

PDF File Format

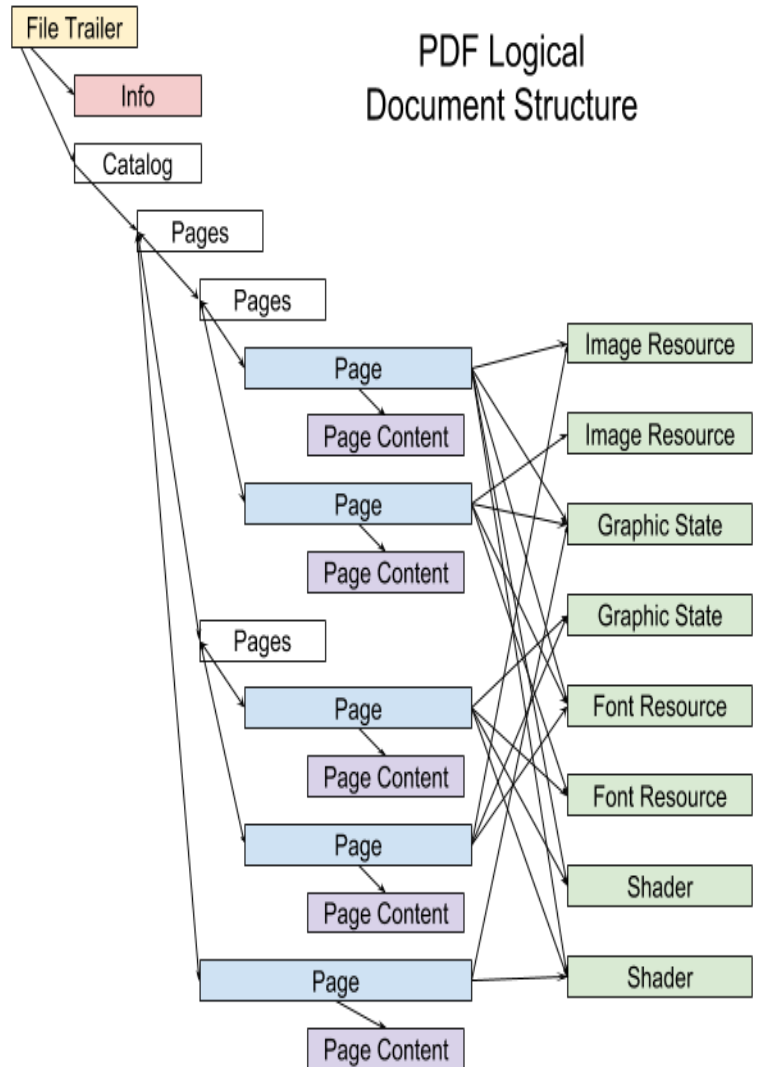
Kitap ve döküman okumaya başladıktan sonra okumak için açtığım PDF dosyalarının zararlı yazılım içerip içermediği konusunda şüpheler duymaya başladım. Bu şüphelerden kurtulabilmek adına birkaç örnek analiz yazısı okuyup videolarını izledikten sonra okumak istediğim PDF dökümanlarını kendim analiz etmeye karar verdim.

PDF dosya formatı 1993 yılında Adobe tarafından oluşturulmuştur. İşletim sisteminden bağımsız belgelerin sunulabilmesini sağlayan metin tabanlı bir yapı olarak oluşturuldu. Metinlerin yanı sıra görseller, videolar ve daha birçok türden içerik PDF dosyaları kullanılarak sunulabilmektedir.

Analiz edebilmek için öncelikle PDF dosyasının yapısını anlamamız gerekiyor. Bir PDF dosyası 4 ana bölümden oluşuyor. Bu bölümler;

- Header, dosyanın bir PDF dosyası olduğunu belirtir ve sürüm bilgisini bulunur.
- The Body, PDF dosyasında kullanıcıya gösterilen metin, görsel, video ve benzeri içeriklerin tutulduğu kısımdır. Bu kısımda bulunan içerikler aslında bir nesne içerisinde tutulmaktadır. Bu nesneler farklı türlerde olabilir.
 - o Names
 - o Numbers
 - o Boolean
 - o String
 - o Arrays
 - o Dictionary
 - o Streams
 - o Indirect Objects
- Cross-reference table, the body kısmında bulunan her nesnenin/içeriğin adres (20 bayt uzunluğunda – 3 parçadan oluşan) bilgisinin tutulduğu tablodur (Bir PDF tablosunda birden fazla Cross-reference tablosu bulunabilir). Bu tablo sayesinde bir nesnenin tutulduğu konumu belirlemek için bütün dosyanın okunmasına gerek kalmıyor ve büyük boyutlu dosyaları açarken de zaman kazandırıyor. Cross-reference tablosunda tutulan her nesne için iki satır bulunmaktadır.
 - o İlk satır iki sayı içermektedir. Bu sayılardan ilki nesnenin sayısal kimliğidir. İkinci sayı ise bu nesnenin altında bulunan nesnelerin sayısıdır.
 - o İkinci satırda ise ilk 10 baytı PDF dosyasının başlangıcından o nesnenin başlangıcına kadar olan uzaklığını tanımlamak için kullanılır. Ardından gelen 5 baytlık parçada nesnenin üretim numarası belirtilir. Son parçada ise “n” ve “f” harfleri kullanılarak nesnenin kullanımda olup olmadığı belirtiliyor.
- Trailer, PDF okuyucular dosyayı sondan okumaya başlar. Trailer kısmıysa dosyanın sonunda bulunur ve PDF okuyucunun nesneleri bulabilmesi için Cross-reference tablosu ve özel nesneler hakkında bilgiler tutar. Trailer kısmı birçok parçadan oluşuyor. Bu parçalar;
 - o Size (Integer), Cross-reference tablosundaki girişlerin sayısını tanımlar.
 - o Prev (Integer), dosyada birden fazla Cross-reference tablosu kullanıldığı durumda dosyanın başlangıcından önceki Cross-reference tablosuna olan uzaklığı tanımlar.
 - o Root, (Dictionary), tabloda kök nesne adresini tutan girişin bilgisini tutar. Bu nesne belgenin kataloğudur. Yani dosyanın nasıl sunulacağı hakkında bilgi ver belgenin içeriğini açıklayan diğer nesnelere referanslar verir.
 - o Info, dosya hakkında genel bilgilerin bulunduğu alandır.

- | Header | | |
|--|--|--|
| <pre>% PDF -1.4 . [hex] 25 504446 2D312E34 0D Ver EOL (CR)</pre> | | |
| Body | | |
| <pre>... <</Length 66/Filter/ FlateDecode/I 86/L 70/S 38>> <</Font<</TT2 10 0 R>> obj<</Subtype/TrueType/ FontDescriptor 12 0 R/LastChar 117/Widths[250 0 0 0 0 0 0 0 0 0 250 0 250 0 500 500 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 667 0 0 0 0 0 0 0 0 0 0 0 556 0 0 556 0 0 0 0 0 0 0 0 0 0 0 444 0 444 500 444 0 0 0 0 0 0 778 500 500 0 0 333 0 278 500] /BaseFontTimesNewRomanPSMT/ FirstChar 32/Encoding/ WinAnsiEncoding/Type/Font>> .endobj 11 0</pre> | | |
| Cross-reference Table | | |
| <pre>xref.. 0 6.. 1 Subsection x 6 0000000000 65535 f.. 3030303030303030303020363535333520 Object #0 Space EOL (CR+LF)</pre> | | |
| <pre>0000001831 00000 n.. 0000001865 00000 n.. 0000001889 00000 n.. 0000001940 00000 n.. 0000005609 00000 n..</pre> | | |
| f = free; n = in-use | | |
| Trailer | | |
| <pre>trailer.. <</Size 6>>.. # Total Xref startxref.. 116.. Offset Xref %%EOF.. End-of-File</pre> | | |



Kaynaklar

- <https://www.intezer.com/blog/incident-response/analyze-malicious-pdf-files/>
- <https://resources.infosecinstitute.com/topic/pdf-file-format-basic-structure/>
- https://labs.appligent.com/pdfblog/pdf_cross_reference_table/
- <https://www.oreilly.com/library/view/pdf-explained/9781449321581/ch04.html>
- <https://medium.com/aia-sg-techblog/basic-structure-of-portable-document-format-pdf-79db682579c9>