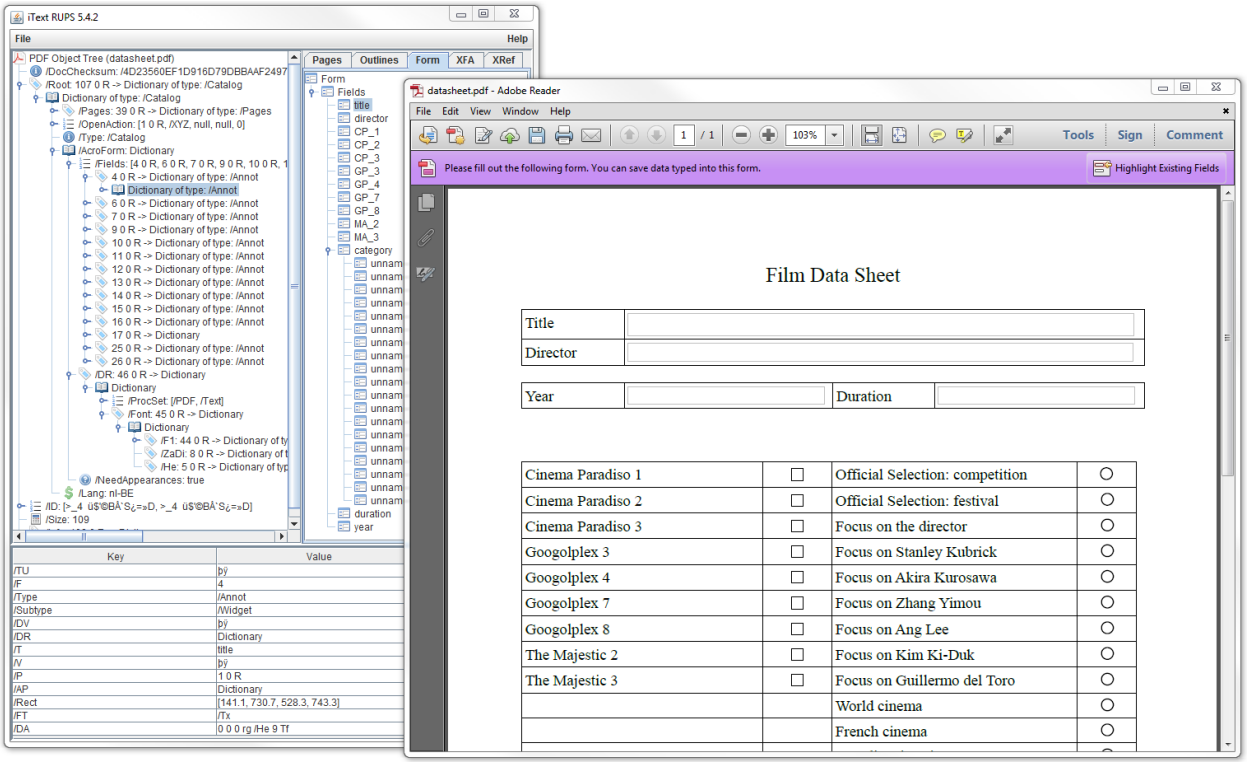


Figure 3.10: AcroForm technology

We selected one specific field in RUPS: the *title* field. In the dictionary panel, we see the combined field and annotation dictionary defining that field. The `/P` value refers to the page on which the widget annotation of

the field will appear; the /Rect entry defines the position of the field's annotation on that page. We'll discuss the other entries of this dictionary in chapter 8. For now, it's important to understand that every placeholder has its fixed place on a fixed page when using AcroForm technology. Placeholders can't resize automatically when their content doesn't fit the designated space.

We could test this form, by getting the AcroForm dictionary from the catalog, retrieving the fields array, and so on, but let's skip that, and use the AcroFields convenience class. See code sample 3.18.

Code sample 3.18: C0306_DestinationsOutlines

```
public static void inspectForm(File file) throws IOException {
    System.out.print(file.getName());
    System.out.print(": ");
    PdfReader reader = new PdfReader(file.getAbsolutePath());
    AcroFields form = reader.getAcroFields();
    XfaForm xfa = form.getXfa();
    System.out.println(xfa.isXfaPresent() ?
              form.getFields().size() == 0 ? "XFA form" : "Hybrid form"
            : form.getFields().size() == 0 ? "not a form" : "AcroForm");
    reader.close();
}
```

This code snippet can be used to check which type of form you're dealing with. It writes the name of the file to the System.out, followed by one of the following results: *XFA form*, *Hybrid form*, *not a form* or *AcroForm*. The AcroFields class inspects the /AcroForm entry, and checks if there's an XFA part. If not, it checks the number of fields. If there is no /AcroForm entry or if the /Fields array is empty, you don't have an AcroForm. In the example shown in figure 3.10, the method returns *AcroForm*.

The AcroForm technology supports four types of fields:

- *Button fields*— representing interactive controls on the screen that can be manipulate using a mouse, such as pushbuttons, check boxes and radio buttons.
- *Text fields*— consisting of boxes or spaces in which the users can enter text from the keyboard.
- *Choice fields*— containing several text items that can be selected, for instance list boxes and combo boxes.
- *Signature fields*— representing digital signatures and optional data for authenticating the name of the signer and the document's contents.

We'll cover button, text and choice fields in chapter 8, but digital signature fields are outside the scope of this book.

A digital signature involves a digital certificate, a timestamp, revocation information and so on. This information is needed as soon as you want to validate the signature. If this information is missing or expired, one can create an incremental update and add a dictionary to the document catalog containing the most up-to-date validation-related information. This dictionary is known as the Document Security Store (DSS) and it's used as the value for the /DSS entry of the root dictionary.

Digital signatures, including the Document Security Store, are discussed in great detail in the book "Sign your PDFs with iText[2]."

3.5.3.2 The XML Forms Architecture

Figure 3.11 shows such an example of a pure XFA form opened in Adobe Reader as well as in RUPS.

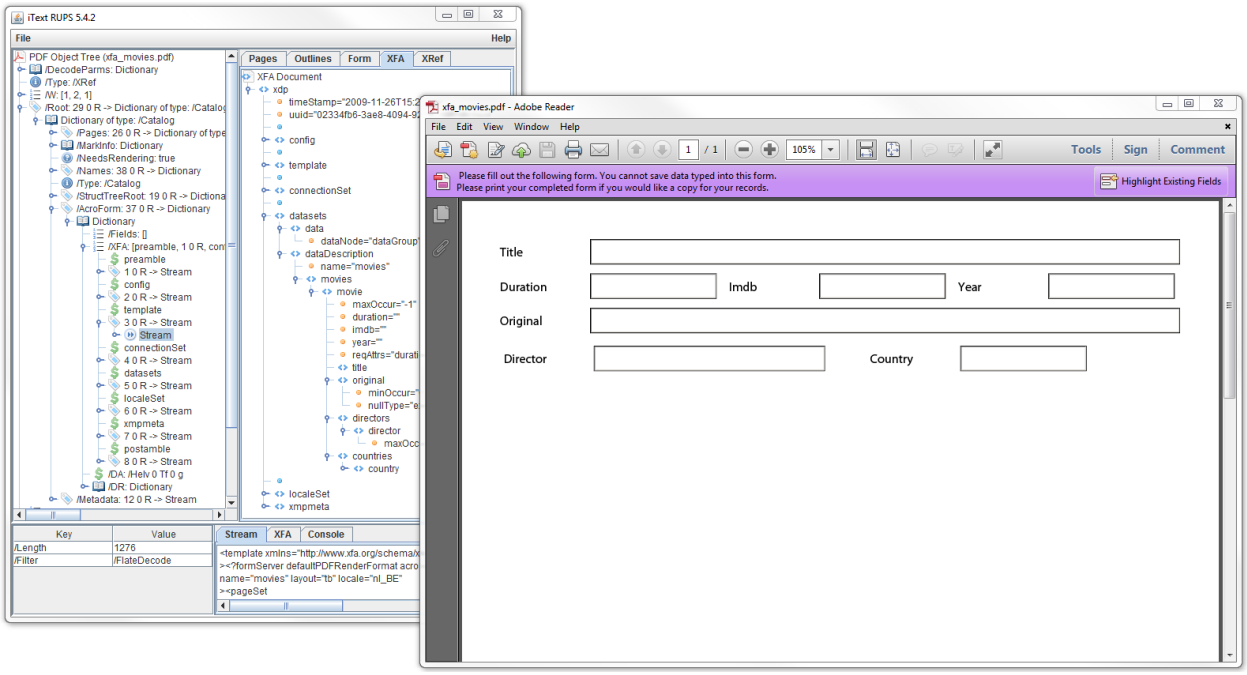


Figure 3.11: An empty XFA form

We detect a `/NeedsRendering` entry with value `true` in the Catalog. This means that the PDF viewer will have to parse XML on order to render the content of the document.

Not all PDF viewers support pure XFA forms. For instance: if you open a pure XFA form in Apple Preview, you'll see whatever content is available in the form of PDF syntax. This is usually a warning saying that your PDF viewer doesn't support this particular version of PDF documents.

The `/AcroForm` entry has an empty `/Fields` array. The form is defined in different streams that are to be concatenated to form a valid XML file.

The form shown in figure 3.11 is empty: it doesn't contain any data. Figure 3.12 shows the same form, but now filled out using an XML file that contains 120 movies. The document that originally counted only one page, now consists of 23 pages. The complete group of fields is repeated 120 times, once for each movie in the XML file. Some fields are repeated too, see for instance the *Country* field.

[2] https://leanpub.com/itext_pdfsign

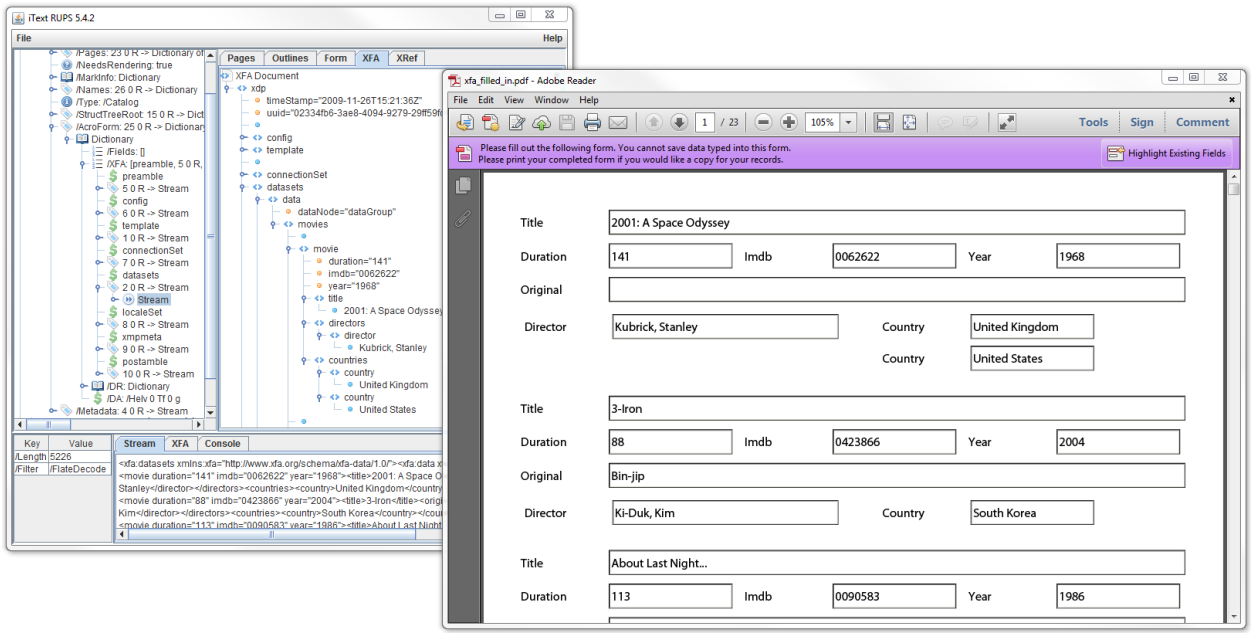


Figure 3.12: A filled out XFA form

As opposed to the static AcroForm documents, XFA forms are truly dynamic. You can use them for forms of templates of which you don't know the amount of data nor the size of the data up-front. XML forms also consist of text fields, choice fields, buttons and signature fields, but they are defined using XML, which provides much more flexibility.

3.5.3.3 Hybrid forms

Figure 3.13 shows a hybrid form. There's no `/NeedsRendering` entry in the Catalog. A PDF viewer doesn't need to parse any XML in order to render the form as the form is also defined using AcroForm technology

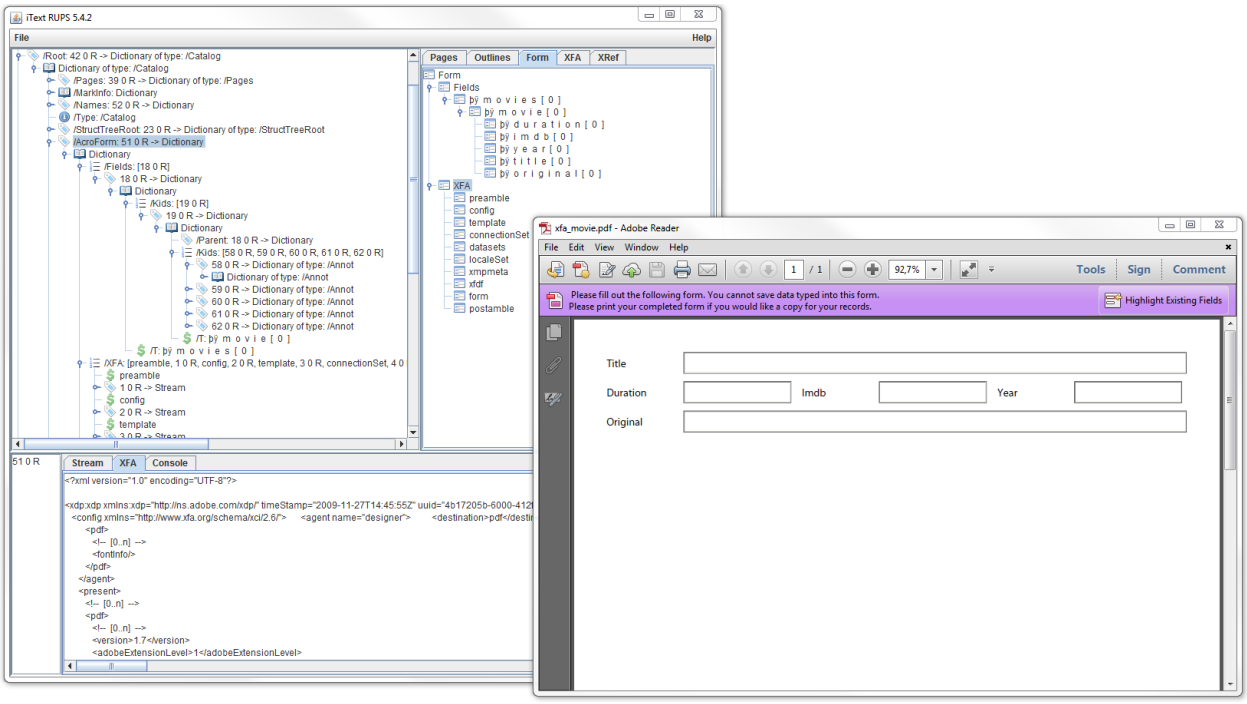


Figure 3.13: A hybrid form

Hybrid forms have the advantage that they can be viewed in PDF viewer that don't support XFA, but the disadvantage that the dynamic nature of XFA is lost.

3.5.4 Marked Content

In the next chapter, we'll take a closer look at the Adobe Imaging Model. We'll learn more about the operators that are needed to draw text, paths, shapes and images to a page. These operators only serve to make sure the visual representation of the document is correct.

There is also a set of operators that allow you to mark this content. Marked-content operators are used to identify a portion of a PDF content stream as an element of interest for a particular goal. For instance: a word drawn on a page doesn't necessarily know which line it belongs too. A line doesn't know which paragraph it belongs too. You can mark a sequence of text operators as a paragraph, so that software can discover the structure of your document. In this case, we add marked content operators to add structure to the document. When specific rules are followed when creating this structure, we say that the PDF is a Tagged PDF document.

FAQ: I've created a PDF with a table and now I want to extract the cells of that table

This is only possible if the PDF is tagged. If the document isn't tagged, it doesn't know there's a tabular structure on the page. What looks like a table to the human eye, is nothing more than a bunch of glyphs, lines and shapes drawn on a page. By introducing *marked content operators* into the content stream, you can mark all the different elements of a table in a way that software can detect rows, columns, headers, and so on.

Using marked content, you can also add object data. For example: if your PDF consists of a blueprint of a machine, you can use marked content operators to add specific properties for each machine part that is drawn. Marked content can also be used to make documents accessible. For instance: for each image in the document, you can add a short textual description so that people who are visually impaired can find out what the image represents. Another typical use case involves optional content. You can mark a sequence of PDF syntax in a way that it becomes visible or invisible, for instance depending on user interaction.

Several entries in the root dictionary refer to Marked Content. The `/MarkInfo` entry refers to a dictionary containing information about the document's use of Tagged PDF conventions. We'll discuss the `/StructTreeRoot` entry in detail in chapter 6. The same goes for the `/OCProperties` key which refers to optional content.

3.5.5 Embedded files

We've looked at the actual content of a document, at navigation information and at the structure of the content, but there's more. A document can also contain attachments, and these attachments can be organized in a special way.

There are different ways to attach a file to a PDF document. You can add an attachment using an annotation. In this case, you'll have a visible object on the page (for instance a paper clip) that can be clicked by the end user to open the attachment. We'll learn more about file attachment annotations in chapter 7.

You can also create document-level attachments also known as embedded files. You need to open the attachments panel in your PDF viewer to see these attachments. When double-clicking an attachment of a different format than PDF, you need an external viewer to open the attachment. See figure 3.14.

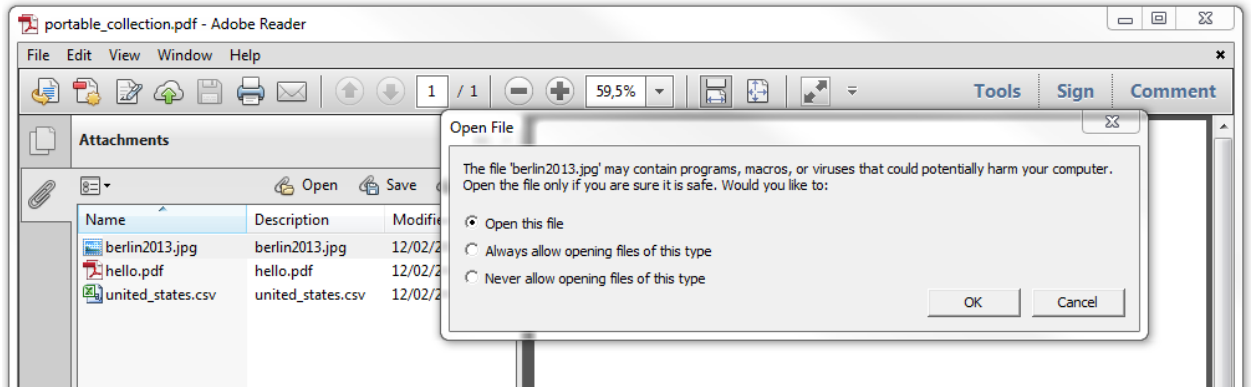


Figure 3.14: Document-level attachments

If you want an approach that is more integrated into the PDF Viewer, you may want to create a portable collection.

3.5.5.1 Portable Collections

Figure 3.15 shows the most simple type of portable collection you can create.

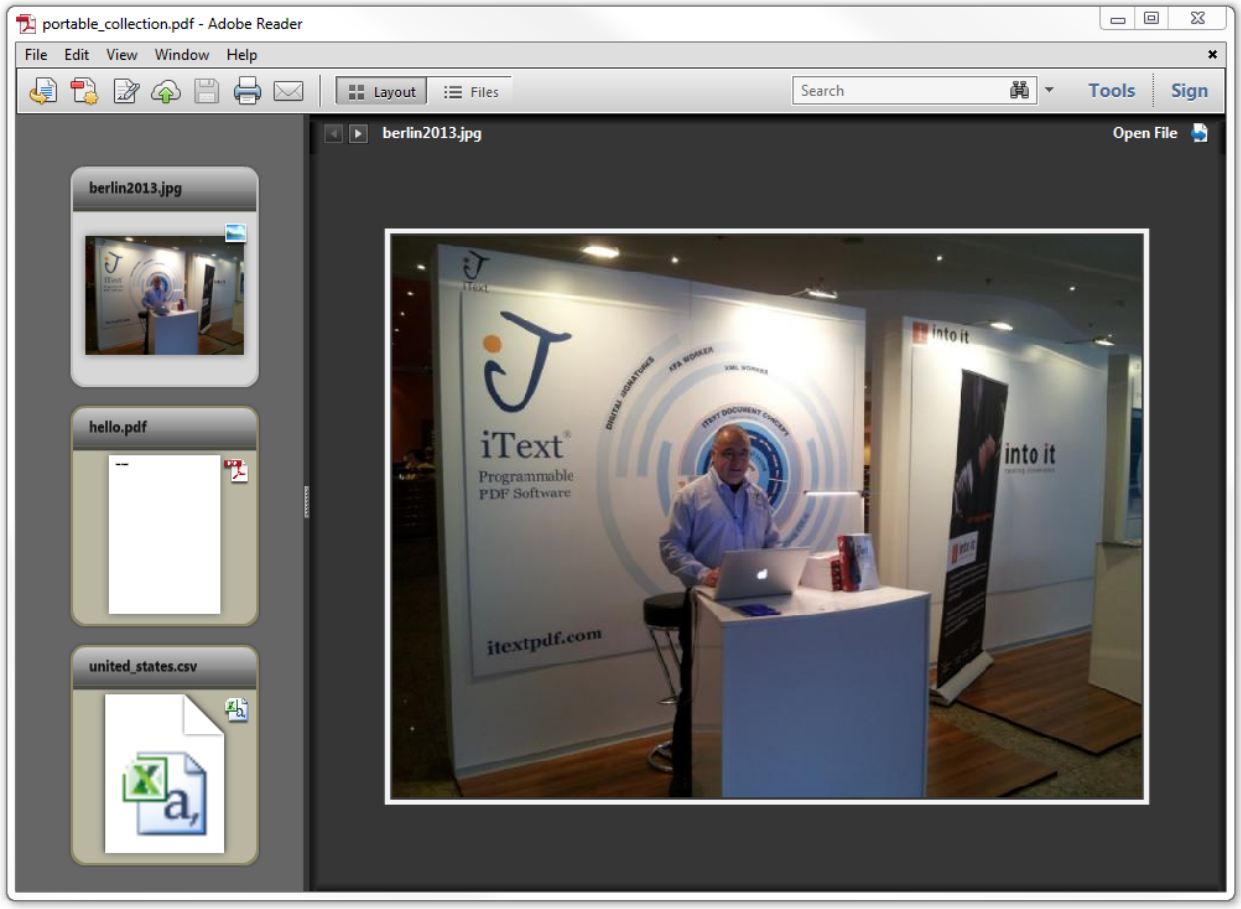


Figure 3.15: A portable collection

In this case, the PDF is defined as a portable collection or a PDF package.

When using a portable collection, the PDF acts as a container for different files, similar to a ZIP file. The advantage of having a PDF package over a ZIP file is the fact that some files can be rendered in Adobe Reader. For instance: you can view the JPG without having to open an external application. To open the CSV file however, you'll need an application such as Excel.

The difference between the PDF shown in figure 3.14 and the one shown in figure 1.15 consists of a single entry in the root dictionary.

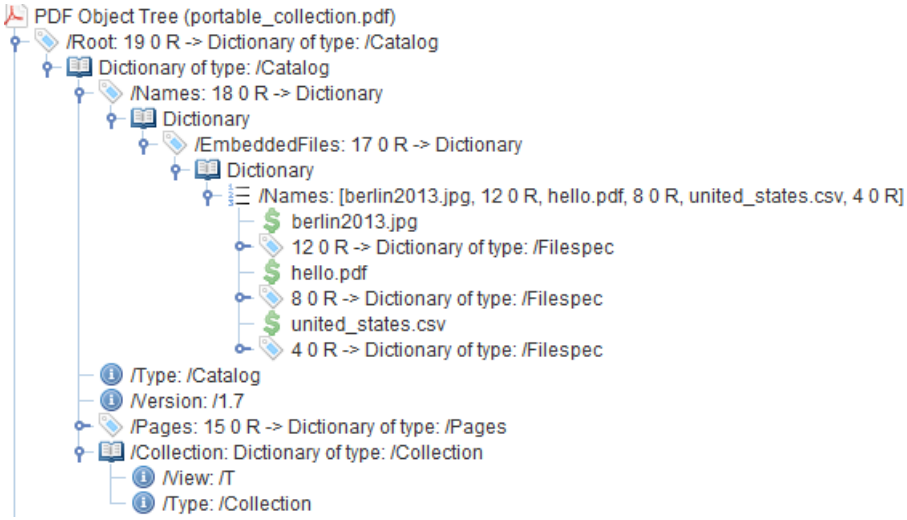


Figure 3.16: A portable collection

In both cases, there is a `/Names` entry with an `/EmbeddedFiels` name tree. In the second case, there is also a `/Collection` entry in the Catalog. In figure 3.16, the view type is `/T` for Tile. There are different types of portable collections. Instead of showing thumbnails, you can provide a table consisting of rows you can populate with data of your choice. You can also create your own Flash component to navigate through the different documents.

3.5.5.2 Associated files

When attaching files to a document, the PDF isn't aware of any relationship between the document and the attachment. To the document, the attachment is merely a sequence of bytes, unless you define an *associated files relationship*. See the `/AFRelationship` key in the filespecification of the file `united_states.csv` in figure 3.17.

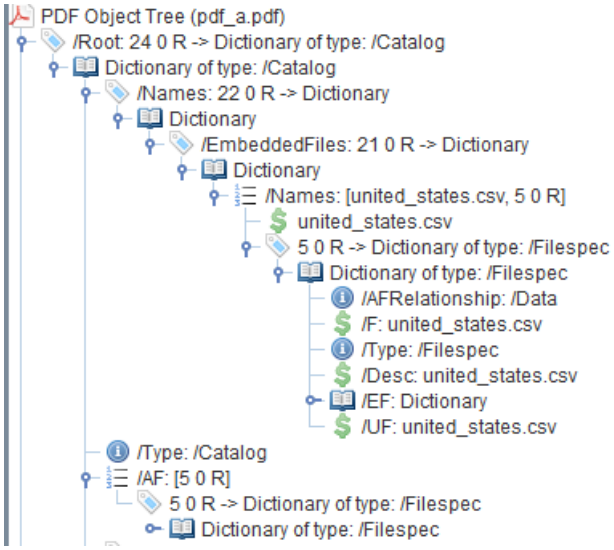


Figure 3.17: A portable collection

The file that is opened in RUPS contains a list of US states that was created based on the data file `united_states.csv`. We add an extra reference to this file specification in the array of associated files; see the value of the `/AF` key in the root dictionary. Defining the relationship between a document and its attachments is a mandatory requirement when you need to comply with the PDF/A-3 standard.

3.5.6 Viewer preferences

The `/PageLayout`, `/PageMode` and `/ViewerPreferences` entries refer to the way the document must be presented on the screen when the document is opened. The `/PageLayout` entry tells the PDF viewer how pages should be displayed. Possible values are:

- `/SinglePage`— Display one page at a time.
- `/OneColumn`— Display the pages in one column.
- `/TwoColumnLeft`— Display the pages in two columns, with the odd-numbered pages on the left.
- `/TwoColumnRight`— Display the pages in two columns, with the odd-numbered pages on the right.
- `/TwoPageLeft`— Display the pages two at a time, with the odd-numbered pages on the left.
- `/TwoPageRight`— Display the pages two at a time, with the odd-numbered pages on the right.

The `/PageMode` entry defines which panel —if any— needs to be opened next to the actual content. You can also use it to have the document opened in full screen view. Possible values are:

- `/UseNone`— Neither document outline nor thumbnail images visible.
- `/UseOutlines`— Neither document outline nor thumbnail images visible.
- `/UseThumbs`— Neither document outline nor thumbnail images visible.
- `/FullScreen`— Neither document outline nor thumbnail images visible.
- `/UseOC`— Neither document outline nor thumbnail images visible.
- `/UseAttachments`— Neither document outline nor thumbnail images visible.

The values of both the page layout and the page mode are expressed as names. The `/ViewerPreferences` are stored in a dictionary. Table 3.15 lists the most important entries involving the viewer application.

Key	Value	Description
`/NonFullScreenPageMode`	Name	The document's page mode when exiting from full screen mode; this entry only makes sense if the value of the `/PageMode` entry is `/FullScreen`. `/UseNone`: no panel is opened `/UseOutlines`: the bookmarks panel is opened `/UseThumbs`: the thumbnails panels is opened `/UseOC`: the optional content panel is opened
`/FitWindow`	Boolean	Changes the zoom factor to fit the size of the first displayed page when `true`.
`/CenterWindow`	Boolean	Positions the document's window in the center of the screen when `true`.
`/DisplayDocTitle`	Boolean	Positions the document's window in the center of the screen when `true`.

Key	Value	Description
/HideToolbar	Boolean	Hides the toolbars in the PDF viewer when true.
/HideMenubar	Boolean	Hides the menubar in the PDF viewer when true.
/HideWindowUI	Boolean	Hides user interface elements in the document's window (scrollbars, navigation controls) when true.

Table 3.16 lists the most important entries with respect to printing the document.

Key	Value	Description
/PrintScaling	Name	Allows you to avoid print scaling by the viewer by setting the value to /None; the default is /AppDefault.
/Duplex	Name	The paper handling option. Possible values are: /Simplex: print single sided /DuplexFlipShortEdge: duplex and flip on the short edge of the sheet. /DuplexFlipLongEdge: duplex and flip on the long edge of the sheet.
/PickTrayByPDFSize	Boolean	If true, the check box in the print dialog associated with input paper tray will be checked.
/PrintPageRange	Array	The page numbers to initialize the print dialog box. The array consists of an even number of integers of which each pair defines a subrange of pages with the first and the last page to be printed.
/NumCopies	Integer	Presets the value of the number of copies that need to be printed.

These viewer preferences preselect or preset a value in the dialog box. They can be used to set parameters, not to actually print the document.

FAQ: How can I print a document silently?

In old versions of Adobe Reader, it was possible to print a PDF without any user interaction. This was known as silent printing. This *feature* can also be seen as a security hazard. A PDF with silent printing activated would start your printer the moment the user opens the document, without asking the user for permission. This *problem* was fixed in the more recent versions of Adobe Reader. Silent printing is no longer possible. The end user always has to confirm that the document can be printed in the print dialog.

Other possible entries in the /ViewerPreferences dictionary are:

- /Direction– to define the predominant order for text (/L2R for left to right and /R2L for right to left).
- /ViewArea, /ViewClip, /PrintArea *and* /PrintClip– to define page boundaries. These entries are deprecated and should no longer be used.
- /Enforce— a new entry introduced in PDF 2.0 with an array of viewer preferences that shall be enforced in the sense that they can't be overridden in the viewer's user interface.

Whether or not setting these viewer preferences has any effect depends on the implementation of the PDF viewer. Not all PDF viewers respect the viewer preferences as defined in the PDF document.

3.5.7 Metadata

In section 2.1.4, we've found a reference to the info dictionary in the trailer. This info dictionry contains metadata such as the title of the document, its author, some keywords, the creation and modification date, and so on. However: this type of storing metadata inside a PDF file will be deprecated in PDF 2.0. Let's find out which type of metadata will remain available in the near future.

3.5.7.1 Version

As explained in section 2.1.1, you can find the PDF version in the document header. However, there are two situations that required an alternative place to store the PDF version.

1. When creating a PDF on the fly, the first bytes can already be sent to the output stream before the document has been completed. Now suppose that your application starts by writing `%PDF-1.4`, but you decide to introduce functionality that didn't exist in PDF 1.4 —for instance Optional Content— during the writing process. You can't change the first bytes anymore. They are sent to an output stream that could be out of reach —for instance a browser on a client machine. In this case, you'll change the version at the level of the catalog. This explains why iText always writes the Document Catalog Dictionary as one of the last objects in the file structure, followed only by the `/info` dictinary. You need to be able to change the keys of the root dictionary up until the very last step in the process.
2. When creating an incremental update, you add an extra body, cross-reference table and trailer. Suppose that you want to add an extra signature to a signed PDF. Suppose that the type of signature you're adding didn't exist in the version of the original PDF. You can't change the existing header in an incremental update; if you tried, you'd break the original signature. In this case, you'll change the version by defining it in the Catalog.

The value of the `/Version` entry in the Catalog is a name object. For instance: `/1.4`, `/1.7`, `/2.0`,...

3.5.7.2 Extensions

Third party vendors can —within certain limits— extend the PDF specification with their own features. When they use these extensions in a document, they'll add an `/Extensions` dictionary that contains a prefix that serves as identification for the vendor or developer, as well as a version number for the extensions that are used in the docment.

3.5.7.3 XMP streams

The `/Info` will be deprecated in favor of using a `/Metadata` entry in the Catalog starting with PDF 2.0 (ISO-32000-2). The value of this entry is a stream of which the dictionary has two additional entries: the `/Type` entry of which the value shall be `/Metadata`, and the `/Subtype` of which the value shall be `/XML`. The metadata is stored as an XML stream. This stream is usually uncompressed so that it can be detected and parsed by applications who aren't PDF aware.

There can be more than one Metadata stream inside a document. One can add a `/Metadata` entry to a page dictionary, or any other object that requires metadata.

The XML grammar that is used for the XML is described in a separate standard (ISO-16684-1:2012) known as the Extensible Metadata Platform (XMP). This standard includes different schemas such as Dublin Core. The XMP specification is outside the scope of this book.

3.5.7.4 The natural language specification

In the context of accessibility, it is recommended that you add a `/Lang` entry to the Catalog. The value of this entry is a string that represents a language identifier that specifies the natural language for all the text in the document. The language defined in the Catalog is valid for the complete document, except where overridden by language specification for marked content or structure elements.

3.5.8 Extra information stored in the Catalog

The Catalog can also be used to store specific information about the content, the producer that created the PDF, and the reader application that will consume the PDF.

The following entries provide more information about the content in some very specific use cases:

- *Threads* — The content of a PDF can consist of different items that are logically connected, but not physically sequential. For instance: an article in a news paper can consist of different blocks of text, distributed over different pages. For instance: a title with some text on the front page, and the rest of the article somewhere in the middle of the news paper. The `/Threads` entry in the Catalog, allows you to store an array of thread dictionaries defining the separate articles.
- *Legal* — JavaScript, optional content,... PDF offers plenty of functionality that can make the rendered appearance of a document vary. This functionality could potentially be used to construct a document that misleads the recipient of the document. These situations are relevant when considering the legal implications of a *signed* PDF document. With the `/Legal` entry, we'll add a dictionary that lists how many JavaScript actions can be found in the document, how many annotations, etc. In case of a legal challenge of the document, any questionable content can be reviewed in the context of the information in this dictionary.

These entries are used by specific software products that produce PDF documents:

- *Private data from the processor* — software that produces PDF as one of its output formats can use the `/PieceInfo` entry to store private PDF processor data. This extra data will be ignored by a PDF viewer.
- *Web Capture data* — if the PDF was the result of a Web Capture operation, the `/SpiderInfo` entry can be used to store the commands that were used to create the document.

These entries are meant to be inspected by software that consumes PDF documents:

- *Permissions* — a PDF can be signed to grant the user specific permissions, for instance to save a filled out form locally. The `/Perms` entry will define the usage rights granted for this document.

- *Requirements* — not all PDF consumers support the complete PDF specification. The `/Requirements` entry allows you to define an array of the minimum functionality a processor must support (optional content, digital signatures, XFA,...) as well as a penalty if these requirements aren't met.
- *Reader requirements* — The `/ReaderRequirements` entry is similar to the `/Requirements` entry, but defines specific reader requirements, for instance related to output intents.
- *Output intents* — The `/OutputIntents` entry consists of an array of output intent dictionaries specifying color characteristics of output devices on which the document might be rendered.

This concludes the overview of possible entries in the root dictionary aka catalog of a PDF document.

3.6 Summary

We've covered a lot of ground in this chapter. After examining the file structure in chapter 2, we've now learned how to obtain an instance of the objects discussed in chapter 1.

We've started exploring the pages of a document starting from the root dictionary and we've gotten used to the concept of using dictionaries to store destinations, outline items, action, and many other elements. While doing so, we've discovered that there's more to a page than meets the eye. We'll elaborate on some concepts such as annotations and optional content in later chapters.

The same goes for the other entries in the document catalog. We've only scratched the surface of what is available in the PDF reference.

In part 2, we'll dive into the content of a page. We'll talk about graphics and text, as well as about structure.

II Part 2: The Adobe Imaging Model

We studied the Carousel Object System and the structure of PDF files and documents in the previous part. While doing so, we briefly looked at a specific type of stream, more specifically a stream containing PDF syntax that draws lines, shapes, text and images to a page. In this part, we'll take a closer look at this syntax.

We'll start by looking at the different operators and operands that can be used to draw lines and shapes and to change properties such as the color, line widths, and so on. We usually refer to the *graphics state* in this context.

In the next chapter, we'll discuss the *text state*, which is a subset of the graphics state. We'll discover how to show text on a page, referring to a font program that knows how to draw each glyph.

Finally, we'll revisit some of the entries in the Catalog dictionary that we discussed only briefly, involving marked content, tagged PDF and optional content.

4 Graphics State

In section 3.4.1.1, we've already seen a glimpse of a content stream when we looked at the content stream of a page. Let's take a look at a similar content snippet:

```
BT
36 788 Td
/F1 12 Tf
(Hello World )Tj
ET
q
0 0 m
595 842 l
S
Q
```

This code snippet writes the words *"Hello Word"* to a pages and strokes a diagonal line.

4.1 Understanding the syntax

Before we start with a syntax overview, let's start by looking at the syntax notation, and find out how the imaging model was implemented in iText.

4.1.1 PDF Syntax Notation

The Portable Document Format evolved from the PostScript language and uses the same syntax notation known as postfix, aka reverse Polish notation. In reverse Polish notation, the operators follow their operands. Table 4.1 shows the different notations that can be used to note down the addition of the integers 10 and 6.

Table 4.1: Mathematical notations

Notation	Example	Description
prefix	+ 10 6	Polish notation
infix	10 + 6	The common arithmetic and logical formula notation
postfix	10 6 +	Reverse Polish notation

Interpreters of the postfix notation are often stack-based. Operands are pushed onto a stack, and when an operation is performed, its operands are popped from a stack and its result is pushed back on. This has the advantage of being easy to implement and very fast.

Let's take a look at the snippet `595 842 l` taken from the content stream in our example. We see the *path construction operator* `l` preceded by its operands, `595` and `842`, which are in this case values for an (x , y) coordinate. You'll find a corresponding method for this operator in iText. There's a `lineTo()` method in the `PdfContentByte` class that is responsible for writing two parameters, `x` and `y`, to a byte buffer, followed by the operator `l`. You can use this method if you want to create PDF at the lowest-level, using PDF syntax instead of the high-level objects described in the book "Create your PDFs with iText[1]."

4.1.2 Creating a PDF using low-level PDF syntax

Creating a PDF using iText always requires five basic steps:

1. Create a `Document` object
2. Get a `Pdfwriter`instance
3. Open the `Document`
4. Add content
5. Close the `Document`

Code sample 4.1 shows the five steps. In the fourth step, we use the `PdfContentByte` object to add some text and some graphics. The corresponding PDF syntax is added as a comment after each line.

Code sample 4.1: C0401_ImagingModel

```
    // step 1
    Document document = new Document();
    // step 2
    PdfWriter writer = PdfWriter.getInstance(document, new FileOutputStream(dest));
    // step 3
    document.open();
    // step 4
    PdfContentByte canvas = writer.getDirectContent();
    canvas.beginText();                              // BT
    canvas.moveText(36, 788);                        // 36 788 Td
    canvas.setFontAndSize(BaseFont.createFont(), 12); // /F1 12 Tf
    canvas.showText("Hello World ");                 // (Hello World )Tj
    canvas.endText();                                // ET
    canvas.saveState();      // q
    canvas.moveTo(0, 0);     // 0 0 m
    canvas.lineTo(595, 842); // 595 842 l
    canvas.stroke();         // S
    canvas.restoreState();   // Q
    // step 5
    document.close();
```

[1] https://leanpub.com/itext_pdfcreate

Figure 4.1 shows the result of this code sample. We see the text *Hello World* positioned more or less at the top of the page (36, 788). We also see a diagonal line going from the lower-left corner (0, 0) to the upper-right corner (595, 842).

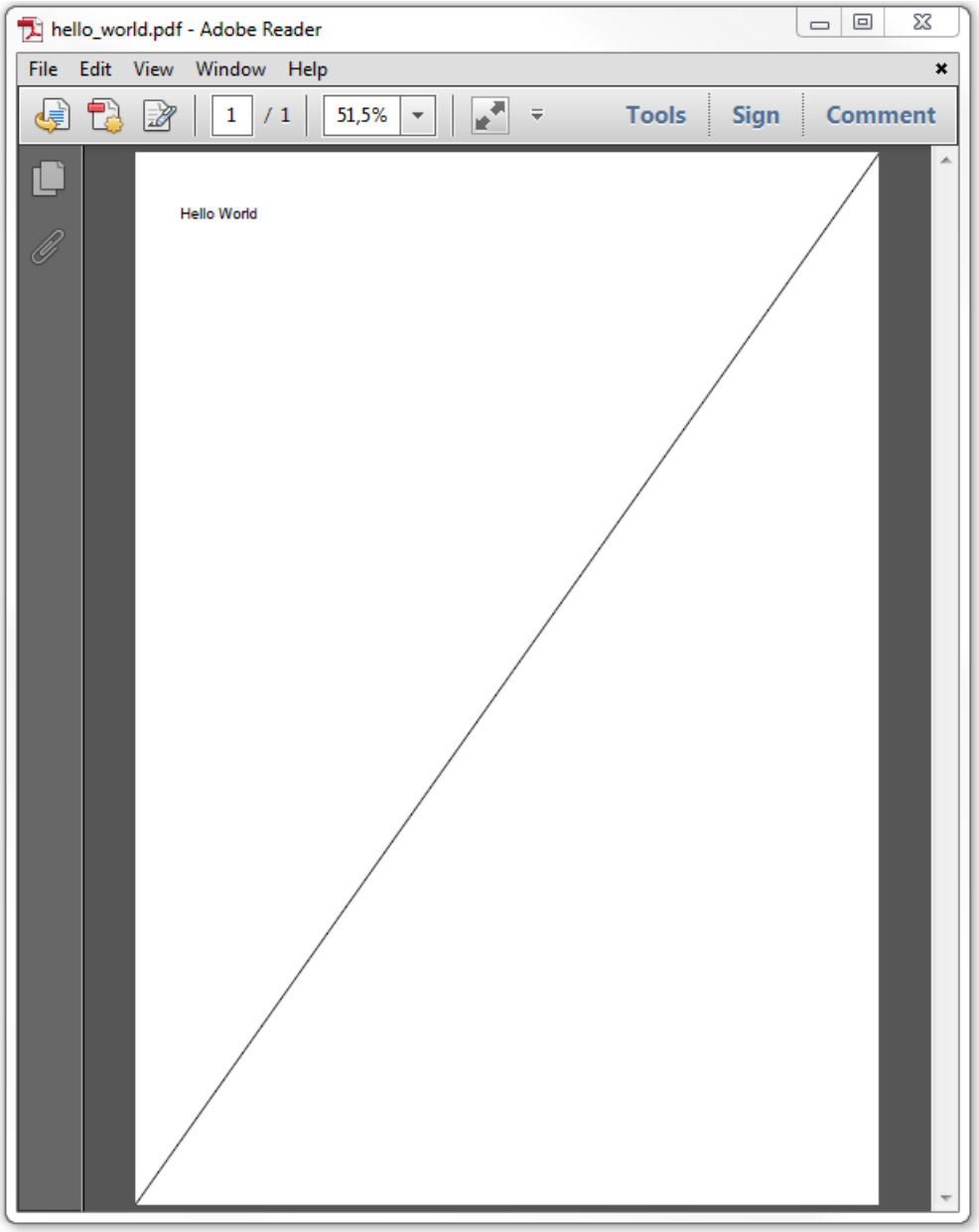


Figure 4.1: Hello World example

Now lets take a look at figure 4.2 where RUPS shows what's under the hood of the PDF.

RUPS inflates the compressed stream to allow us to see the PDF syntax. It removes or introduces spaces and newlines: all operands are separated by a single space character, each operator is shown on a separate line. It highlights the syntax in different colors: text state operators are shown in blue; pure graphics state operators are shown in orange; operands are shown in black. The *begin text* and *end text* operators, as well as the *save*

state and *end state* operators are also highlighted differently.

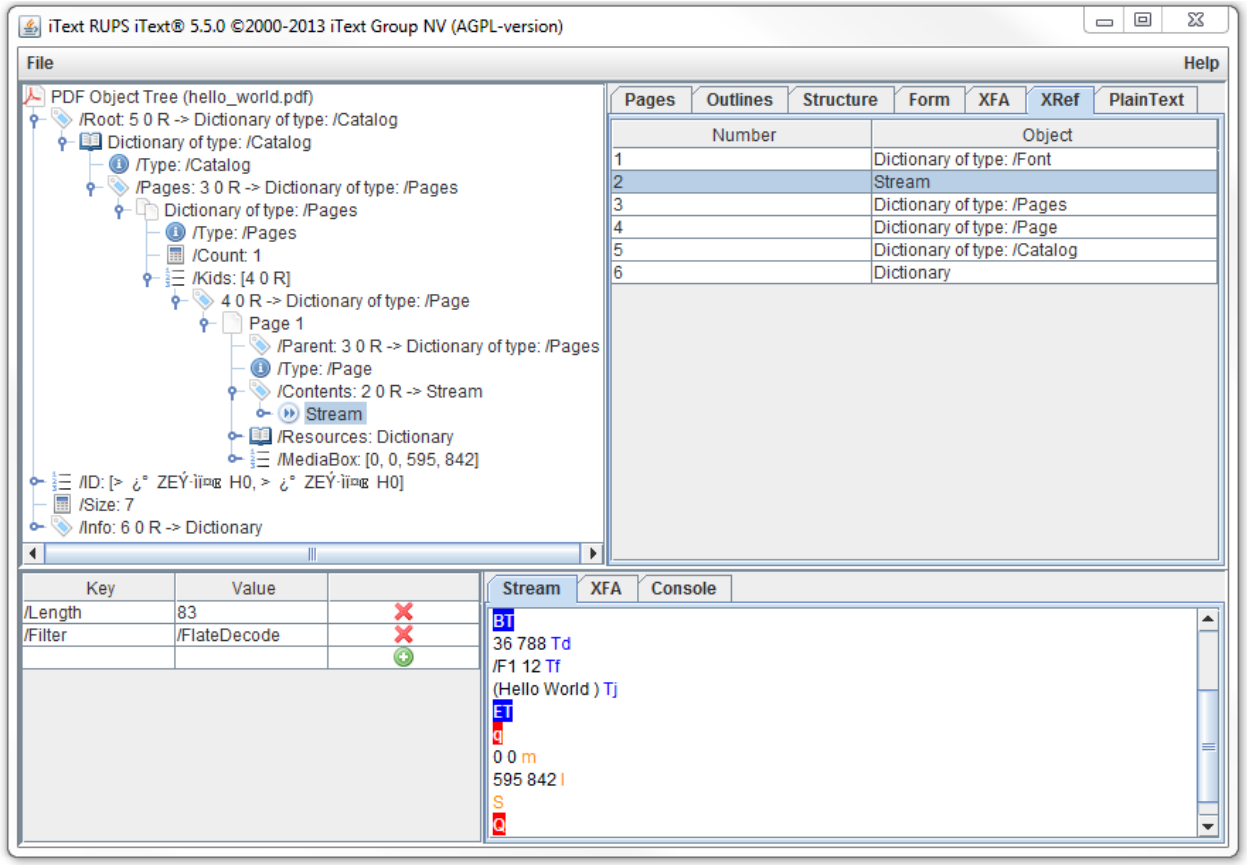


Figure 4.2: Syntax of the Hello World example

As you can see, RUPS makes it really easy for PDF-savvy people to read the syntax of different graphics objects.

4.1.3 Graphics objects

There are 5 types of graphics objects in PDF:

1. A *path object* is a shape created using path construction and painting operators. The path construction operators allow you to define lines, rectangles and curves. With the path painting operators you can fill or stroke the paths. You can also use a path to clip content.
2. A *text object* consists of a sequence of operators enclosed between the begin text and end text operator. A text object always refers to a font program that knows how to draw glyphs. Just like paths, these glyphs can be filled, stroked or used to clip content.
3. An *external object* (XObject) is an object defined outside the content stream and referenced by its name. The most common types of XObjects are *form XObjects* refering to another content stream that is to be considered as a single graphics object, and *image XObjects* refering to a raster image such as a JPEG, a CCITT image, and so on.

4. An *inline image object* uses special syntax to include raster image data within the content stream. This is only allowed for images with a size up to 4096 bytes.
5. A *shading object* describes a geometric shape whose color at a specific position is defined by a function, for instance a function defining a gradient transitioning from one color to another. A shading can also be used as a color when painting other graphics objects. It's not considered to be a separate graphics object in that case.

These objects are created using graphics state operators.

4.2 Graphics State Operators

There are different categories of graphics state operators: general graphics state operators, special graphics state operators, color operators, shading operators, path construction operators, path painting operators, clipping path operators, XObjects operators, inline images operators, text objects operators, text state operators, text positioning operators, text showing operators, Type 3 fonts operators, marked content operators and compatibility operators.

All of these operators are supported in iText, except for the compatibility operators. The `BX` and `EX` operators are used to begin and end a sequence of operators that may not be recognized by a PDF processor. Such a PDF processor will ignore unrecognized operators without reporting an error.

Lets start with the operators that allow us to draw a path object.

4.2.1 Constructing path objects

A path object always starts with one of the following operators: `m` or `re`. It ends either with a path painting or a path clipping operator. All the available path construction operators are shown in figure 4.3.

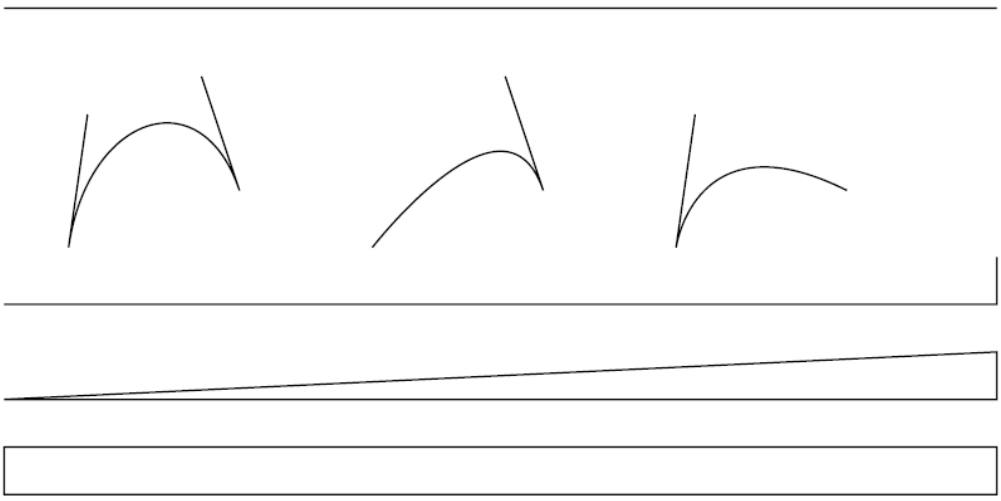


Figure 4.3: Path construction operators

Let's take a look at the overview of all the available path construction operators and find out if we can recognize them in figure 4.3.

In the first column, we have the PDF operator; in the second column you'll find the corresponding iText method; the parameters for those methods (the operands needed by the operator) are listed in the third column; the fourth column gives us a description.

Table 4.2: PDF path construction operators and operands

PDF	iText	Parameters	Description
`m`	`moveTo`	`(x, y)`	Moves the current point to coordinates `(x, y)`, omitting any connecting line segment. This begins a new (sub)path.
`l`	`lineTo`	`(x, y)`	Moves the current point to coordinates `(x, y)`, appending a line segment from the previous to the new current point.
`c`	`curveTo`	`(x1, y1, x2, y2, x3, y3)`	Moves the current point to coordinates `(x3, y3)`, appending a cubic Bézier curve from the previous to the new current point, using `(x1, y1)` and `(x2, y2)` as Bézier control points
`v`	`curveTo`	`(x2, y2, x3, y3)`	Moves the current point to coordinates `(x3, y3)`, appending a cubic Bézier curve from the previous to the new current point, using the previous current point and `(x2, y2)` as Bézier control points
`y`	`curveTo`	`(x1, y1, x3, y3)`	Moves the current point to coordinates `(x3, y3)`, appending a cubic Bézier curve from the previous to the new current point, using `(x1, y1)` and `(x3, y3)` as Bézier control points
`h`	`closePath`	`()`	Closes the current subpath by appending a straight line segment from the current point to the starting point of the subpath.
`re`	`rectangle`	`(x, y, w, h)`	Starts a new path with a rectangle or appends this rectangle to the current path as a complete subpath. The `x` and `y` parameter define the coordinate of the lower-left corner; `w` and `h` define the width and the height of the rectangle.

Code sample 4.2 shows the code that was used to create the PDF in figure 4.3.

Code sample 4.2: C0302_PathConstruction

```
    PdfContentByte canvas = writer.getDirectContent();
    // a line
    canvas.moveTo(36, 806);
    canvas.lineTo(559, 806);
    // lines and curves
    canvas.moveTo(70, 680);
    canvas.lineTo(80, 750);
    canvas.moveTo(140, 770);
    canvas.lineTo(160, 710);
    canvas.moveTo(70, 680);
    canvas.curveTo(80, 750, 140, 770, 160, 710);
    canvas.moveTo(300, 770);
    canvas.lineTo(320, 710);
    canvas.moveTo(230, 680);
    canvas.curveTo(300, 770, 320, 710);
    canvas.moveTo(390, 680);
    canvas.lineTo(400, 750);
    canvas.moveTo(390, 680);
    canvas.curveTo(400, 750, 480, 710);
    // two sides of a triangle
    canvas.moveTo(36, 650);
    canvas.lineTo(559, 650);
    canvas.lineTo(559, 675);
    // three sides of a triangle
    canvas.moveTo(36, 600);
    canvas.lineTo(559, 600);
    canvas.lineTo(559, 625);
    canvas.closePath();
    // a rectangle
    canvas.rectangle(36, 550, 523, 25);
    // nothing is drawn unless we stroke:
    canvas.stroke();
```

We start with a `moveTo()` and a `lineTo()` operation. This draws the first line.

Then we draw some more lines followed by the three flavors of the `curveTo()` method. These `curveTo()` methods create *Bézier curves*.

Bézier curves are parametric curves developed in 1959 by Paul de Casteljau (using *de Casteljau's algorithm*). They were widely publicized in 1962 by Paul Bézier, who used them to design automobile bodies. Nowadays they're important in computer graphics.

Cubic Bézier curves are defined by four points: the two *endpoints* —the current point and point `(x3, y3)`— and two *control points* —`(x1, y1)` and `(x2, y2)`. The curve starts at the first endpoint going onward to the

first control point. In general, the curve doesn't pass through the control points. They're only there to provide directional information. The distance between an endpoint ad its corresponding control point determines how long the curve moves toward the control point before turning toward the other endpoint. In figure 4.3, we've added lines that connect the endpoints with their corresponding control point. In the second curve, the endpoint to the left coincides with the first control point (the `v` operator was used instead of `c`). In the third curve, the endpoint to the right coincides with the second control point (the `y` operator was used).

Right under the curves, we see a subpath consisting of two lines. It is followed by another subpath that was constructed by a single `moveTo()` and two `lineTo()` operators, but instead of two lines, we now see a triangle. That's because we've used the `closePath()` operator. This operator adds a linear segment to the subpath that connects the current endpoint with the original startpoint of the subpath that was started with a `moveTo()` operation. Finally, we've also used the `rectangle()` method to draw a rectangle.

FAQ: I've constructed a path, but I can't see any line or shape in my document

The operators we've discussed so far can be used to *construct* a path. This doesn't mean the path is actually drawn. To draw the path, you need a path painting operator. In the example, we used the `stroke()` method to stroke the paths.

The PDF specification doesn't have any operator that allows you to draw a circle or an ellipse. Instead, you're supposed to combine the path construction operators listed in table 4.2. For instance: a circle consists of one `moveTo()` and four `curveTo()` operations. This isn't trivial. Fortunately, iText provides a handful of convenience methods, as listed in table 4.3.

Table 4.3: Convenience methods for specific shapes

iText method	Parameters	Description
`ellipse()`	`(x1, y1, x2, y2)`	Constructs the path of an ellipse inscribed within the rectangle `[x1 y1 x2 y2]`.
`arc()`	`(x1, y1, x2, y2, a, e)`	Constructs a path of a partial ellipse inscribed within the rectangle `[x1 y1 x2 y2]`; starting at `a` degrees (the start angle) and covering `e` degrees (the extent). Angles start with `0` to the right and increase counterclockwise.
`circle()`	`(x, y, r)`	Constructs the path of a circle with center `(x, y)` and radius `r`.
`roundRectangle()`	`(x, y, w, h, r)`	Constructs the path of a rounded rectangle: `(x, y)` is the coordinate of the lower-left corner; `w` and `h` define the width and the height. The radius used for the rounded corners is `r`.

Now that we know how to construct paths, let's find out how to paint them.

4.2.2 Painting and Clipping Path Objects

We've already used one painting operator in the previous examples: the `stroke()` operator `S`. To explain all the possible painting operators, we'll work with a series of pathes that represent a set of triangles as shown in figure 4.4.

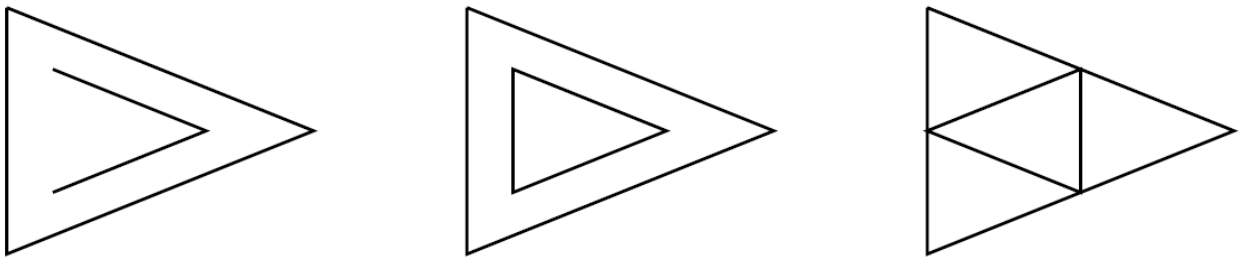


Figure 4.4: stroke() and closePathStroke()

Code sample 4.3 shows how we constructed the paths rendered in figure 4.4.

Code sample 4.3: C0403_PathPainting

```
1     protected void triangles1(PdfContentByte canvas) {
2         canvas.moveTo(50, 760);
3         canvas.lineTo(150, 720);
4         canvas.lineTo(50, 680);
5         canvas.lineTo(50, 760);
6         canvas.moveTo(65, 740);
7         canvas.lineTo(115, 720);
8         canvas.lineTo(65, 700);
9     }
10    protected void triangles2(PdfContentByte canvas) {
11        canvas.moveTo(200, 760);
12        canvas.lineTo(300, 720);
13        canvas.lineTo(200, 680);
14        canvas.lineTo(200, 760);
15        canvas.moveTo(215, 740);
16        canvas.lineTo(265, 720);
17        canvas.lineTo(215, 700);
18    }
19    protected void triangles3(PdfContentByte canvas) {
20        canvas.moveTo(350, 760);
21        canvas.lineTo(450, 720);
22        canvas.lineTo(350, 680);
23        canvas.lineTo(350, 760);
24        canvas.moveTo(400, 740);
25        canvas.lineTo(350, 720);
26        canvas.lineTo(400, 700);
27    }
```

In the `triangles1()` and `triangles2()` method, we draw one large triangle using three `lineTo()` methods, one for each side of the triangle. We start with the upper-left corner, draw a line that goes down to the right, followed by a line that returns down to the left. We close the path by connecting the lower-left corner with the upper-left corner. Inside this large triangle, we draw two sides of a smaller rectangle. Again we start with the upper-left corner, we draw a line to the right, followed by a line that returns to the left.

The `triangles3()` method is slightly different. The outer triangle is drawn in exactly the same way as before, but when we draw the inner triangle, we start to the right, we add a line that moves down to the right, followed by a line that moves down to the left. In `triangles1()` and `triangles2()` the two triangles are drawn using the clockwise orientation. In `triangles3()` one triangle is drawn clockwise, the other one counterclockwise.

The orientation of the paths doesn't matter when we merely stroke the paths. Code sample 4.4 shows how we've painted the paths shown in figure 4.4.

Code sample 4.4: C0403_PathPainting

```
1     PdfContentByte canvas = writer.getDirectContent();
2     triangles1(canvas);
3     canvas.stroke();
4     triangles2(canvas);
5     canvas.closePathStroke();
6     triangles3(canvas);
7     canvas.closePathStroke();
```

This code snippet explains why the second and third triangle have three sides in figure 4.4 in spite of the fact that we only constructed two lines. The `stroke()` method will only stroke two lines; the `closePathStroke()` method will close the path first, then stroke it.

Figure 4.5 shows what happens if we fill the path instead of stroking it.

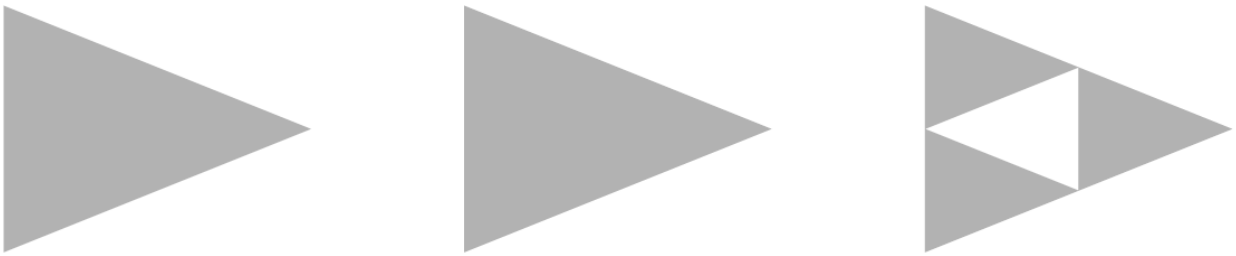


Figure 4.5: fill()

In the first two sets of triangles, both the other and the inner triangle are filled. This isn't the case in the third set: the inner triangle made a hole in the outer triangle. Let's take a look at code sample 4.5 and then discover why that hole is there.

Code sample 4.5: C0403_PathPainting

```
    triangles1(canvas);
    canvas.fill();
    triangles2(canvas);
    canvas.fill();
    triangles3(canvas);
    canvas.fill();
```

We have filled the different sets of triangles using the `fill()` method. This method uses the *nonzero winding number rule* to determine whether or not a given point is inside a path.

With the nonzero winding number rule, you need to draw a line from that point in any direction, and examine every intersection of the path with this line. Start with a count of zero; add one each time a subpath crosses the line from left to right; subtract one each time a subpath crosses from right to left. Continue doing this until there are no more path segments to cross. If the final result is zero, the point is outside the path; otherwise it's inside.

This explains why the orientation we used to draw the segments of the triangles matters. The winding number count for the points inside the inner triangle is 2: when drawing a line from inside the inner triangle to outside the outer triangle, we encounter two segments drawn from left to right. In the third set, the count is 0 because we encounter a segment drawn from right to left (subtract one), followed by a segment drawn from left to right (add one).

Figure 4.6 shows two more methods that use the nonzero winding number rule:

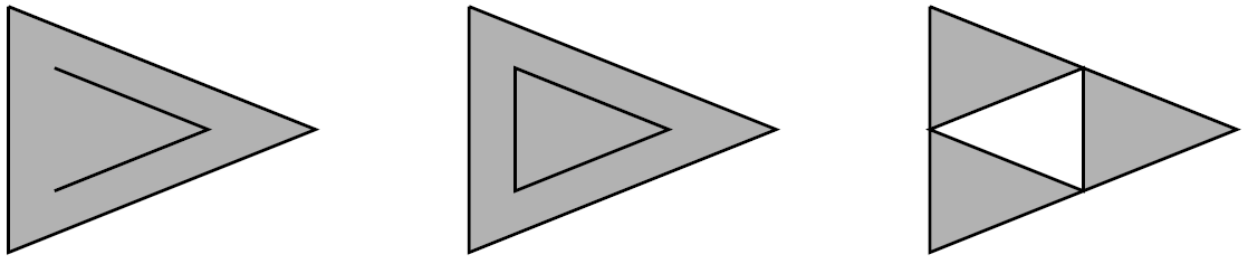


Figure 4.6: fillStroke() and closePathFillStroke()

Code sample 4.6 shows how these triangles were drawn.

Code sample 4.6: C0403_PathPainting

```
    triangles1(canvas);
    canvas.fillStroke();
    triangles2(canvas);
    canvas.closePathFillStroke();
    triangles3(canvas);
    canvas.closePathFillStroke();;
```

The `fillStroke()` method is a combination of the `fill()` and the `stroke()` method. The `closePathFillStroke()` method combines `closePath()`, `fill()` and `stroke()`.

Figure 4.7 shows a different way to fill the paths. In this case there's also a hole in the first two sets of triangles.

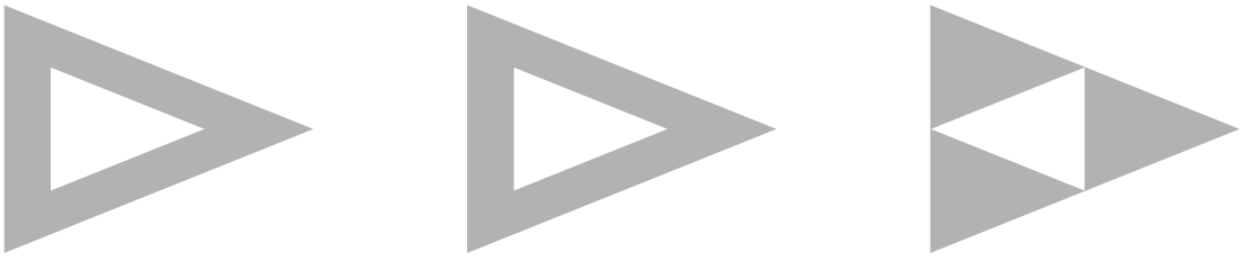


Figure 4.7: eoFill()

Please compare code sample 4.7 with code sample 4.5.

Code sample 4.7: C0403_PathPainting

```
    triangles1(canvas);
    canvas.eoFill();
    triangles2(canvas);
    canvas.eoFill();
    triangles3(canvas);
    canvas.eoFill();
```

The `eoFill()` method uses the *even-odd rule* to determine whether or not a given point is inside a path.

With the even-odd rule, you draw a line from the point that's being examined to infinity. Now count the number of path segments that are crossed, regardless of their orientation. If this number is odd, the point is inside; if even, the point is outside.

In this case, the orientation we used to draw the triangles doesn't matter.

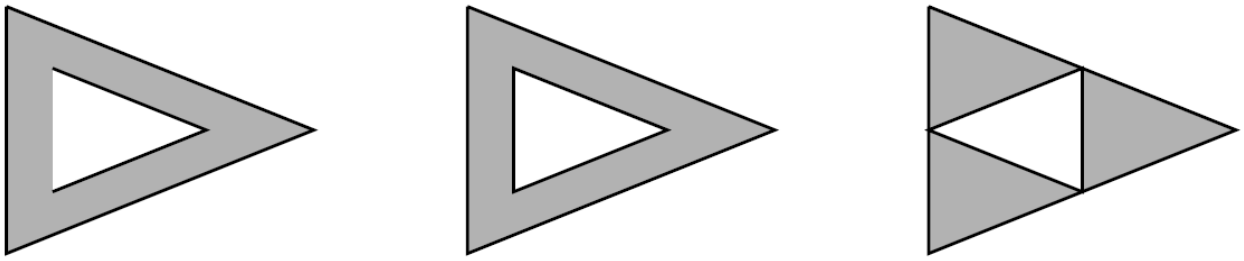


Figure 4.8: eoFillStroke() and closePathEoFillStroke()

Figure 4.8 shows two more methods using the even-odd rule. These methods are used in code sample 4.8.

Code sample 4.8: C0403_PathPainting

```
    triangles1(canvas);
    canvas.eoFillStroke();
    triangles2(canvas);
    canvas.closePathEoFillStroke();
    triangles3(canvas);
    canvas.closePathEoFillStroke();
```

The nonzero winding number rule and the even-odd rule can also be used for clipping. Figure 4.9 looks identical to figure 4.5.

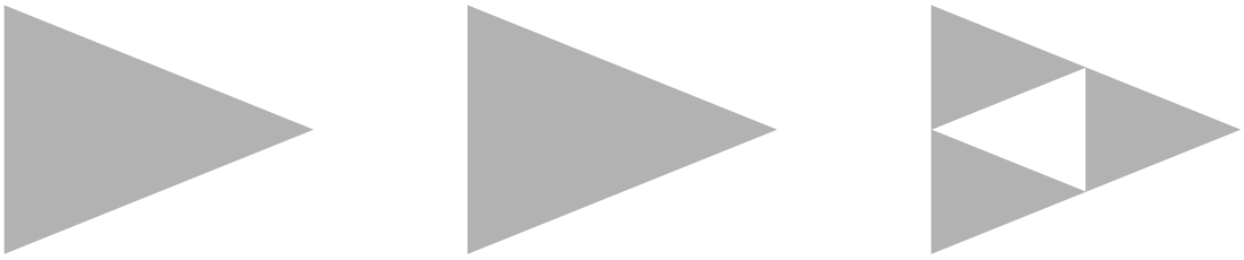


Figure 4.9: clip()

In spite of the resemblance, the code is completely different. See code sample 4.9.

Code sample 4.9: C0403_PathPainting

```
    canvas.saveState();
    triangles1(canvas);
    canvas.clip();
    canvas.newPath();
    canvas.rectangle(45, 675, 120, 100);
    canvas.fill();
    canvas.restoreState();
    canvas.saveState();
    triangles2(canvas);
    canvas.clip();
```

```
    canvas.newPath();
    canvas.rectangle(195, 675, 120, 100);
    canvas.fill();
    canvas.restoreState();
    canvas.saveState();
    triangles3(canvas);
    canvas.clip();
    canvas.newPath();
    canvas.rectangle(345, 675, 120, 100);
    canvas.fill();
    canvas.restoreState();
```

In this snippet, we draw the same paths, but we use them as clipping paths by invoking the `clip()` method. It's not our intention to draw the paths we've constructed, hence we start a new path with the `newPath()` method. Then we draw a rectangle and we fill that rectangle. The result a rectangle that is clipped using the path of the triangles. As we used the `clip()` method, the nonzero winding number rule is used.

Figure 4.10 shows what happens when we use the even-odd rule for the clipping path.

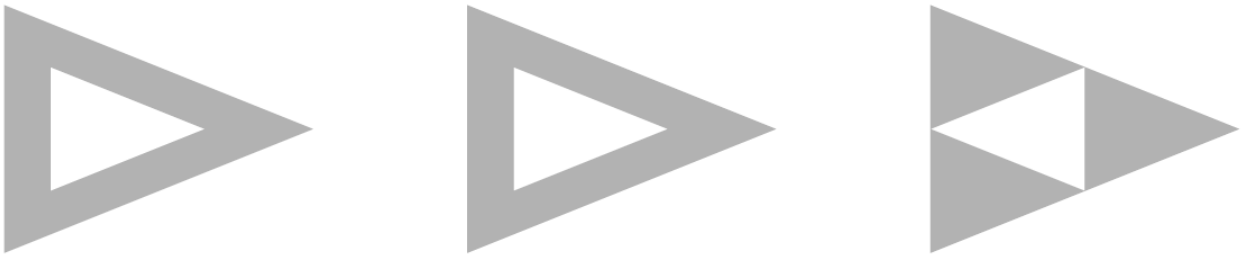


Figure 4.10: eoClip()

Code sample 4.10 shows how it's done.

Code sample 4.10: C0403_PathPainting

```
    canvas.saveState();
    triangles1(canvas);
    canvas.eoClip();
    canvas.newPath();
    canvas.rectangle(45, 675, 120, 100);
    canvas.fill();
    canvas.restoreState();
    canvas.saveState();
    triangles2(canvas);
    canvas.eoClip();
    canvas.newPath();
    canvas.rectangle(195, 675, 120, 100);
    canvas.fill();
    canvas.restoreState();
```

```
15      canvas.saveState();
16      triangles3(canvas);
17      canvas.eoClip();
18      canvas.newPath();
19      canvas.rectangle(345, 675, 120, 100);
20      canvas.fill();
21      canvas.restoreState();
```

In these code snippets, we introduced some methods we haven't discussed yet. We've also silently changed the fill color: the default color is black, not gray. Before we continue with more types of graphics states operators, let's look at an overview all the path painting operators in table 4.4.

Table 4.4: PDF path painting and clipping path operators

PDF	iText	Description
S	`stroke()`	Strokes the path: lines only; the shape isn't filled.
s	`closePathStroke()`	Closes and strokes the path. This is the same as doing `closePath()` and `stroke()`.
f	`fill()`	Fills the path using the nonzero winding number rule. Open subpaths are closed implicitly.
F	—	Deprecated! Equivalent to `f`, and included for compatibility. ISO-32000-1 says that PDF writer applications should use `f` instead.
f*	`eoFill()`	Fills the path using the even-odd rule.
B	`fillStroke()`	Fills the path using the nonzero winding number rule, and then strokes the path. This is equivalent to `fill()` followed by `stroke()`.
B*	`eoFillStroke()`	Fills the path using the even-odd rule, and then strokes the path. This is equivalent to `eoFill()` followed by `stroke()`.
b	`closePathFillStroke()`	Closes, fills, and strokes the path, as is done with `closePath()` followed by `fillStroke()`.
b*	`closePathEoFillStroke()`	Closes, fills, and strokes the path, as is done with `closePath()` followed by `eoFillStroke()`.
n	`newPath()`	Ends the path object without filling or stroking it. Used primarily after defining a clipping path.
W	`clip()`	Modifies the current clipping path by intersecting it with the current path, using the nonzero winding number rule.
W*	`eoClip()`	Modifies the current clipping path by intersecting it with the current path, using the even-odd rule.

These operators paint a path, but how the path is painted depends on the current graphics state.

4.2.3 Graphics state operators

When we talk about graphics state, we refer to an internal data structure that holds current graphics control parameters. These parameters will have an impact on the graphics objects we draw. For instance: when we have constructed a path using a `moveTo()` and `lineTo()` operator, and we stroke that path using the `stroke()` operator, the line will be drawn using the line width and stroke color as currently defined in the graphics state. If we didn't explicitly set a line width or color, default values will be used. For example: the default line width is 1 and the default stroke color is black.

5 Text State

6 Marked Content

III Part 3: Annotations and form fields

7 Annotations

8 Interactive forms