



Plongée au cœur des PDF



SOMMAIRE

- Pourquoi s'intéresser aux PDF?
- D'où vient le format PDF?
- Comment est construit un PDF?
 - Format de données générique
 - Langage graphique
 - Flux et filtres
- Lecture d'un PDF
- Optimisation!
 - Chaînes de caractères
 - Supprimer le superflu
 - Optimiser le code graphique
 - Améliorer l'efficacité des filtres
 - Utiliser le meilleur filtre
 - Regrouper pour compresser
 - Cumuler les filtres
- Et DietPDF?

**Pourquoi encore
s'intéresser aux PDF ?**

On n'est pas prêt de s'en débarrasser

- Inconvénients

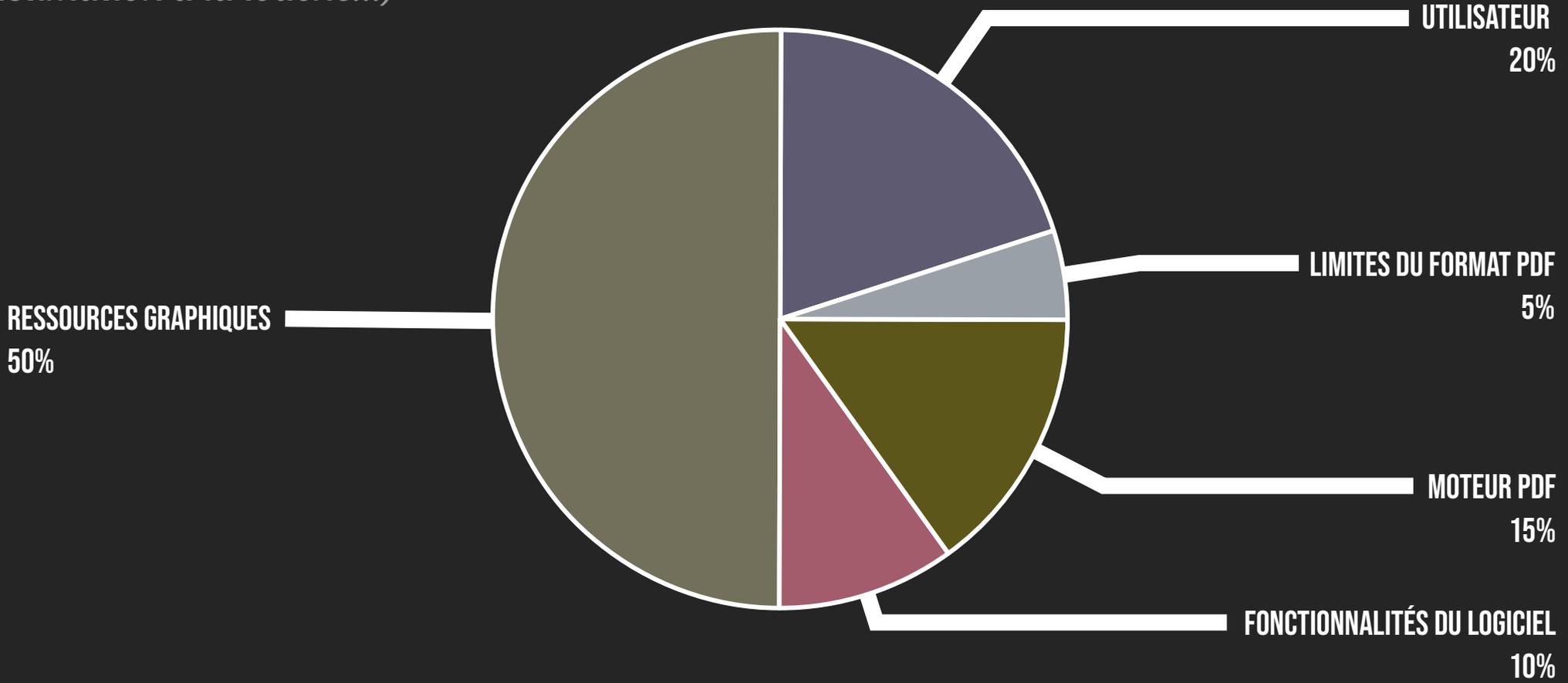
- Mise en page figée
- Peu d'évolutions
- Peu interactif
- Problèmes d'accessibilité

- Avantages

- Structure monolithique
- Adapté à l'archivage
- Signature numérique



(Estimation à la louche...)



Qui est responsable du poids d'un PDF ?

Problèmes des compresseurs de PDF

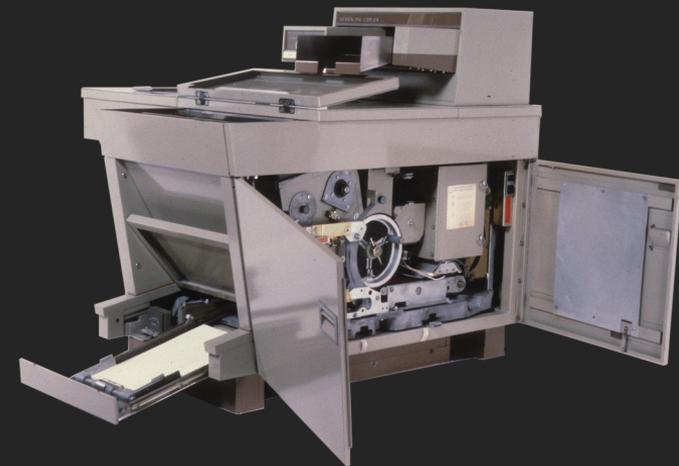
- Optimisation ciblée
- Images maltraitées
 - Compression forte
 - Résolution ≤ 200 ppp
- Conformité
- Métadonnées
- Side-project : **DietPDF**



**D'où vient
le format PDF ?**

Laser game

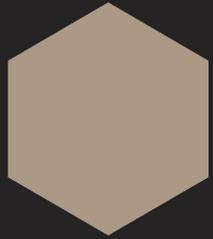
- **1960**, Xerox 914, 1^{er} photocopieur sur papier standard
- **1968**, 1 000 000 000 \$ de CA
- **1970**, ouverture du Palo Alto Research Center
- **1973**, Xerox Alto, le bureau du futur
- **1978**, Création de JaM
... qui deviendra InterPress



Ainsi naquit Adobe

- **1982**
 - Départ de Warnock et Geschke, déçus par Xerox
 - Création d'Adobe
 - Création de PostScript, héritier d'InterPress
- **1985**
 - Apple LaserWriter, 1^{re} imprimante PostScript
 - Décollage du marché de la PAO

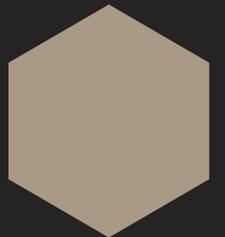




Ce problème concerne notre capacité à communiquer du matériel visuel entre différentes applications et systèmes informatiques.

Il n'existe pas de moyen universel de communiquer et de visualiser ces informations imprimées par voie électronique.

**LE PROJET CAMELOT
JOHN WARNOCK (1990)**

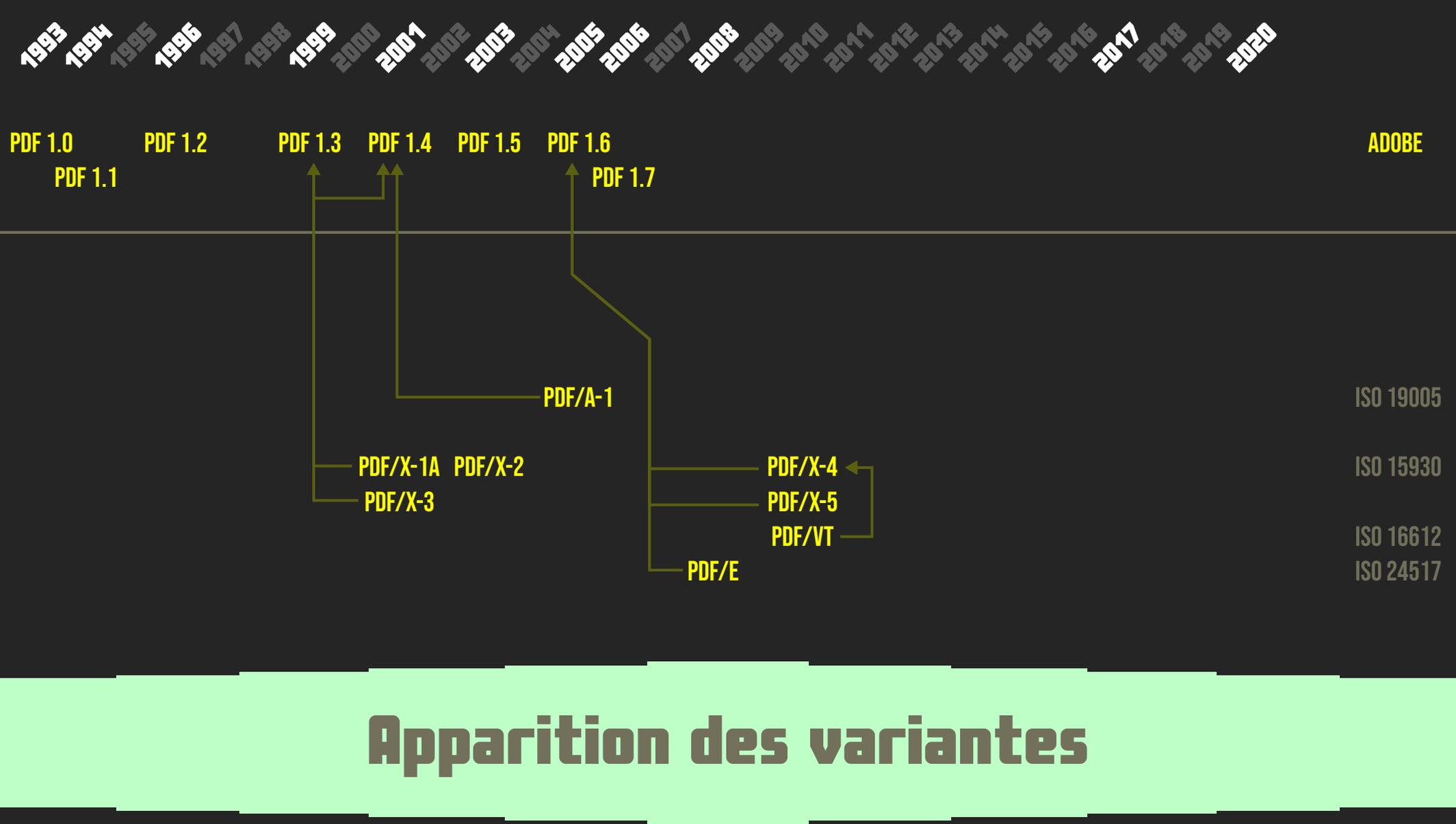




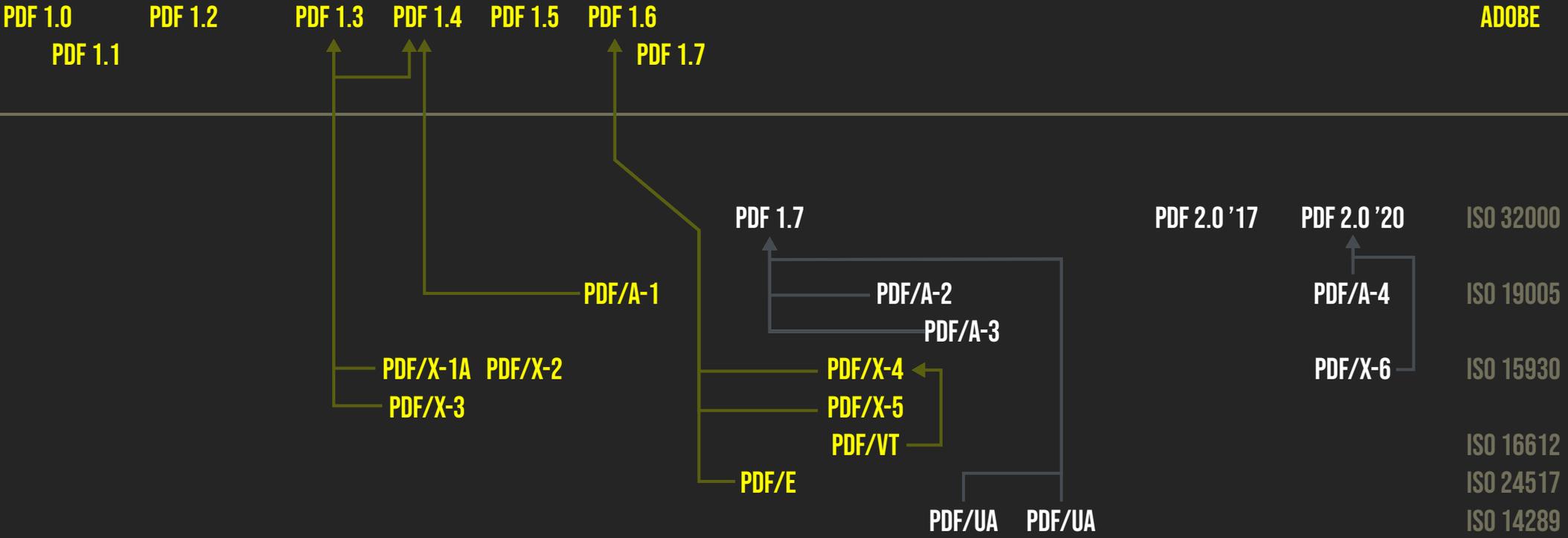
1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020

PDF 1.0 PDF 1.1 PDF 1.2 PDF 1.3 PDF 1.4 PDF 1.5 PDF 1.6 PDF 1.7 ADOBE

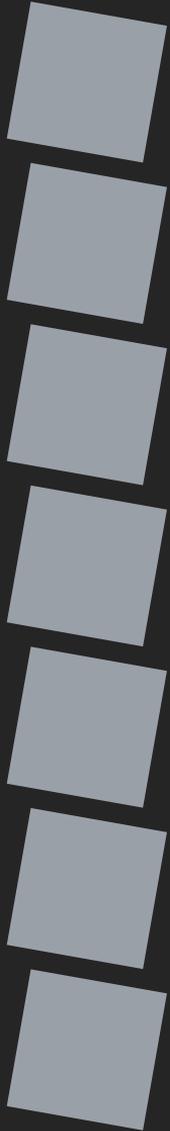
Période Adobe



1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020



Période ISO



Variantes du PDF

- **Archivage..... PDF/A**
JavaScript/multimédia interdits, profil ICC, polices embarquées, balisage strict...
- **Échange de données graphiques.....PDF/X**
Prêt pour l'impression, pas de multimédia, profil ICC, polices embarquées...
- **Impression variable et transactionnelle..... PDF/VT**
Limitation mémoire, zones variables...
- **Documentation technique..... PDF/E**
Contenu 3D, dessins complexes, transfert de données techniques, précision...
- **Accessibilité..... PDF/UA**

**Comment est construit
un PDF ?**



UN FORMAT DE DONNÉES GÉNÉRIQUE



```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Un langage inspiré de PostScript

```
% Un commentaire
```

```
4 0 obj
```

```
<<
```

```
  /Type      /Page
```

```
  /Parent    3 0 R
```

```
  /MediaBox  [0 0 842 595]
```

```
  /Title     (Hello)
```

```
  /Creator   <48656C6C6F>
```

```
  /Resources << /Font 6 0 R >>
```

```
>>
```

```
stream ... endstream
```

```
endobj
```

Commentaires

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Mots-clés

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Nombres

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Chaînes de caractères

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Chaînes hexadécimales

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Noms

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Objets

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Flux

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Dictionnaires

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Clés de dictionnaires

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Valeurs de dictionnaires

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Tableaux

```
% Un commentaire
4 0 obj
  <<
    /Type      /Page
    /Parent    3 0 R
    /MediaBox  [0 0 842 595]
    /Title     (Hello)
    /Creator   <48656C6C6F>
    /Resources << /Font 6 0 R >>
  >>
  stream ... endstream
endobj
```

Références

Table des références croisées

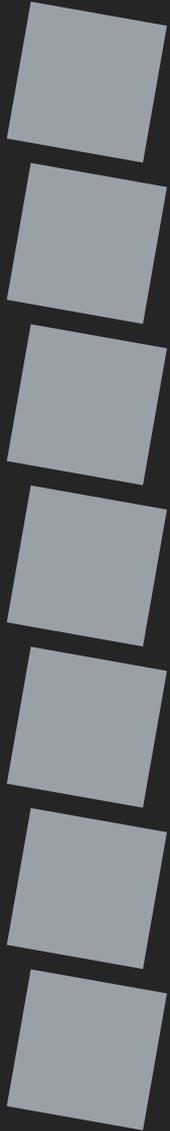
- Accès direct aux objets
- Offsets dans le fichier
- Deux formats de table
 - historique, en clair
 - Compressée, PDF ≥ 1.5

xref

1 7

0000000028	00000	n
0000000257	00000	n
0000000339	00000	n
0000000433	00000	n
0000000605	00000	n
0000000926	00000	n
0000000990	00000	n

Une table de références croisées historique apparaît telle qu'elle dans le fichier PDF



Trailer (annonce)

- Un dictionnaire non référencé
- Juste après la table des références croisées
- Références
 - Vers le catalogue du fichier PDF
 - Vers les métadonnées (titre, auteurs, dates de création...)
- Identifiant du PDF



UN LANGAGE GRAPHIQUE



```
% Débute un bloc de texte
BT

% Dessine en noir en mode nuances de gris
0 g

% Utilise la fonte R7 à 96 points
/R7 96 Tf

% Définit la matrice de transformation du texte
1 0 0 1 170 270 Tm

% Écrit « Hello world! »
(Hello world!) Tj

% Fin du bloc de texte
ET
```

Écrire « Hello world! »

```
% Débute un bloc de texte
```

```
BT
```

```
% Dessine en noir en mode nuances de gris
```

```
0 gg
```

```
% Utilise la fonte R7 à 96 points
```

```
/R7 96 Tf
```

```
% Définit la matrice de transformation du texte
```

```
1 0 0 1 170 270 Tm
```

```
% Écrit « Hello world! »
```

```
(Hello world!) Tj
```

```
% Fin du bloc de texte
```

```
ET
```

Commandes graphiques

```
% Débute un bloc de texte
```

```
BT
```

```
% Dessine en noir en mode nuances de gris
```

```
0 g
```

```
% Utilise la fonte R7 à 96 points
```

```
/R7 96 Tf
```

```
% Définit la matrice de transformation du texte
```

```
1 0 0 1 170 270 Tm
```

```
% Écrit « Hello world! »
```

```
(Hello world!) Tj
```

```
% Fin du bloc de texte
```

```
ET
```

Paramètres

Hello world!

Résultat du code « Hello world! »



DES FLUX ET DES FILTRES



La ressource est dans le flux (stream)

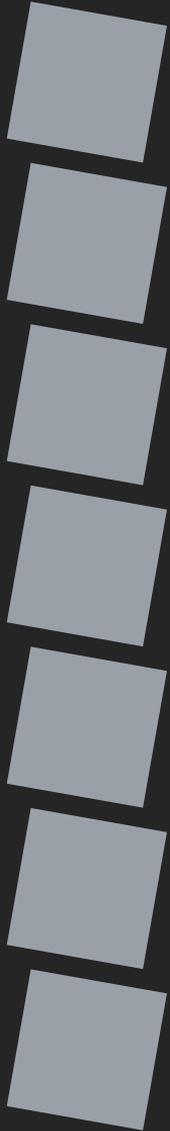
- Images
 - Vectorielles
 - Matricielles
- Fichiers embarqués
 - Fichiers bureautiques
 - Métadonnées XMP
 - Vidéos
- Polices de caractères
 - TrueType
 - PostScript



Chaque flux est « filtré »

- **Encodage**
Ascii85, Hex
- **Compression sans perte**
RLE, LZW, Flate, CCITTFax, JPEG2000
- **Compression avec perte**
JPEG, JPEG2000, JBIG2
- **Prédicteurs pour certains filtres**
TIFF, PNG
- **Les filtres peuvent se cumuler!**





Exemple de flux filtré

```
36 0 obj
```

```
<<
```

```
  /Length 40
```

```
  /Filter /ASCIIHexDecode
```

```
>>
```

```
stream
```

```
48 65 6c 6c 6f 2c 20 57 6f 72 6c 64 21 >
```

```
endstream
```

```
endobj
```

Longueur du flux avant décodage

36 0 obj

<<

/Length 40

Le flux fait 40 octets

/Filter /ASCIIHexDecode

>>

stream

48 65 6c 6c 6f 2c 20 57 6f 72 6c 64 21 >

endstream

endobj

Encodage du flux

```
36 0 obj
```

```
<<
```

```
  /Length 40
```

```
  /Filter /ASCIIHexDecode
```

```
>>
```

```
stream
```

```
48 65 6c 6c 6f 2c 20 57 6f 72 6c 64 21 >
```

```
endstream
```

```
endobj
```

Le flux a été encodé
en hexadécimal

Contenu du flux décodé

```
36 0 obj
```

```
<<
```

```
  /Length 40
```

```
  /Filter /ASCIIHexDecode
```

```
>>
```

```
stream
```

```
48 65 6c 6c 6f 2c 20 57 6f 72 6c 64 21 >
```

```
endstream
```

```
endobj
```

Le flux contient la chaîne de caractères «Hello, World!»

Lecture d'un PDF

Un format conçu en 1993 pour...

- La transmission
 - Fun fact: un PDF peut être **entièrement** encodé en ASCII!
- Des tailles de fichier énormes
 - PDF < 1.5: maximum de ~9,3 Gio
 - PDF ≥ 1.5: pas de limite théorique
- La mise à jour incrémentale
 - CRUD = Simple ajout à la fin du fichier





Vérification de l'entête

L'entête d'un PDF

- Signature %PDF - {version}
1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7 et 2.0
- Commentaire «détrompeur» (optionnel)
4 caractères 8 bits non ASCII (> 127)

```
00000000 25 50 44 46 2d 31 2e 37 0a 25 e2 e3 cf d3 0d 0a |%PDF-1.7.%.....|
00000010 39 30 37 39 33 20 30 20 6f 62 6a 0a 3c 3c 2f 4c |90793 0 obj.<</L|
00000020 69 6e 65 61 72 69 7a 65 64 20 31 2f 4c 20 31 34 |linearized 1/L 14|
00000030 37 32 31 30 38 38 2f 4f 20 39 30 37 39 35 2f 45 |721088/0 90795/E|
00000040 20 31 30 30 36 33 39 2f 4e 20 31 30 30 33 2f 54 | 100639/N 1003/T|
00000050 20 31 34 37 30 39 36 34 36 2f 48 20 5b 20 33 34 | 14709646/H [ 34|
00000060 31 34 20 39 39 35 35 5d 3e 3e 0a 65 6e 64 6f 62 |14 9955]>>.endob|
```



ENTÊTE

ANNONCE
MISE À JOUR

Recherche de l'annonce

Repérer les références croisées

- Fin de fichier `%%EOF`

Peut se trouver n'importe où dans les 1024 derniers octets!

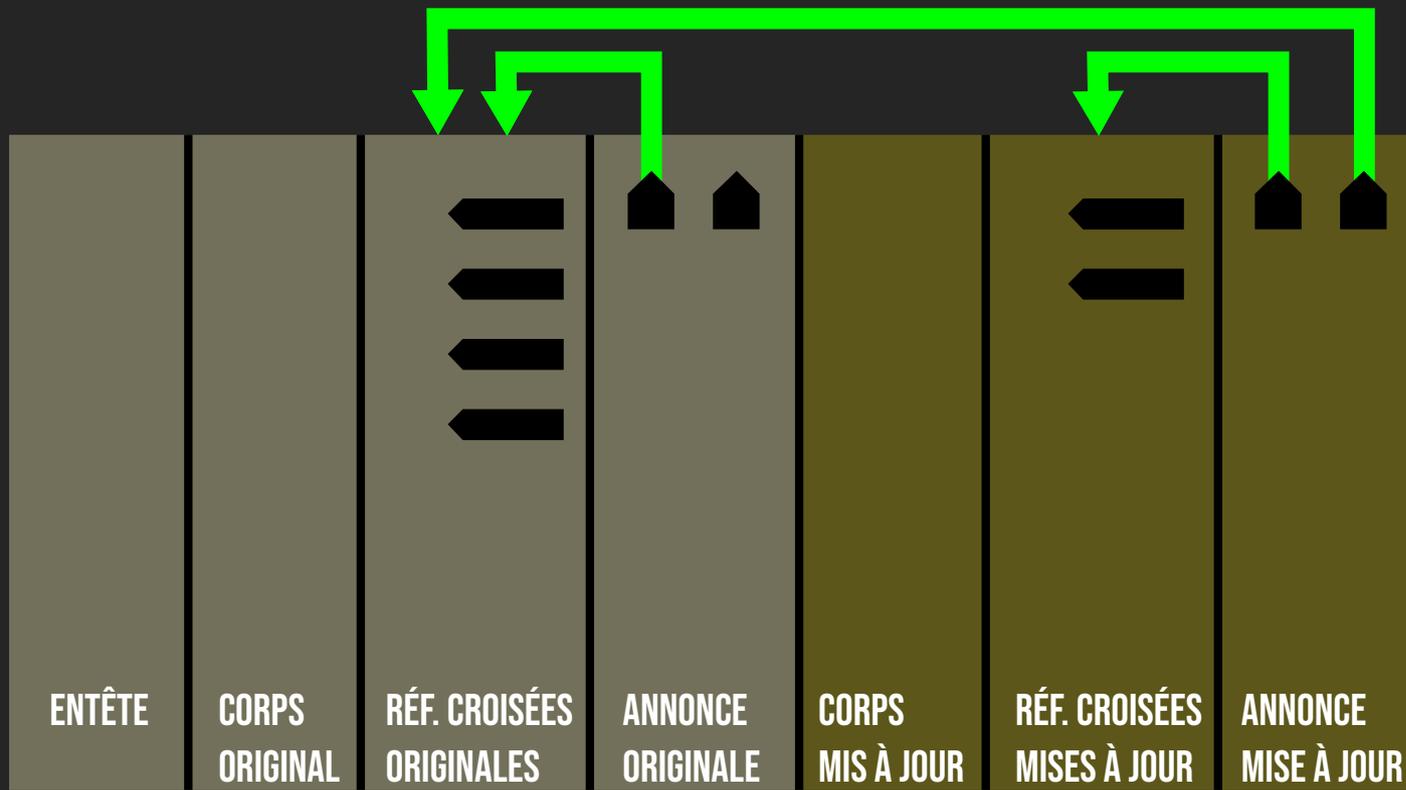
- Offset de la table des références croisées `startxref`

En base 10!

```
00e09fe0 f1 0a 42 bc 82 78 05 f1 0a 42 bc 82 78 05 f1 0a |..B..x...B..x...|
00e09ff0 42 bc 82 78 05 f1 0a 42 bc 82 78 05 f1 0a c2 e9 |B..x...B..x.....|
00e0a000 7a 55 fc c1 b7 ab 4f d2 7a 90 28 07 f1 ea af ba |zU....0.z.(.....|
00e0a010 dd be f3 ff 06 00 79 9f aa 6a 0d 65 6e 64 73 74 |.....y..j.endst|
00e0a020 72 65 61 6d 0d 65 6e 64 6f 62 6a 0d 73 74 61 72 |ream.endobj.star|
00e0a030 74 78 72 65 66 0a 31 31 36 0a 25 25 45 4f 46 0a |txref.116.%%EOF. |
```



Recherche des références croisées



Décodage des références croisées

Décoder les références croisées

```
00000070 6a 0d 20 0d 39 30 38 32 34 20 30 20 6f 62 6a 0d |j. .90824 0 obj.|
00000080 3c 3c 2f 44 65 63 6f 64 65 50 61 72 6d 73 3c 3c |<</DecodeParms<<|
00000090 2f 43 6f 6c 75 6d 6e 73 20 35 2f 50 72 65 64 69 |/Columns 5/Predi|
000000a0 63 74 6f 72 20 31 32 3e 3e 2f 46 69 6c 74 65 72 |ctor 12>>/Filter|
000000b0 2f 46 6c 61 74 65 44 65 63 6f 64 65 2f 49 44 5b |/FlateDecode/ID[|
000000c0 3c 32 42 35 35 31 44 32 41 46 45 35 32 36 35 34 |<2B551D2AFE52654|
000000d0 34 39 34 46 39 37 32 30 32 38 33 43 46 46 31 43 |494F9720283CFF1C|
000000e0 34 3e 3c 33 43 44 41 38 42 42 36 44 35 38 33 34 |4><3CDA8BB6D5834|
000000f0 45 34 31 41 35 45 32 41 41 31 36 43 33 35 45 34 |E41A5E2AA16C35E4|
00000100 43 34 37 3e 5d 2f 49 6e 64 65 78 5b 39 30 37 39 |C47>]/Index[9079|
00000110 33 20 31 30 31 34 5d 2f 49 6e 66 6f 20 39 30 37 |3 1014]/Info 907|
00000120 39 32 20 30 20 52 2f 4c 65 6e 67 74 68 20 31 38 |92 0 R/Length 18|
00000130 35 2f 50 72 65 76 20 31 34 37 30 39 36 34 37 2f |5/Prev 14709647/|
00000140 52 6f 6f 74 20 39 30 37 39 34 20 30 20 52 2f 53 |Root 90794 0 R/S|
00000150 69 7a 65 20 39 31 38 30 37 2f 54 79 70 65 2f 58 |ize 91807/Type/X|
00000160 52 65 66 2f 57 5b 31 20 33 20 31 5d 3e 3e 73 74 |Ref/W[1 3 1]>>st|
```

Et si on simplifiait la lecture de ce dump hexadécimal ?

Un flux de références croisées

```
90824 0 obj <<
```

```
  /Type      /Xref  
  /Size      91807  
  /Index     [90793 1014]  
  /Filter    /FlateDecode  
  /DecodeParms <</Columns 5 /Predictor 12>>  
  /Length    185  
  /W        [1 3 1]  
  /Info     90792 0 R  
  /Root     90794 0 R  
  /Prev    14709647  
  /ID      [...]
```

```
>>
```

Cette ligne indique des références croisées compressées

Numéro de départ des futurs objets

```
90824 0 obj <<  
  /Type      /Xref  
  /Size      91807  
  /Index     [90793 1014]  
  /Filter    /FlateDecode  
  /DecodeParms <</Columns 5 /Predictor 12>>  
  /Length    185  
  /W         [1 3 1]  
  /Info      90792 0 R  
  /Root      90794 0 R  
  /Prev      14709647  
  /ID        [...]  
>>
```

Utilisé pour de potentielles futures mises à jour du PDF

Numéro du 1^{er} objet et nombre d'objets

```
90824 0 obj <<
  /Type      /Xref
  /Size      91807
  /Index     [90793 1014]
  /Filter    /FlateDecode
  /DecodeParms <</Columns 5 /Predictor 12>>
  /Length    185
  /W        [1 3 1]
  /Info      90792 0 R
  /Root      90794 0 R
  /Prev      14709647
  /ID        [...]
>>
```

$$90793 + 1014 = 91807$$

Flux compressé par Flate (Zlib)

```
90824 0 obj <<
  /Type      /Xref
  /Size      91807
  /Index     [90793 1014]
  /Filter    /FlateDecode
  /DecodeParms <</Columns 5 /Predictor 12>>
  /Length    185
  /W        [1 3 1]
  /Info     90792 0 R
  /Root     90794 0 R
  /Prev    14709647
  /ID      [...]
>>
```

Prédicteur Up et entrée = 5 octets

```
90824 0 obj <<
  /Type      /Xref
  /Size      91807
  /Index     [90793 1014]
  /Filter    /FlateDecode
  /DecodeParms <</Columns 5 /Predictor 12>>
  /Length    185
  /W        [1 3 1]
  /Info     90792 0 R
  /Root     90794 0 R
  /Prev    14709647
  /ID      [...]
>>
```

1 entrée de la table = 5 octets
Prédicteur utilisé = «Up»

Flux compressé = 185 octets

```
90824 0 obj <<
  /Type      /Xref
  /Size      91807
  /Index     [90793 1014]
  /Filter    /FlateDecode
  /DecodeParms <</Columns 5 /Predictor 12>>
  /Length    185
  /W        [1 3 1]
  /Info     90792 0 R
  /Root     90794 0 R
  /Prev    14709647
  /ID      [...]
>>
```

Taille des champs en octets

```
90824 0 obj <<  
  /Type          /Xref  
  /Size          91807  
  /Index         [90793 1014]  
  /Filter        /FlateDecode  
  /DecodeParms  <</Columns 5 /Predictor 12>>  
  /Length        185  
  /W             [1 3 1]  
  /Info          90792 0 R  
  /Root          90794 0 R  
  /Prev         14709647  
  /ID            [...]  
>>
```

1 octet pour le type de l'entrée
3 octets pour un offset/n° d'objet
1 octet pour la génération/index

Métadonnées du document

```
90824 0 obj <<
  /Type      /Xref
  /Size      91807
  /Index     [90793 1014]
  /Filter    /FlateDecode
  /DecodeParms <</Columns 5 /Predictor 12>>
  /Length   185
  /W        [1 3 1]
  /Info     90792 0 R
  /Root     90794 0 R
  /Prev     14709647
  /ID      [...]
>>
```

Les métadonnées sont
dans l'objet 90792

Catalogue du document

```
90824 0 obj <<
  /Type      /Xref
  /Size      91807
  /Index     [90793 1014]
  /Filter    /FlateDecode
  /DecodeParms <</Columns 5 /Predictor 12>>
  /Length   185
  /W        [1 3 1]
  /Info     90792 0 R
  /Root     90794 0 R
  /Prev    14709647
  /ID      [...]
>>
```

Le catalogue est dans
l'objet 90794

Offset des références précédentes

```
90824 0 obj <<
  /Type      /Xref
  /Size      91807
  /Index     [90793 1014]
  /Filter    /FlateDecode
  /DecodeParms <</Columns 5 /Predictor 12>>
  /Length    185
  /W        [1 3 1]
  /Info      90792 0 R
  /Root      90794 0 R
  /Prev      14709647
  /ID        [...]
>>
```

Une précédente table de références croisées existe à l'offset 14709647. La table courante est prioritaire sur elle.

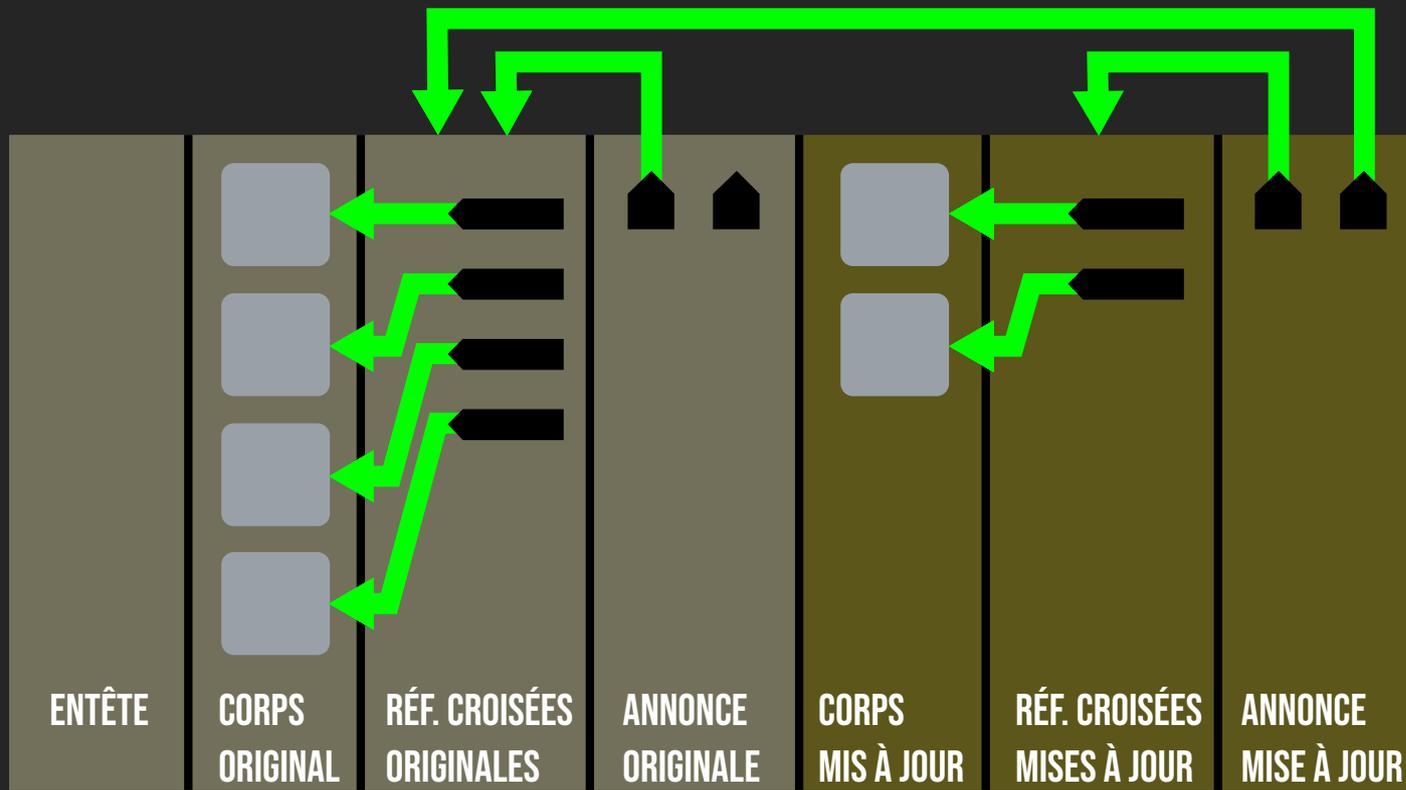
Identifiant du fichier PDF

```
90824 0 obj <<
  /Type      /Xref
  /Size      91807
  /Index     [90793 1014]
  /Filter    /FlateDecode
  /DecodeParms <</Columns 5 /Predictor 12>>
  /Length    185
  /W        [1 3 1]
  /Info     90792 0 R
  /Root     90794 0 R
  /Prev     14709647
  /ID      [...]
>>
```

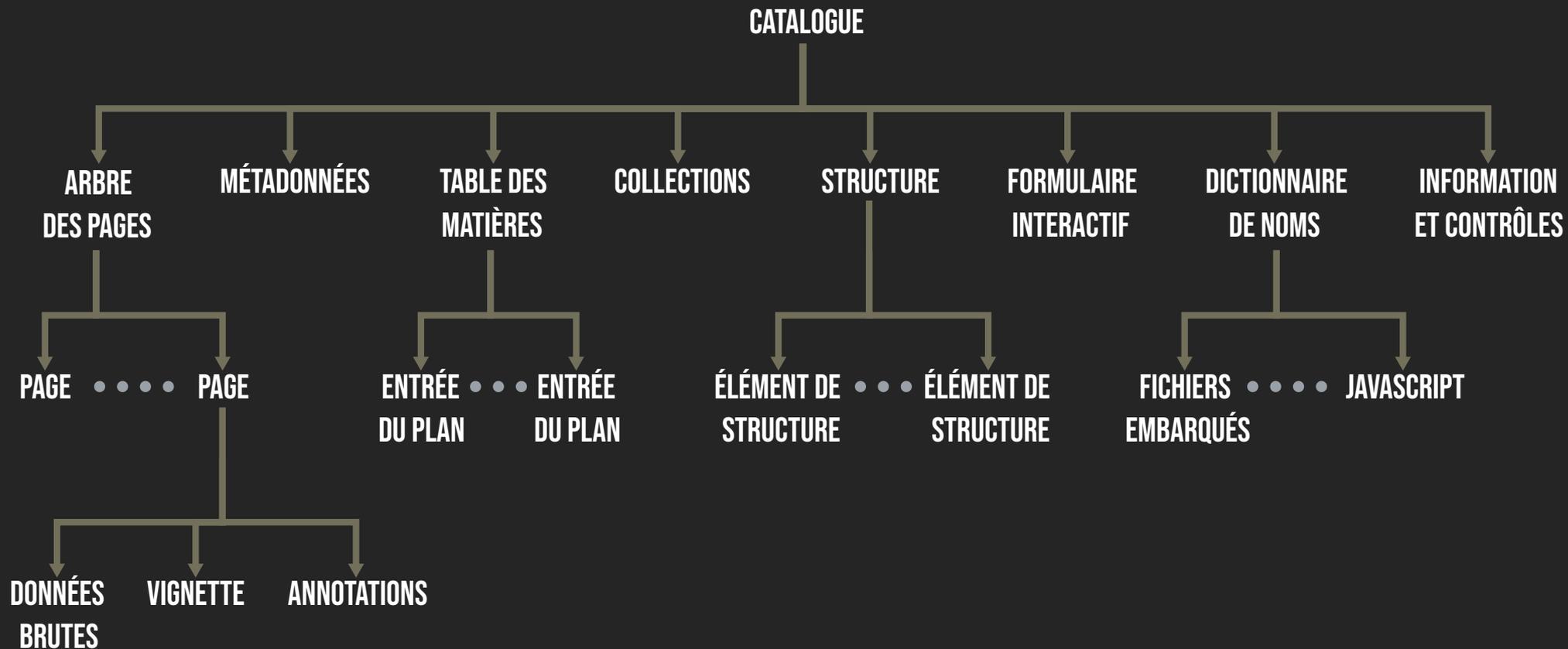
Trouver l'offset du catalogue

- Décoder la table des références croisées
- 3 colonnes par entrée
 - Type d'entrée: 1
 - Offset: 0x3439 ou 13369
 - Numéro de génération: 0

90793	01	000010	00
90794	01	003439	00
90795	01	00363E	00
90796	01	003E53	00
90797	01	004228	00
...			
90825	02	0162AC	00
90826	02	0162AC	01
90827	02	0162AC	02
90828	02	0162AC	03
90829	02	0162AC	04
...			



Recherche des objets



Structure logique d'un document PDF

The image features a dark, almost black background. In the center, the word "Optimisation!" is written in a bold, white, sans-serif font. On either side of the text, there is a light blue square. The square on the left is positioned to the left of the text, and the square on the right is positioned to the right of the text. Both squares are of equal size and are partially cut off by the edges of the frame.

Optimisation!



LES CHAÎNES DE CARACTÈRE



2 façons de coder des chaînes

- Chaînes littérales

(A string)

((word))

(P)

Pas besoin d'échappement
si les parenthèses sont
équilibrées

- Chaînes hexadécimales

<4120737472696E67>

<28 77 6F 72 64 29>

<5>

Le zéro terminal est
optionnel, 5 = 50

2 encodages, 5 types de chaînes

Type	Caractères	Litérale	Hexadécimale
Chaîne d'octets <i>byte string</i>	256	n + échappements	2n
ASCII <i>pas d'accents!</i>	127	n + échappements	2n
UTF-16BE <i>big endian</i>	Unicode	$2+2n \leq \text{taille} \leq 2+4n$ + échappements	$4+4n \leq \text{taille} \leq 4+8n$
UTF-8 <i>PDF 2.0+</i>	Unicode	$3+n \leq \text{taille} \leq 3+4n$ + échappements	$6+2n \leq \text{taille} \leq 6+8n$
PDFDocEncoded <i>~Latin-1</i>	256	n + échappements	2n



SUPPRIMER LE SUPERFLU



Que supprimer ?



- **Éléments inutiles**
 - Retours chariots
 - Espaces
 - Commentaires
 - Caractères inutiles
- **Précision inutile**
- **Objets inutilisés**

- **Éléments**
 - Cachés
 - Hors-cadres
 - Dupliqués
- **Instructions redondantes**
- **Valeurs identiques aux valeurs par défaut**



Code non minifié

```
4 0 obj
```

```
<<
```

```
  /Type      /Page
```

```
  /Parent    3 0 R
```

```
  /MediaBox  [0 0 842 595]
```

```
  /Contents  5 0 R
```

```
  /Resources << /Font 6 0 R >>
```

```
>>
```

```
endobj
```

148 octets dont 45 inutiles

```
00000000 34 20 30 20 6f 62 6a 0a 20 20 3c 3c 0a 20 20 20 |4 0 obj. <<. |
00000010 20 2f 54 79 70 65 20 20 20 20 20 20 2f 50 61 67 | /Type /Pag|
00000020 65 0a 20 20 20 20 2f 50 61 72 65 6e 74 20 20 20 |e. /Parent |
00000030 20 33 20 30 20 52 0a 20 20 20 20 2f 4d 65 64 69 | 3 0 R. /Medi|
00000040 61 42 6f 78 20 20 5b 30 20 30 20 38 34 32 20 35 |aBox [0 0 842 5|
00000050 39 35 5d 0a 20 20 20 20 2f 43 6f 6e 74 65 6e 74 |95]. /Content|
00000060 73 20 20 35 20 30 20 52 0a 20 20 20 20 2f 52 65 |s 5 0 R. /Re|
00000070 73 6f 75 72 63 65 73 20 3c 3c 20 2f 46 6f 6e 74 |sources << /Font|
00000080 20 36 20 30 20 52 20 3e 3e 0a 20 20 3e 3e 0a 65 | 6 0 R >>. >>.e|
00000090 6e 64 6f 62 6a |ndobj|
```

Un dump hexadécimal permet de mieux se rendre compte

Précision inutile, exemple 1

- Ex.: 0.14509 0.14509 0.14509 RG
 rouge vert bleu couleur du tracé
- 5 chiffres de précision
 - 100 000 niveaux de rouge, de vert et de bleu
 - 1 000 000 000 000 000 de couleurs possibles!
 - L'œil humain voit 10 000 000 de couleurs et 30 niveaux de gris
<https://royalsocietypublishing.org/doi/full/10.1098/rsif.2012.0601>

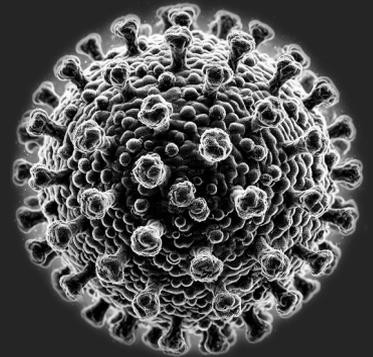
Précision inutile, exemple 2

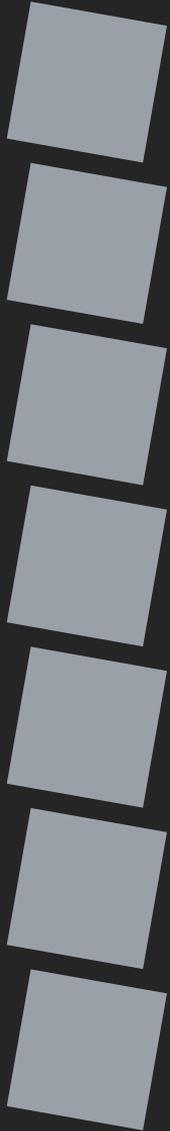
- Ex.: 0 0.028 793.672 446.428 re

x	y	largeur	hauteur	trace un rectangle
---	---	---------	---------	--------------------

- L'unité utilisée est le point

- 1 point, 1/72^e pouce.....0,353 mm
- 0,001 point..... 0,000 353 mm
- Virus de la grippe.....0,000 100 mm





Réduction de la précision

- Dépend des commandes
 - Couleur
 - Tracé
 - Texte
- Dépend du contexte
 - Matrice de transformation
 - Pile d'état graphique
- Solution
 - Précision variable
 - Pseudo-interpréteur



Le point dans la matrice

- Tous les tracés subissent une matrice de transformation
- $12.345 \times 1 == 12345 \times 0.001$
 - 1 point décimal = 1 octet
 - 1 coordonnée = 2 points décimaux
= 2 octets

Valeurs par défaut inutiles

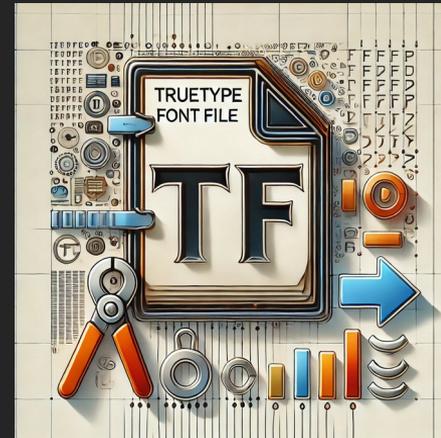
```
4 0 obj
<<
  /Length 37
  /Filter [/ASCIIHexDecode /FlateDecode]
  /DecodeParms [
    null
    << /Predictor 11 /Columns 26 /Components 8 /Colors 1 >>
  ]
>>
stream
78 da 63 74 64 c4 05 00 08 15 00 5c >
endstream
endobj
```

Valeurs par défaut supprimées

```
4 0 obj
<<
  /Length 37
  /Filter [/ASCIIHexDecode /FlateDecode]
  /DecodeParms [
    null
    << /Predictor 11 /Columns 26 >>
  ]
>>
stream
78 da 63 74 64 c4 05 00 08 15 00 5c >
endstream
endobj
```


Optimiser les fontes True Type

- **Suppression des informations de « hinting »**
Des informations dédiées à l'amélioration du rendu sur écran mais souvent ignorées
- **Suppression des commentaires**



Supprimer les éléments inutiles

- Tout élément ou image
 - Dupliqué
 - Utilisant des masques de découpe
 - Sortant du cadre
 - Masqué par d'autres éléments
 - Utilisant une résolution ≥ 300 ppp
- Un problème complexe !

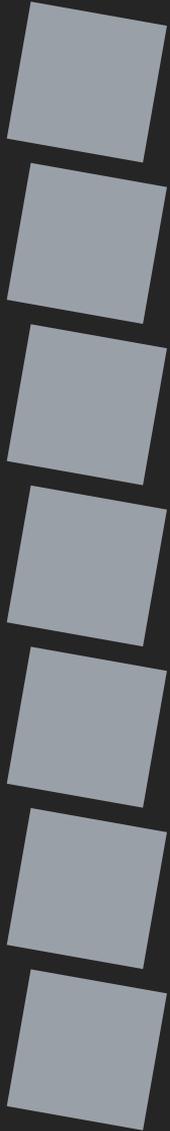


PowerPoint vs Impress

- Comparatif des stratégies face à des images
- Test effectué sur

– PowerPoint	Microsoft Office 365	16.0.16904.40516
– Impress	LibreOffice	24.8.0.3





Création des échantillons

1) Créer la 1^{re} diapo

- Créer un document vide
- Importer un JPEG
- Rogner le JPEG
- Créer une zone de texte
- Insérer du texte

2) Créer la 2^e diapo

- Dupliquer la 1^{re} diapo
- Supprimer le rognage

3) Créer la 3^e diapo

- Dupliquer la 2^e diapo
- Déplacer le JPEG

4) Exporter en PDF

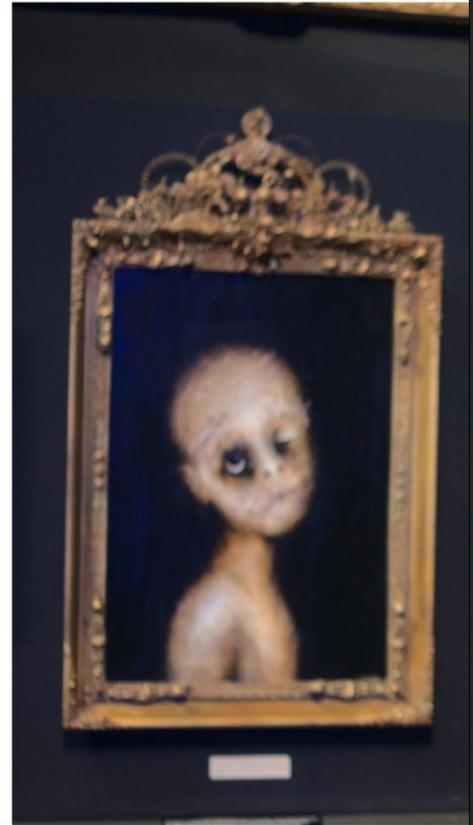
La plupart des hommes ne s'étonnent point assez. En présence des plus grands phénomènes, des inventions les plus admirables, on les voit trop souvent indifférents, impassibles. C'est le propre de la matière d'être impassible, et non pas de l'esprit. Ceux dont la curiosité est toujours en éveil, qui aiment à s'expliquer ce qu'ils voient, qui recherchent les causes, ceux-là seuls parviennent à s'instruire, à s'éclairer, à augmenter leurs jouissances intellectuelles, et peuvent, s'ils sont doués de quelque supériorité, contribuer à l'avancement des sciences et de leurs applications, c'est-à-dire au progrès du bien-être de leurs semblables et de la civilisation. Voici, par exemple, les chemins de fer et le télégraphe électrique qui ne datent que de peu d'années : on s'y est déjà si bien habitué qu'il ne semble que ces merveilleuses inventions aient existé de tout...



Diapo 1, JPEG rogné

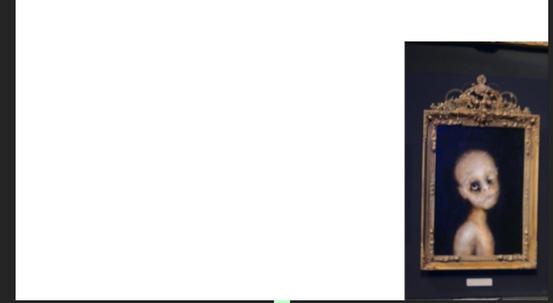


Diapo 2, JPEG complet



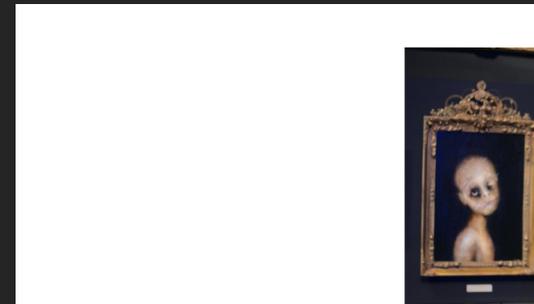
Diapo 3, JPEG débordant du cadre

La plupart des hommes ne s'étonnent point assez. En présence des plus grands phénomènes, des inventions les plus admirables, on les voit trop souvent indifférents, impassibles. C'est le propre de la matière d'être impassible, et non pas de l'esprit. Ceux dont la curiosité est toujours en éveil, qui aiment à s'expliquer ce qu'ils voient, qui recherchent les causes, ceux-là seuls parviennent à s'instruire, à s'éclairer, à augmenter leurs jouissances intellectuelles, et peuvent, s'ils sont doués de quelque supériorité, contribuer à l'avancement des sciences et de leurs applications, c'est-à-dire au progrès du bien-être de leurs semblables et de la civilisation. Voici, par exemple, les chemins de fer et le télégraphe électrique qui ne datent que de peu d'années : on s'y est déjà si bien habitué qu'il ne semble que ces merveilleuses inventions aient existé de tout...

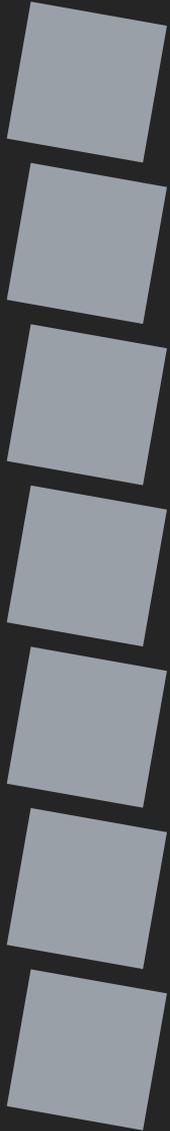


PDF généré par PowerPoint

La plupart des hommes ne s'étonnent point assez. En présence des plus grands phénomènes, des inventions les plus admirables, on les voit trop souvent indifférents, impassibles. C'est le propre de la matière d'être impassible, et non pas de l'esprit. Ceux dont la curiosité est toujours en éveil, qui aiment à s'expliquer ce qu'ils voient, qui recherchent les causes, ceux-là seuls parviennent à s'instruire, à s'éclairer, à augmenter leurs jouissances intellectuelles, et peuvent, s'ils sont doués de quelque supériorité, contribuer à l'avancement des sciences et de leurs applications, c'est-à-dire au progrès du bien-être de leurs semblables et de la civilisation. Voici, par exemple, les chemins de fer et le télégraphe électrique qui ne datent que de peu d'années : on s'y est déjà si bien habitué qu'il ne semble que ces merveilleuses inventions aient existé de tout...



PDF généré par Impress



Les stratégies de génération du PDF

- Rognage

- PowerPoint rogne
- Impress masque

- Duplication

- Pas de duplication mais rognage = dupliquer

- Débordement

- Aucune gestion du débordement

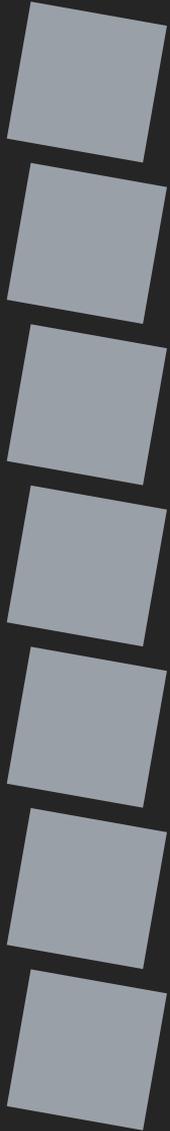
- Résolution

- Adaptation sauf pour Office 365 web



OPTIMISER LE CODE GRAPHIQUE





Tout un programme !

- Espaces inutiles
- Saut de ligne → espace
- Commandes ineffectives
- Commandes plus courtes
- Tri de commandes
- Fusion de commandes
- Sauvegardes et restauration inutiles
- Chaînes de caractères
- Noms des ressources
- Précision des nombres
- Factorisation
- Etc.

Utilisation d'opérateurs dédiés

```
MoveTo( 0, 0)
```

```
LineTo( 0, 20)
```

```
LineTo(50, 20)
```

```
LineTo(50, 0)
```

```
LineTo( 0, 0)
```

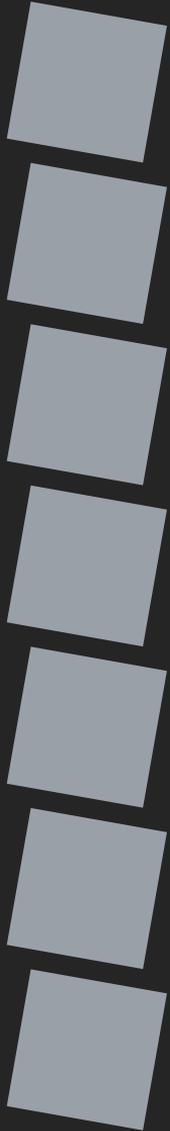
```
CloseSubpath
```

```
FillPath
```

```
Rectangle(0, 0, 50, 20)
```

```
FillPath
```

(Du pseudo-code est utilisé pour simplifier la compréhension)



Opérations inutiles

SaveGS

Rectangle(0, 0, 5, 2)

FillPath

SaveGS

MoveTo(0, 0)

LineTo(0, 20)

StrokePath

RestoreGS

RestoreGS

SaveGS

Rectangle(0, 0, 5, 2)

FillPath

MoveTo(0, 0)

LineTo(0, 20)

StrokePath

RestoreGS

(Du pseudo-code est utilisé pour simplifier la compréhension)



AMÉLIORER L'EFFICACITÉ DES FILTRES

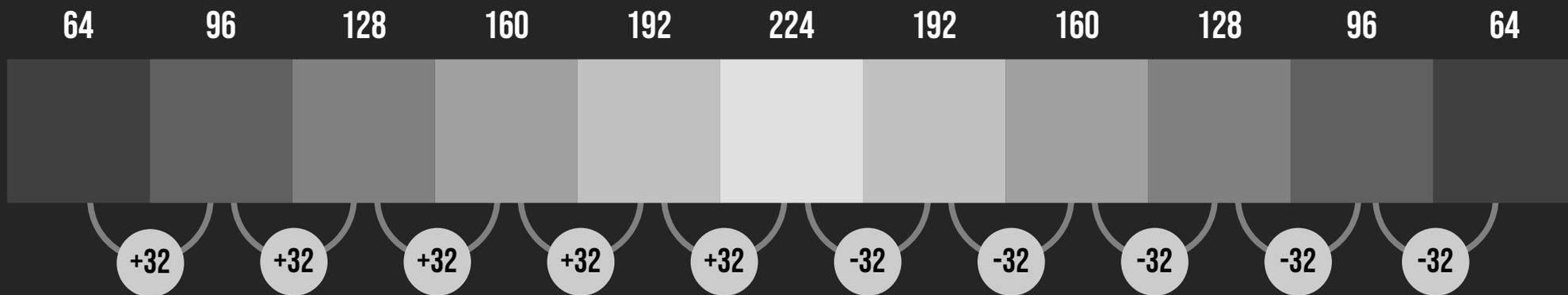


Un dégradé a peu de redondance de données
Une compression par dictionnaire sera inefficace



Faible taux de compression

Le prédicteur Sub calcule la différence entre chaque pixel



Prédicteur Sub

Des répétitions apparaissent après application du prédicteur
Seule la première valeur est conservée telle qu'elle



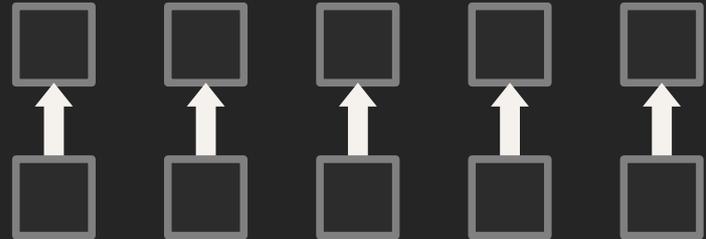
Fort taux de compression

Prédicteurs disponibles (1/2)

- Sub

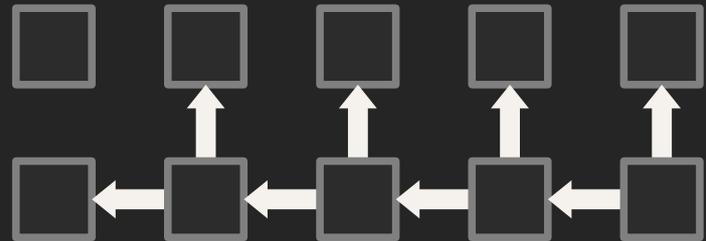


- Up

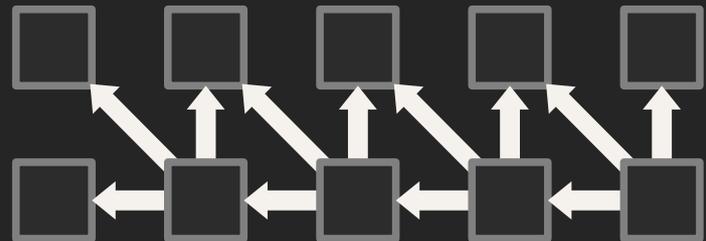


Prédicteurs disponibles (2/2)

- Average

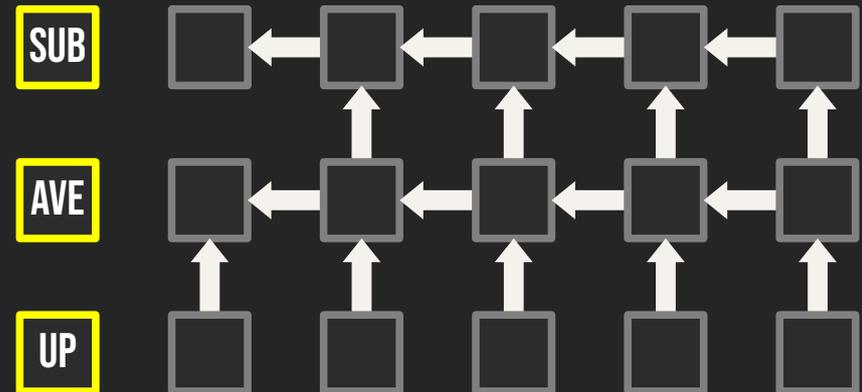


- Paeth



Un prédicteur PNG par ligne

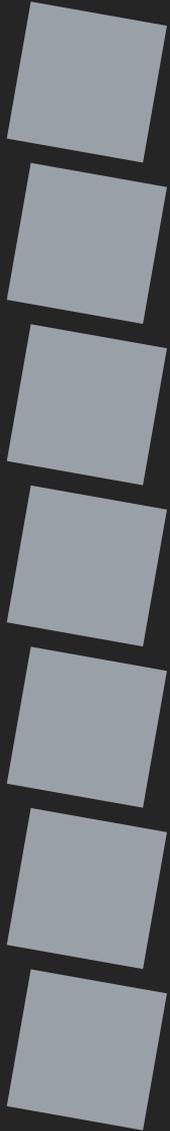
- 1 octet en début de ligne
0 = aucun / 1 = sub / 2 = up / 3 = average / 4 = paeth
- Comment trouver la combinaison optimale?
 - Test de toutes les combinaisons possibles
 - Calcul d'entropie
 - Réseau neuronal?





UTILISER LE MEILLEUR FILTRE





Histoire de filtres

- **PDF 1.0 (1993)**
 - JPEG
 - CCITT Group 3/4
 - LZW
 - RLE
 - Ascii85
 - Hex
 - Prédicteur TIFF
- **PDF 1.2 (1996)**
 - Flate
 - Prédicteurs PNG
- **PDF 1.4 (2001)**
 - JBIG2
- **PDF 1.5 (2003)**
 - JPEG 2000

Remplacement de filtres

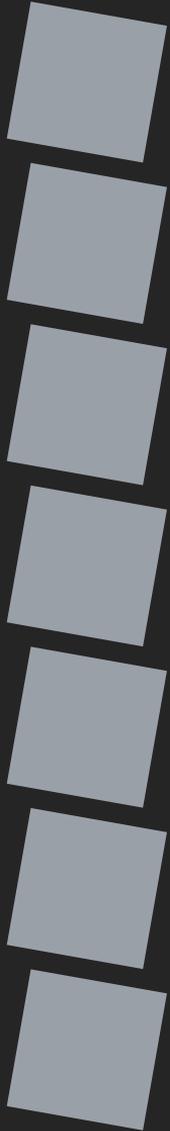
- Ascii85 → *Rien!*
 - Hex → *Rien!*
 - LZW → Flate
 - Flate → JPEG 2000?
lossless
 - JPEG → JPEG 2000?
lossy
 - CCITT → JBIG2
- Ascii85 et Hex
 - Pas de chiffrement!
 - Pour des PDF ASCII
 - Surpoids



JPEG 2000 or not JPEG 2000 ?

- Réencodage JPEG
cumul d'artéfacts, limites de gain
- JPEG 2000 lossless
peu efficace pour les masques
- Limites d'utilisation
autorisé à partir de PDF/A-2





Optimisation du JPEG

- **Progressive vs baseline**
2 à 10 % de gain en général
- **Codage arithmétique**
5 à 10 % de gain, mais non supporté
- **Vignette**
inutilisée par le PDF
- **CMJN → RVB**
25 à 30 % de gain, mais perte de fonctionnalité
- **Tables de quantisation et Huffman**
jusqu'à 30 % de gain
- **Taux de compression**
gain mais perte d'information
- **Commentaires inutiles**
- **JpegTran, MozJPEG...**

Optimisation de Flate

- Utilisation de Zopfli
 - 👍 Meilleur **taux** de compression que Zlib
 - 👍 Les données compressées par Zopfli restent **lisibles** par Zlib
 - 👎 Temps de compression très important





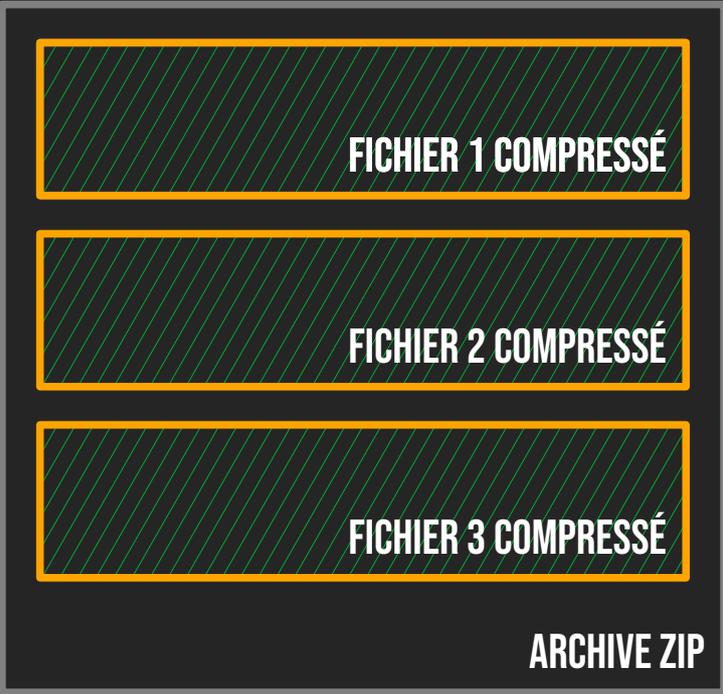
REGROUPER POUR MIEUX COMPRESSER



A vertical decorative bar on the left side of the slide, consisting of eight light blue squares stacked vertically, slightly offset to the right.

Une fonctionnalité introduite en 2003

- Avant PDF 1.5
 - Seuls les flux peuvent être compressés
 - Références croisées et objets sans flux restent en clair
- À partir de PDF 1.5
 - On peut regrouper références croisées et objets sans flux dans le flux d'un objet
 - ... et utiliser les prédicteurs !
 - ... et les compresser !



ZIP

archive de fichiers compressés

accès direct aux fichiers

TAR.GZ

archive compressée de fichiers

compression optimisée



Archive ZIP vs archive TAR.GZ



CUMULER LES FILTRES



Combinaisons intéressantes

- Certaines combinaisons peuvent être gagnantes
 - 👍 image-brute → RLE → flate → image-compressée
 - 👍 image-brute → JPEG-baseline → flate → image-compressée
- Quand certaines sont interdites
 - 👎 image-brute → JPEG-progressif → flate → ❌



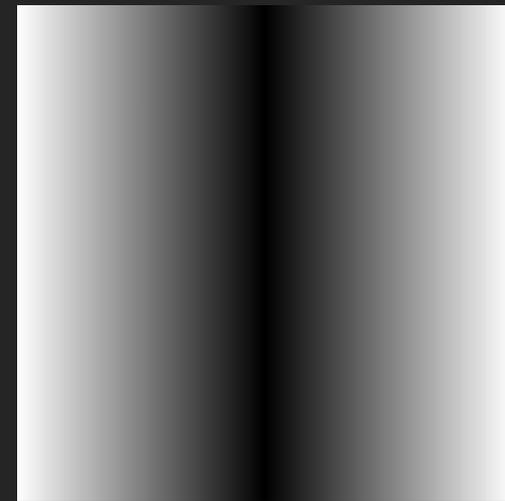
Des prédicteurs pour RLE ?

- Problèmes
 - Les prédicteurs sont des paramètres, pas des filtres
 - Ils sont réservés à LZW et Flate
 - Ils sont effectifs sur des données à 2 dimensions
- Solution
 - Flate dispose d'un mode « NoCompression »

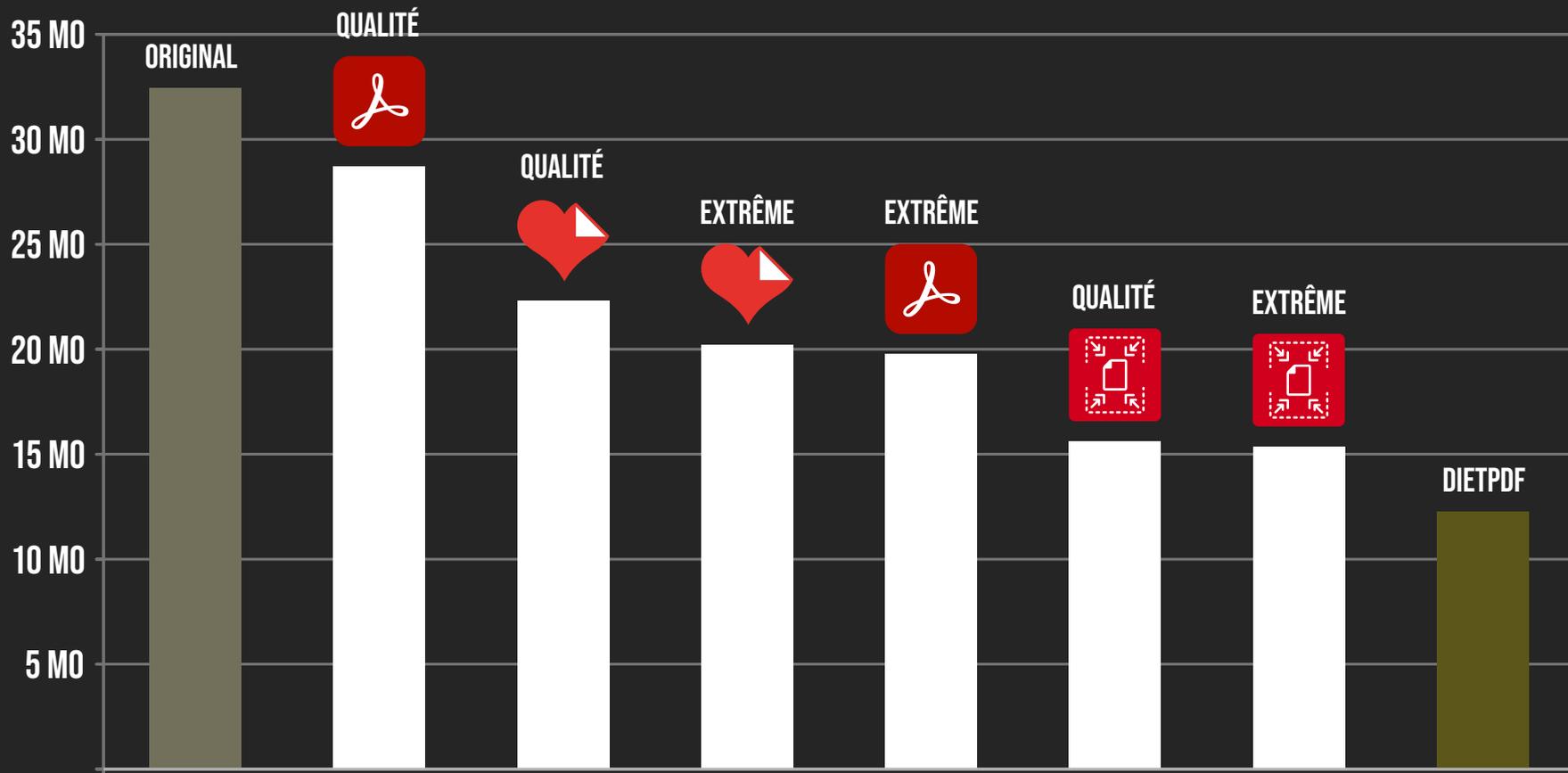
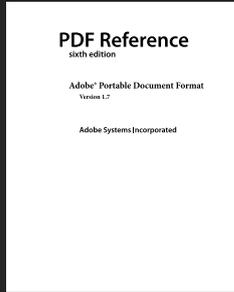


Exemple de prédicteurs avec RLE

- Image test : 512×512 pixels, dégradé blanc → noir → blanc
- Comparatif des poids
 - Brut = 262 144 octets
 - Prédicteur + RLE = 5 703 octets
 - Flate = 1 765 octets
 - Prédicteur + Flate = 877 octets
 - Prédicteur + RLE + Flate = 134 octets
 - *WebP* = 654 octets
 - *JPEG 2000* = 513 octets
 - *JPEG XL* = 414 octets



Et DietPDF ?



PDF Reference 1.7, 32 Mo

MERCI!

- Merci aux organisateurs et à Virginie Férey
- DietPDF : <https://github.com/Zigazou/dietpdf-haskell>
- Moi sur les internets
 - Github : <https://github.com/zigazou>
 - Twitter : [@zigazou](https://twitter.com/zigazou)
 - Mail : zigazou@protonmail.com