

1 PDF

Poznámka (Historie)

Portable Document Format (Adobe, 90 léta) měl nahradit PostScript (Ten byl příliš komplikovaný. Dokonce i s doplněním o konvence DSC.). Na rozdíl od PostScriptu PDF není míněno jako programovací popis.

1.0 ani Medvěd neviděl, 1.7 už bylo v ISO standardu (ale lepší je čist PDF Reference přímo od Adobe než ISO standard).

Postupem vznikly profily (sady pravidel) pro různé použití: PDF / X = pro tiskárnu (zakazuje JS apod.), PDF / A = pro archivaci (univerzálně srozumitelná téměř jistě i za 30 let).

1.1 Lexikální struktura

Definice 1.1 (Obsah PDF)

PDF je key sensitive.

Znaky jsou nadmnožina ASCII. Dělí se na mezery (mezera, TAB, FF, NUL, CR, LF, řádek končí LF / CR LF), oddělovače (() [] {} <> / %), ostatní (u těch není specifikováno, jaké je kódování).

% značí komentář do konce řádku (formálně mezera).

Klíčová slova (začínají písmenem a jdou až do mezery): null, true, false, f*.

Čísla jsou buď celá nebo floaty a píší se standardně.

Stringy se píší do kulatých závorek (ne uvozovek). Mohou být víceřádkové, mohou obsahovat i znak 0, mohou obsahovat vnořené kulaté závorky (správně uzavřené), escapes (\ \ \n \t \newline se ignoruje \ooo kód znaku v 8 soustavě). Nebo může být uzavřen do menšítek, pak je ale celý v hexadecimálním formátu.

Jména jsou znaky kromě mezer a oddělovačů předcházené /, hexadecimální znak se dá zadat #xx (ale nelze znak s kódem 0), konvence utf-8.

Pole jsou [prvky ...], tedy třeba 0 false (str) [1 2]. Slovník je << /jméno hodnota ... >> (slušnost je, aby se slova neopakovala). Konvencí je, že pokud slovník obsahuje /Type, tak je správně, např. /Page. Stream je tvaru << slovník + /Length bytes >> stream <eol> ...data stream slovník navíc může obsahovat /Filter /jméno nebo /Filter [/jméno ...] (např. /ASCIIHexDecode, /ASCII85Decode, /FlatDecode, /LZWDecode?).

Nepřímé objekty: definice = {číslo objektu} {číslo generace} obj ... endobj odkazování se = {číslo} {generace} R

Definice 1.2 (Struktura)

První řádek verze, druhý řádek nesmyslné znaky (aby se neinterpretovalo jako textový soubor), obojí zakomentované.

Mělo by (musí tam někde být) končit %%EOF (i když zatím občas bývá nějaký balast), před tím jsou (v pořadí odshora): xref (pro rychlé přečtení dat o objektech, 2 čísla a f- nebo n-?), trailer (metadata o soboru, slovník) a startxref (kde začíná xref, číslo).

Mezi tím jsou objekty.

Definice 1.3 (xref)

Začínají xref a končí (už tam nepatří) trailer. Má sekce (začínají 2 čísla, pořadí sekce? a počtem znaků). Navíc každý řádek musí mít 20 bytů (takže TeX doplňuje mezeru kvůli CR LF (TeX používá jen LF)).

Položka (tj. 1 řádek) může být n, pak obsahuje pozici (10 číslic), generaci (5 číslic) a n. Nebo může být f (později, souvisí s generacemi a aktualizacemi PDF).

Definice 1.4 (trailer)

Slovník obsahující /Size + počet objektů, /Prev + odkaz (pozice) na předchozí xref, /Rot nepřímý odkaz, /Info nepřímý odkaz, občas obsahuje /Encrypt, nepovinné, ale silně doporučené /ID + číslo verze při vytvoření + číslo aktuální verze.

Definice 1.5 (Updatování)

PDF je stavěné na update připsáním na konec. (Zajímavé třeba při podepisování.) Navíc xref má položku f (smazáno, nebo také free = volný obsahující odkaz (číselný) na další f objekt (resp. u 0 na 0) jako spojový seznam a číslo poslední generace (a f)). Navíc objekty mají generace, aby se daly recyklovat jejich čísla.

Definice 1.6

V novějších verzích se objekty ukládají do object streamů (aby se zmenšil text okolo). Ten vypadá: všechno generace 0, žádné další streamy, výjimky (nesmí obsahovat velikost jiného objektu atd.).

Číslo gen obj << /Type /objStm /N #objektů /First pozice 1. objektu [/Extends] [/Fi kde stream obsahuje N párů čísel (čísla objekth a pozice), cokoliv, N objektů bez obj a endobj.

Následně také obsahuje Xref Stream (místo xref a traileru): číslo gen obj << /Type /xref polož

Poznámka

Pro lepší práci s pdf se hodí program `qpdf`. Speciálně `qpdf --qpdf --object-stream=disable` (výsledek se ukládá do souboru a pak ho lze po upravení zase 'zkomprimovat').

Definice 1.7 (Stranky)

Root objekt (viz trailer) obsahuje seznam dětí = stránek.

Definice 1.8 (High-level struktura)

Z traileru získáme info, ale získáme i odkaz na catalog (root), který obsahuje všechno. Obsahuje odkazy (page tree, page labels, ...), viewer prefs (nějaké nastavení prohlížeče, např. zazoomování), jazyk (může být až na úrovni slov, když je dokument vícejazyčný, udává se např. kvůli ligaturám), version override (při updatu můžeme chtít použít novější verzi PDF, tak ta se uvádí zde).

page tree (počítá se klidně i s tisíci stránek a málem paměti zařízení) se skládá z jednotlivých objektů, které mají typ pages, parent ..., kids [...], count počet (aby se dalo rychle listovat). Listy pak mají typ page, parent ..., resources (slovník), contents (stream), mediabox [velikost stránky^a], cropbox [rozměry, na které ořezáváme veškeré kresby], bleedbox [kam může barva prosakovat] + trimbox [na co se bude ořezávat papír po opuštění tiskárny], rotate 90 (ne všechny prohlížeče respektují). (Boxy jdou definovat už v Pages, stejně tak další vlastnosti, jako přechody v prezentaci, ...).

resources a contents: v resources jsou odkazy na objekty, protože nemohou být contents (protože je to stream a nemusel by být čitelný všem programům). resources můžou být nepřímý objekt, tedy je mohou stránky sdílet.

Content stream (vypadá jak postscript) – částečně zásobníkový jazyk (ale po provedení operátoru musí být zásobník prázdný), obsahuje čísla a operátory (vždy seznam argumentů a operátor). Operátory jsou např. (m = move, l = line (nekreslí, jen ji vytvoří), S = stroke (vykreslí)).

^aSouřadnice jsou v bp (big pointech). Papír je popsán obdélníkem $[x_1y_1x_2y_2]$.

Definice 1.9 (Operátory)

Nastavení parametrů kreslení (grafický stav = gstate)^a: q / Q (save / restore – zásobník grafických stavů), *abcdef* cm (definuje transformační matici (násobí se s předchozí?)), *x w* (line width = šířka obdélníku kolem úsečky, při *x* = 0 se nastaví 1 pixel),

Konce a napojení úseček: (butt ending = konec kolmý na úsečku, round ending = konec oblý se středem v konci, square ending = konec čtverec se středem v konci, zároveň existují 3 typy napojení (jedné lomené čáry) round = vyplní se kruhovou úsečí, tj. v podstatě round, bevel = spojí se vrcholy obdélníků, miter = protáhne se do špičky (pokud je moc daleko (tzn. víc jak miter limit), udělá se bevel)): typ J (line cap: 0 = butt, 1 = round, 2 = square), typ j (line join: 0 = miter, 1 = round, 2 = bevel), limit M (miter limit).

Dále čárkování: [pole délek] fáze d (dash pattern, pole délek = délky čáry, mezery, čáry,

mezery, ..., fáze = kde v seznamu má začít, pole délek může mít i lichou délku, prostě se nekonečněkrát zopakuje za sebou, čárkový pattern pokračuje i za zlomem, pokud je to jedna čára, line cap a line join se aplikují na každou čárku zvlášť).

Další věci se dají reprezentovat ExtGState objektem, na který se pak odkazujeme operátorem /jméno gs.

Konstrukce cest: xy m (move to, začíná úsek), xy l (line to), $x_1y_1x_2y_2x_3y_3$ c (curve to = bézierova křivka 3. řádu (x_3y_3 je konec, další dva body jsou směry odchodu a příchodu do koncových bodů)), h (close = nakreslí úsečku do počátku úseku), $xywh$ re (založí nový úsek a zkonstruuje obdélník), S (stroke, aktuálními parametry gstate se obtáhne cesta a zruší se), s (close & stroke = uzavře a obtáhne), f/f* (close? & fill = vyplní cestu, f* počítá počet průsečíků a podle parity určí, zda vyplnit, f počítá počet orientovaných průsečíků a porovnává s 0), b/b* (close & fill & stroke), n (new = discard).

^aMáme user space, grafický stav pak musí definovat tzv. CTM (current transform matrix?) a ta zobrazuje user space na device space. Obecně jsou PDF vektory $(x, y, 1)^T$ a CTM je 3×3

Definice 1.10 (Barvy)

Součást gstatu. Existují různé (přesně nedefinované) prostory: g G/g (device gray, $g \in [0, 1]$, větší hodnota jasnější), rgb RG/rg (device RGB), $cmymk$ (device CMYK), malé jsou výplň, velké hranice (stroke).

Poznámka (TeX)

TeX nastaví počátek userspace na aktuální referenci, pokud použijeme samotné `\pdfliteral{...}`. Pokud použijeme `\pdfliteral{...}` začne na začátku stránky.

Pozor

Transformace je lepší dělat přímo TeXem, jinak se rozsypou například odkazy.