# BAB II

# LANDASAN TEORI

Keberadaan format file PDF tidak lepas dari sejarah yang berdiri di belakangnya. Berawal dari sebuah impian untuk mewujudkan sebuah kantor yang bebas dari penggunaan kertas atau *paperless*, adalah John Warnock yang juga merupakan salah satu dari pendiri Adobe yang mempunyai ide untuk membuat format file khusus agar dokumen dapat disebarluaskan dalam perusahaan dan dapat ditampilkan oleh semua komputer dengan berbagai jenis sistem operasi yang digunakan.

## 2.1 Definisi PDF

Beberapa definisi dari PDF :

1. *PDF is a file format that retains a document's true formatting across platforms and is useful for documents with complex formatting such as newsletters or financial statements. PDF files can be viewed and printed exactly as the author intended using a free PDF file reader* [EDU06].
2. ***Portable Document Format (PDF)** is a file format proprietary to Adobe Systems for representing two-dimensional documents in a device independent and resolution independent fixed-layout document format. Each PDF file encapsulates a complete description of a 2D document (and, with the advent of Acrobat 3D, embedded 3D documents) that includes the text, fonts, images, and 2D vector graphics that compose the document. PDF files do not encode information that is specific to the application software, hardware, or operating system used to create or view the document. This feature ensures that a valid PDF will render exactly the same regardless of its origin or destination (but depending on font availability)* [WIK06].
3. *Suatu format file yang didesain agar dokumen dapat dilihat dan dicetak sekaligus formasi dokumennya (bentuk huruf, gambar, layout, dll.) ditampilkan sama sebagaimana sistem operasi yang digunakan menampilkannya, sehingga dokumen PDF memiliki tampilan yang sama pada Windows, Macintosh, Linux, OS/2, dll. Format PDF pada dasarnya merupakan penggunaan luas dari bahasa Postcript document-description. PDF and Postscript dikembangkan oleh Adobe Corporation* [TEL06].

Dari beberapa pengertian di atas, bisa diambil suatu pengertian secara umum yaitu PDF merupakan suatu format file khusus dimana dokumen dapat dibaca dan dicetak sesuai format aslinya (bentuk huruf, gambar, layout, dll) dengan berbagai macam sistem operasi yang digunakan seperti Windows, Macintosh, Linux dan lain sebagainya. Aplikasi pembaca format file PDF yang tersedia secara gratis dapat didownload/diunduh dari situs Adobe sendiri, namun untuk dapat menggunakan fitur tambahan dikenakan biaya khusus untuk memilikinya.

## 2.2 Struktur PDF

Sebuah file PDF memiliki struktur file yang unik di dalamnya. Umumnya sebuah dokumen PDF versi awal terdiri dari 4 bagian yang terdiri dari : kepala dokumen, badan dokumen, *cross reference* dan *trailer*. Bagian – bagian tersebut memiliki pengertian dan fungsi tertentu, yaitu :

1. Kepala dokumen : bagian paling awal dari susunan struktur dokumen PDF ini bertugas untuk mengidentifikasi versi dari dokumen PDF serta memaparkan versi tersebut ke dalam struktur dokumen PDF yang sedang dianalisis. Bagian ini dapat dilihat dengan kondisi penulisan seperti berikut : %PDF-1.2. Angka 1.2 dapat berubah tergantung dari versi dokumen PDF yang digunakan.
2. Badan dokumen : berisi berbagai macam objek yang terdapat pada halaman PDF sehingga membentuk kumpulan objek yang terorganisasi membentuk struktur pohon yang mendeskripsikan sebuah struktur halaman dan elemen-elemen halaman, seperti bentuk teks, grafik, bentuk

tabel dan sebagainya. Sebuah objek dapat digambarkan sebagai daun yang ada pada struktur pohon PDF. Setiap objek memiliki tiga komponen penting; pertama memiliki susunan angka, posisi yang tetap pada dokumen file PDF, serta memiliki *content* atau isi.

3. *Cross reference* : *cross reference* mengijinkan aplikasi pembaca file PDF seperti Acrobat Reader untuk mengakses sebuah objek dengan cepat. Ciri dari bagian *cross reference* yaitu dimulai dengan kata kunci *xref* diikuti dengan sekumpulan angka objek Contoh dari bagian *cross reference* :

```
Xref
0 20
0000000000   65535   F
0000000009   00000   N
0000000116   00000   N
.....
```

Sekumpulan daftar objek tersebut dipakai sebagai referensi keberadaan objek-objek yang ditandai dengan huruf n yang berarti *unused* atau tidak digunakan dan huruf f yang berarti *free* atau bebas.

4. *Trailer* : merupakan masukan singkat menuju bagian yang paling penting dari dokumen, berisi sebuah pointer untuk memulai bagian *cross reference* dan angka objek [PRE07]. Contoh dari *trailer* :

```
Trailer
<<
/Size 20
/Root 19 0 R
/Info 18 0 R
>>
Startxref
2354
%%EOF
```

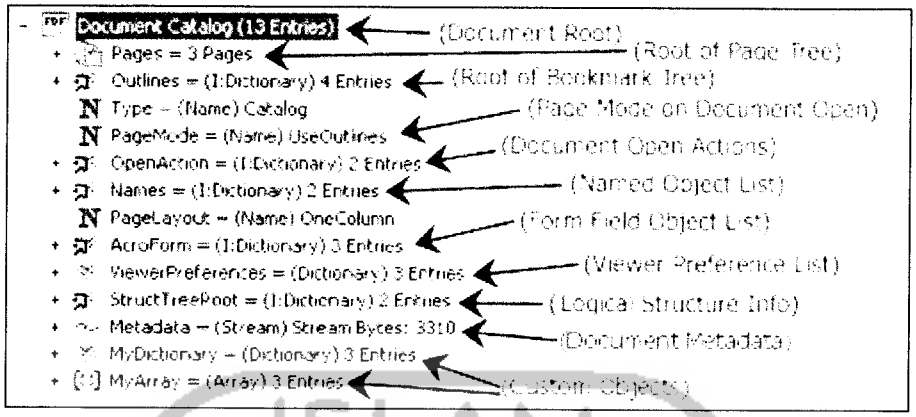
Format file PDF adalah sebuah teks dengan campuran data biner di dalamnya. Pada level awal atau level yang paling rendah terdapat 8 tipe objek yang terbagi menjadi 2 bagian yaitu *Scalar* dan *Container*. Yang dikategorikan pada tipe *Scalar* adalah *integer*, *boolean*, *real number*, *name* dan *string*. Sedangkan yang termasuk dalam tipe *Container* adalah *dictionary*, *array* dan *stream*. Berikut pengertian dari beberapa tipe objek *Scalar* dan *Container* :

1. ***Integer***, yaitu tipe data yang terdiri dari bilangan bulat dalam rentang tertentu.
2. ***Boolean***, pada bagian ini tipe data dinyatakan '*true*' or '*false*'
3. ***Real number***, yaitu tipe data yang terdiri dari bilangan pecahan atau desimal dalam rentang tertentu.
4. ***Name***, yaitu tipe data dengan format "*/text*" dimana dimulai dengan karakter *slash* diikuti teks dibelakangnya tanpa spasi maupun tanda baca.
5. ***String***, yaitu tipe data berupa karakter atau *hexadecimal* karakter.
6. ***Dictionary***, format penulisan seperti ini <<...*other objects*...>>. Objek *Dictionary* selalu berdampingan dengan objek lain seperti *Name* objek.
7. ***Array***, format penulisan *[...other objects...]*. Sederetan daftar dari kumpulan objek yang dipisahkan oleh spasi jika memang diperlukan.
8. ***Stream***, format penulisan <<...*stream attribute objs*...>>*stream...binary data...endstream*. Sebenarnya objek ini merupakan suatu penggabungan dari objek *Dictionary* dengan objek *String*. *Dictionary* memuat informasi yang diperlukan untuk mengakses sebuah data pada objek *String*. *Stream*

merupakan objek tidak langsung, sehingga objek *Stream* selalu di awali dengan referensi objek sebelumnya [PLA07].

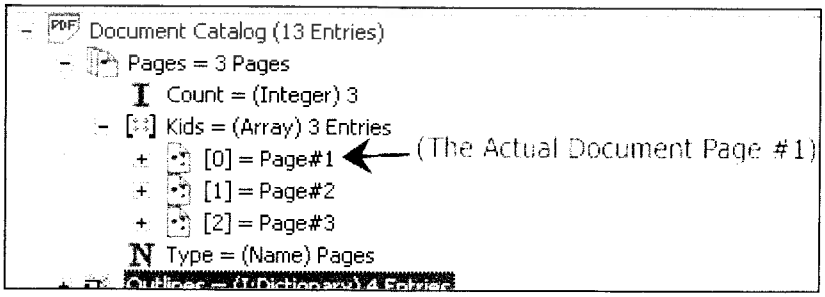
Sebuah dokumen PDF memiliki berbagai macam layer di dalamnya. Pada layer terendah dari sebuah file PDF memuat sejumlah data dokumen yang masih bersifat mentah. Selanjutnya layer COS (layer inti dari dokumen PDF) mengorganisir data tersebut menjadi sebuah struktur pohon objek sederhana. Pada layer selanjutnya yaitu *Portable Document*, kumpulan beberapa objek sederhana ini membentuk menjadi sebuah kesatuan agar dapat diimplementasikan pada struktur level menengah seperti huruf-huruf dan gambar. Objek yang telah terbentuk ini kemudian membangun layer yang lebih tinggi lagi menjadi beberapa keterangan dan halaman-halaman. Beberapa dari objek tersebut juga digunakan untuk menentukan struktur logis, seperti paragraf dan urutan artikel.

Dengan bantuan plug-in tambahan yaitu PDF CanOpener pada Acrobat, dapat dilihat perbedaan bagian-bagian dari pohon objek COS. Pada tampilan yang muncul struktur pohon yang utama dari setiap dokumen PDF dinamakan katalog dokumen. Katalog ini merupakan kamus dari berbagai macam objek yang ada. Apapun masukan di dalamnya, pada umumnya merupakan keseluruhan PDF atau keterangan umum dari level dokumen. Beberapa inti bagian di dalam sebuah katalog adalah *scalar*, *type*, *pageMode*, dan *PageLayout*



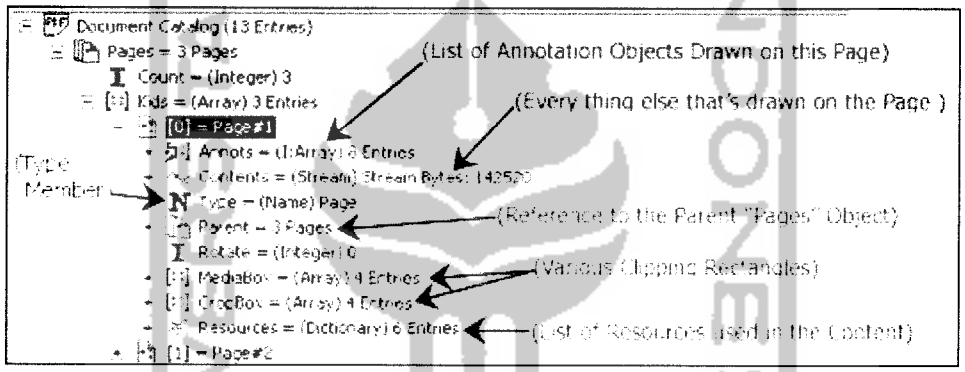

**Gambar 2.1** *Document Catalog*

Banyak objek yang terlihat atau ditemukan pada struktur PDF, salah satunya akar pada struktur pohon halaman yang merupakan sebuah objek kamus. Secara spesifik kamus objek tersebut merujuk pada kamus halaman. Pada objek ini terdapat 3 sub objek yaitu *Type*, *Count*, dan *Kids*. Sub objek *Type* biasanya terdapat pada beberapa tingkat objek yang lebih tinggi. Pada kasus ini tingkat objeknya adalah halaman. Pada sub objek *Count* mengindikasikan berapa jumlah total dari keseluruhan halaman yang terdapat pada struktur pohon halaman. Sementara array *Kids* adalah sebuah daftar dari *Pages* atau *Page* objek. *Pages* Objek adalah objek tingkat lanjut sementara *Page* merupakan objek awal dari struktur pohon halaman.




**Gambar 2.2** *Page Tree Root Dictionary*

Pada sub objek halaman berikutnya, dicontohkan pada objek halaman pertama terdapat objek-objek yang menerangkan *property* dari halaman tersebut. Objek *Annots* adalah daftar anotasi objek gambar yang terdapat pada halaman. Objek *Content* adalah objek yang memuat semua isi yang ada pada halaman. Sedangkan objek-objek seperti *MediaBox*, *CropBox* merupakan objek berbentuk bujur sangkar yang memiliki 4 titik tepat di mana koordinatnya menggunakan sisi-sisi sebelah kiri, kanan, atas dan bawah dari bujur sangkar tersebut [PLB07].




**Gambar 2.3** *Pages Dictionary*