# Guide Méthodologique

# Les outils de conversion vers le format PDF : traitement de texte, dessins techniques, édition scientifique

**Version**: **1.0**
**Date: 14/01/2014**
**Document: NUMEN-SIAF-HUMANUM-CINES-GM-OCPDF-1.0**
**Confidentialité**: **Public**

**Guide méthodologique**
**Les outils de conversion**
**vers le format PDF**

# Table des Matières

# Introduction

L'archivage de fichiers numériques requiert l'utilisation de formats de fichiers pérennes. Le format PDF est un des formats les plus répandus. Le SIAF (Service Interministériel des Archives de France) et la TGIR HumaNum (UMS CNRS 3598, anciennement le TGE Adonis) ont souhaité initier une étude sur ce format afin de conseiller les utilisateurs qui souhaiteraient utiliser ce format. Mais, pour un néophyte, le domaine seul du PDF est complexe à comprendre dans toutes ses spécificités et ses nuances.

Cette étude comprend l'étude théorique du format PDF mais aussi des tests de logiciels de conversion au format PDF et de validateurs du format PDF. Après avoir défini dans une première étape ce qu'est le format PDF[1], la deuxième étape consiste à étudier différents logiciels qui peuvent créer des fichiers PDF.

Il est difficile de créer directement des documents dans le format PDF. Toutefois, l'utilisation du logiciel Adobe Acrobat Pro permet de créer un fichier PDF natif mais son utilisation n'est pas pratique pour des documents volumineux et complexes. La méthode la plus courante est de créer un fichier source avec un éditeur (ou un logiciel spécifique adapté à son activité). Mais dans une perspective d'archivage, le format de fichier alors créé peut s'avérer être très dépendant du logiciel de création, ce qui n'offre que peu de garantie pour une conservation à long terme. C'est pourquoi, on a l'habitude de recourir ensuite à une conversion du format de fichier initial au format PDF.

## 1.1 Périmètre de l'étude

Cette deuxième partie de l'étude sur le format PDF se concentre sur les logiciels de conversion existants pour créer du PDF. Les objectifs de cette étude sont :

- présenter les principes de conversion existants actuellement dans les outils du marché ;
- présenter les choix effectués en termes
  a) de formats de fichiers avant conversion au format PDF
  b) de logiciels de conversion ;
- présenter la méthodologie de test des logiciels sélectionnés ;
- évaluer la qualité des outils sélectionnés (en particulier par le critère de fidélité des informations présentes dans les fichiers PDF créés par rapport à celles du document originel).

Face à la quantité de formats de fichiers source et d'outils de conversion existants, le périmètre de l'étude a volontairement été limité à plusieurs niveaux :

- **Les types de fichiers testés :**

Une large partie de l'étude a été consacrée au test de logiciels bureautiques de type traitement de texte et des outils de conversions associés, car ils produisent les formats de fichiers les plus répandus.
Le format de fichier TeX a également été étudié parce qu'il est très utilisé dans la communauté Enseignement Supérieur et Recherche, notamment pour la production des thèses. De plus, dans le cadre de l'archivage au CINES des thèses de doctorat soutenues en France, il a été constaté que de nombreux fichiers PDF générés à partir de fichiers sources en LateX sont rejetés au moment de la validation du format.
Enfin, la dernière partie de cette étude est consacrée à un format de fichiers pivot pour la création de plans, le DWG produit par le logiciel de Création Assistée par Ordinateur (CAO) Autocad.

[1] Voir le document « *Guide Méthodologique : le format de fichiers PDF* », accessible en ligne : http://www.archivesdefrance.culture.gouv.fr/static/6189

- **Les logiciels testés :**

Les outils de conversion ont été sélectionnés en fonction de leur utilisation par la communauté scientifique et dans le cadre des archives publiques, en privilégiant ceux qui étaient directement intégrés dans les logiciels de création des fichiers sources.

- **Les versions de PDF générées en sortie :**

Toutes les versions de PDF n'ont pas été testées. Dans la mesure du possible, les tests ont ciblé les dernières versions de PDF et de PDF/A disponibles dans l'outil de conversion testé.
La méthodologie suivie pour la réalisation de l'ensemble des tests est précisée dans cette étude afin de permettre à chacun de poursuivre l'expérimentation en fonction de ses propres besoins.

Cette étude s'adresse principalement aux producteurs d'archives ou à ceux qui ont la charge de leur conservation :

- pour les aider à choisir un format cible d'archivage parmi la famille des formats PDF, selon les fonctionnalités mises en œuvre dans le fichier original et qu'ils souhaitent préserver ;
- pour les guider dans le choix d'un outil de conversion afin qu'ils puissent produire ce format cible avec une qualité correcte à partir de leurs autres formats source.

## 1.2 Références

| Nom du Document | Version | Localisation du Document |
|---|---|---|
| Guide Méthodologique – le format de fichier PDF | 1.0 | http://www.archivesdefrance.culture.gouv.fr/static/6189 |
| Guide pratique du Pdf-A | 2 | Raymond Schiano – IRSTEA, juin 2013 |
| Rapport de synthèse :<br>Etude des outils de conversion en PDF/A | 1.0 | Ministère de la Défense – projet ARCHIPEL, février 2013 |
| PDF/A in a Nutshell<br>(Long Term Archiving with PDF) | 1.b | Olaf Drümmer, Alexandra Oettler, Dietrich von Seggern<br>http://www.pdfa.org/wp-content/uploads/2011/08/PDFA-in-a-Nutshell_1b.pdf |
| PDF/A in a Nutshell<br>(PDF for a long-term archiving) | 2.0 | Alexandra Oettler<br>http://www.pdfa.org/2013/04/pdfa-in-a-nutshell-2_0/ |
| PDF/A from Wikipedia | | http://en.wikipedia.org/wiki/PDF/A |
| | | |

# Types d'outils de conversion

Quatre types de convertisseur ont été identifiés. Les possibilités de conversion de chaque outil vont largement dépendre du type de convertisseur utilisé :

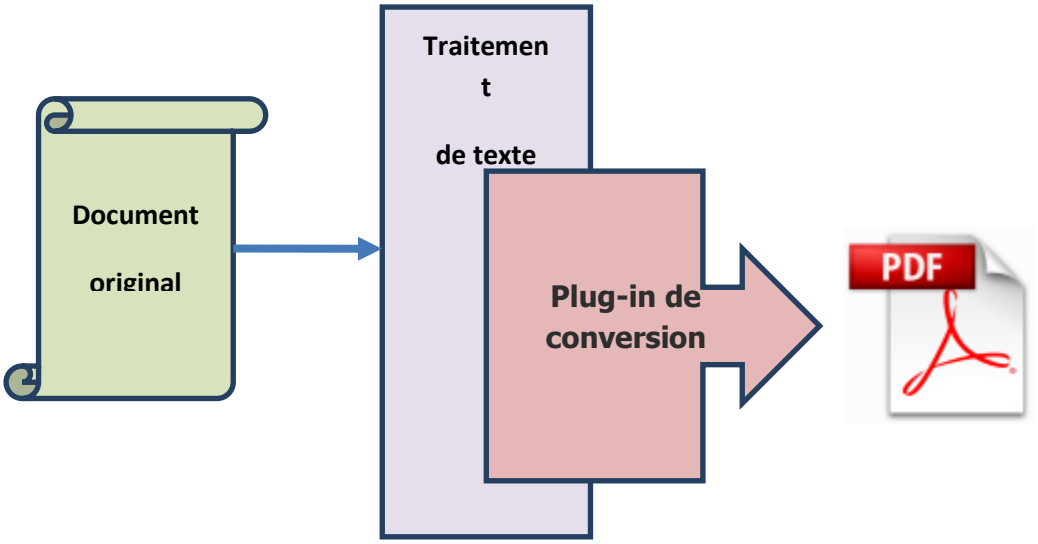
- plug-in du logiciel original
- logiciel capable de comprendre le format du fichier original et le convertir en PDF sans recours au logiciel utilisé pour le créer
- pilote d'impression qui interprète le flux d'impression pour en créer un fichier PDF
- logiciel qui comprend du PostScript et qui le convertit en PDF

Dans les sections ci-dessous, chacun de ces types de conversion est expliqué, ainsi que leurs avantages et inconvénients.

## Type 1 : Plug-in du logiciel original

En prenant l'exemple d'un traitement de texte, ce type de conversion fonctionnerait de la façon suivante :

**Figure 1 :** Conversion de fichier avec un plug-in dans le logiciel éditeur




Le document original (fichier du traitement de texte) est chargé dans le traitement de texte, et formaté par celui-ci. Une option de menu présente dans le traitement de texte (typiquement « Exporter comme PDF ») permet de lancer l'outil de conversion. Étant un plug-in, cet outil va avoir accès à une API interne du traitement de texte pour l'interroger sur le contenu du document et de sa mise en forme. Il utilisera ces informations pour créer le fichier PDF de sortie.

**Note :** Certains logiciels, comme *Microsoft Word* ou *Open Office* par exemple, ont déjà une telle option (typiquement « Enregistrer sous PDF » ou « Exporter comme PDF »). Ces options peuvent être considérées comme étant des convertisseurs de type 1.

**Avantages**

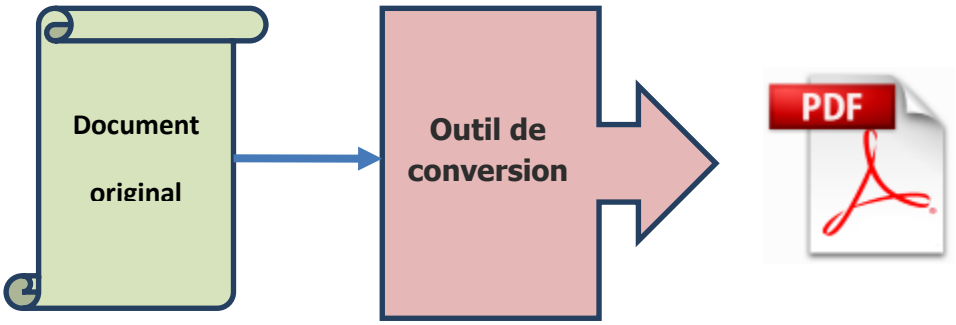
- Si l'API est assez puissante, l'outil a accès à toutes les informations du document ainsi que la façon dont le logiciel d'origine a formaté le document.
- Comme il s'agit d'un logiciel original, l'API est susceptible d'être plus complète qu'avec un outil tiers.

**Inconvénients**

- Il faut avoir accès au logiciel d'origine pour faire la conversion.

# Type 2 : Conversion depuis le fichier source

**Figure 2 :** Conversion de fichier avec un outil de conversion




Ce type d'outil de conversion est capable d'interpréter le format du fichier original et de le formater complètement pour en créer un fichier PDF de sortie.

**Avantages**

- Il n'y a pas besoin d'avoir accès au logiciel qui a été utilisé pour créer le fichier.
- Toutes les informations contenues dans le document original sont disponibles pour l'outil de conversion.

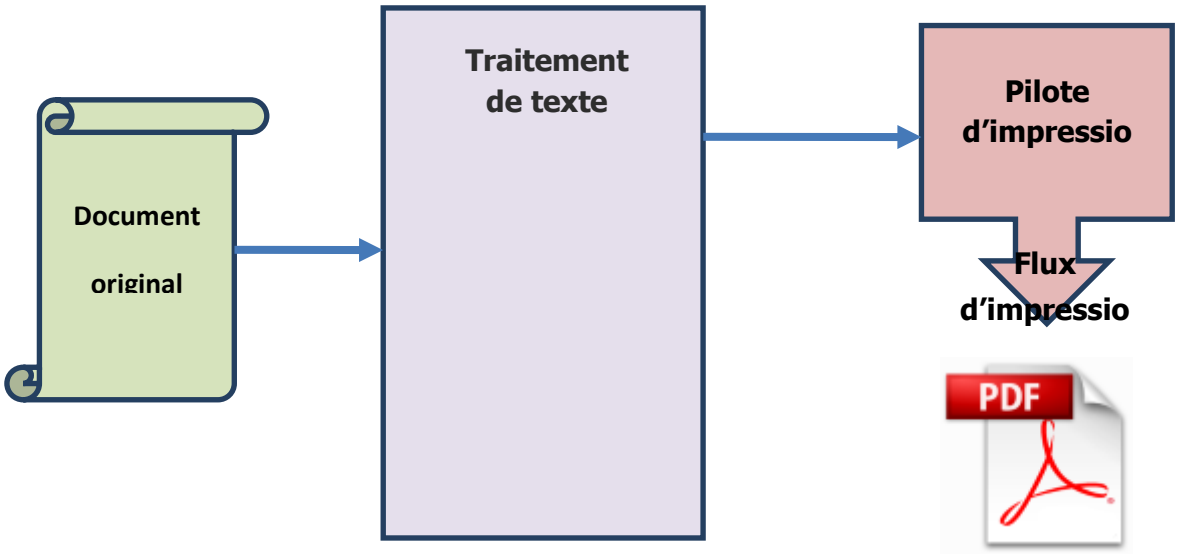
**Inconvénients**

- L'outil de conversion doit faire le travail de mise en forme. Puisque les algorithmes utilisés dans les logiciels sont rarement publics, la mise en forme peut être différente de celle faite par le logiciel original.
- Si la structure logique du document original n'est pas publiée, il est aussi possible que l'interprétation du fichier soit différente de celle de l'outil original.

# Type 3 : Pilote d'impression

En prenant l'exemple d'un traitement de texte, ce type de conversion fonctionnerait de la façon suivante :

**Figure 3 :** Conversion de fichier avec un pilote d'impression




L'impression depuis le logiciel utilisé pour créer le document se fait en choisissant un pilote spécifique à la création de PDF. Ce pilote interprète le flux d'impression (typiquement défini par le système d'exploitation) pour créer le fichier PDF.

**Avantages**

- La mise en forme est entièrement faite par le logiciel original. Le pilote applique seulement le texte et les images aux endroits indiqués par le flux d'impression. Le fichier PDF reflèterait donc bien la mise en forme du document original.

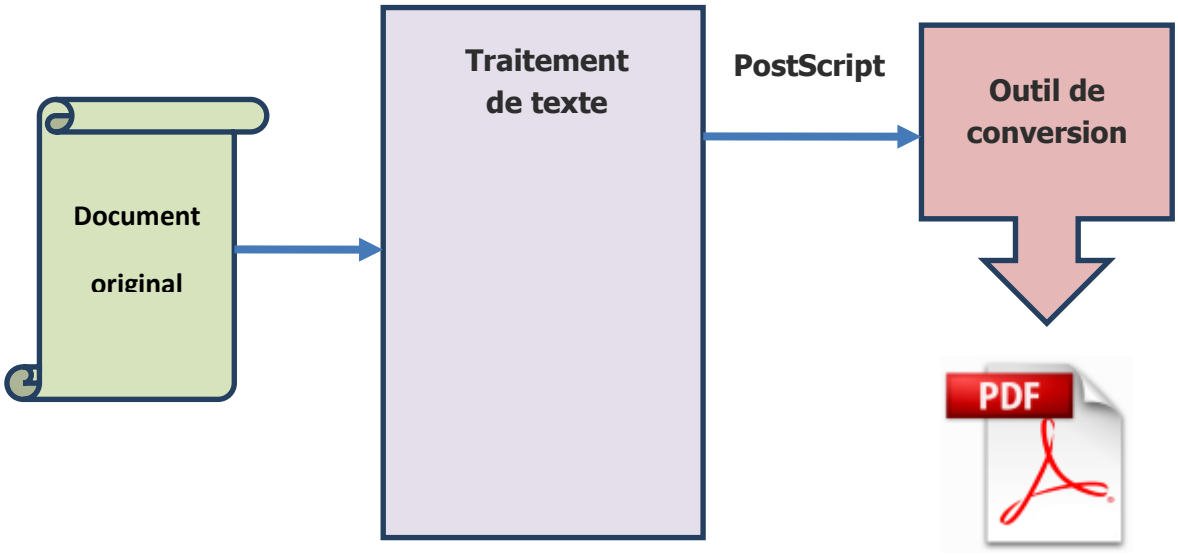
**Inconvénients**

- Le pilote n'a accès qu'à un flux d'impression. Il n'aura donc pas accès aux informations de structure ou aux métadonnées du document original.

# Type 4 : Convertisseur PostScript

Un convertisseur PostScript fonctionne exactement selon le même principe qu'un pilote d'impression, sauf que le flux d'impression est un flux PostScript.

**Figure 4 :** Conversion de fichier avec un convertisseur PostScript




Ce type de convertisseur est traité séparément car le PostScript peut contenir bien plus d'informations qu'un flux d'impression du système d'exploitation et il est donc possible de récupérer plus d'informations.

Ce qui est important dans ce schéma est la façon dont le PostScript a été créé. S'il est créé par un logiciel qui comprend ce langage, le PostScript peut être très riche. Par contre, s'il est créé par un pilote d'impression PostScript, il risque d'avoir les mêmes limites qu'un pilote d'impression PDF.

**Avantages**

- Le PostScript peut contenir plus d'informations qu'un flux d'impression du système d'exploitation et peut être créé pour indiquer des informations PDF directement.
- Le modèle de description de pages de PostScript est très proche de PDF, donc le rendu sera de bonne qualité.

**Inconvénients**

- La qualité du PDF dépendra de la qualité du PostScript créé.
- PostScript n'est pas capable de transmettre toutes les informations nécessaires à un fichier PDF.

# Critères de choix du type de convertisseur pour l'archivage

Ayant décrit les caractéristiques des différents types de convertisseurs, l'étude s'intéresse maintenant à la manière d'en choisir un pour les besoins de l'archivage. Les critères énoncés ci-dessous correspondent aux principales attentes que l'on peut avoir envers un fichier PDF à archiver.

## Mise en forme

Dans le cadre de l'archivage, il est crucial que la mise en forme du fichier PDF généré soit fidèle à celle du fichier original. Tous les types de convertisseur cherchent à atteindre cet objectif, mais les risques de transformation sont plus importants avec un convertisseur qui prend le fichier original en entrée sans utiliser le logiciel qui l'a créé (type 2). Il ne faut pas pour autant écarter ce type de convertisseurs, mais accorder une attention particulière à la qualité de la mise en forme.

## Texte Unicode

Il est également important de pouvoir récupérer le texte du document original depuis le fichier PDF. Les convertisseurs de type 1 et 2 sont capables de mettre le texte correctement dans le fichier PDF, ce qui n'est pas le cas des types 3 et 4.

Voici un exemple de problématique rencontrée avec un convertisseur qui utilise un flux d'impression :

**Texte original :**

« La condition mentionnée dans l'article n'est pas applicable aux logements. »

**Mise en forme :**

Dans ce cas, le logiciel qui fait la mise en forme décide de couper le mot « applicable ». Les instructions suivantes (a minima) sont transmises au flux d'impression :

1. Définir la position sur la page
2. Imprimer le texte « La condition mentionnée dans l'article n'est pas appli- »
3. Définir la position sur la page
4. Imprimer le texte « cable aux logements. »

Le texte qui est reçu par le convertisseur n'est donc pas le même que le texte original. Il existe un caractère de plus et les deux parties d'un mot (« applicable ») ont été séparées.

L'identification des fins de mots est un autre problème récurrent. Plusieurs logiciels positionnent chaque mot individuellement pour appliquer une justification horizontale ou des parties de mots pour appliquer le crénage. Les fins de mots ne sont donc pas mises explicitement dans le flux et il peut être très difficile d'identifier les mots.

Le texte cité ci-dessus pourrait être envoyé comme une séquence de chaînes positionnée individuellement ainsi :

« La » « co » « ndi » « tion » « me » « nti » « onn » « ée » etc.

Il devient donc impossible pour les convertisseurs de type 3 ou 4 de trouver et identifier les mots présents dans le document.

Adobe Acrobat interprète généralement de manière satisfaisante ce genre de cas afin de restituer le texte, mais il n'y a aucune garantie car il se base sur des dictionnaires pour chaque langue utilisée.

Dans le cadre de l'archivage, se baser sur ce genre d'algorithmes pour récupérer le texte peut représenter un risque notoire car il faut s'assurer préalablement d'avoir, grâce à ce procédé, la garantie d'obtenir le résultat escompté. Il est donc préférable d'utiliser des convertisseurs de type 1 ou 2.

## Structure du document

Il est également important de récupérer la structure du document. Par exemple, un convertisseur de bonne qualité doit être capable de baliser les paragraphes qui utilisent un style de titre dans un document Word, comme des titres dans le PDF, et les autres paragraphes comme des paragraphes.

Seuls les types 1 et 2 seront capables de garder cette structure. Les convertisseurs de type 3 basés sur un fichier intermédiaire d'impression n'auront pas accès à la structure et par conséquent ne pourront pas la mettre dans le fichier PDF.

Les convertisseurs de type 4 pourraient avoir accès à la structure si le fichier PostScript a été créé spécifiquement pour une conversion PDF (avec l'utilisation de l'extension PostScript appelé pdfmark) mais ce cas d'utilisation reste très rare.

## Conclusion

Du point de vue théorique, le convertisseur de « type 1 » semble plus fiable car la conversion se fait sans intermédiaire et donc sans risque de perte d'informations. Cela dépend évidemment aussi de la qualité de l'API du logiciel effectuant la conversion.

De même, le convertisseur de « type 2 » semble aussi fiable, à condition que la mise en forme reste conforme à l'original, après conversion.

Les convertisseurs de « type 3 » et de « type 4 » sont moins performants car ils utilisent un fichier intermédiaire avec un format différent.

Aussi, le lecteur notera que les convertisseurs de type 1 et 2 ont été privilégiés dans le cadre de cette étude pour la génération de fichiers au format PDF.

# Tests des fichiers bureautiques de type « traitement de texte »

## 3.1 Définition de la stratégie de tests

### Les outils testés

Il est très difficile de choisir les logiciels à tester, mais la grande quantité de produits disponibles sur le marché oblige à effectuer une sélection.

Conformément au résultat de la comparaison entre les différents types de convertisseur (cf. chapitre précédent), les outils retenus appartiennent majoritairement au « type 1 » et au « type 2 », et sont tous des logiciels exploitables sous Windows. Au vu des arguments présentés précédemment, les pilotes d'impression (type 3) et les convertisseurs PostScript (type 4) n'ont pas été retenus. Toutefois, l'outil PDF Creator étant largement utilisé, il a été étudié à des fins de comparaison.

Tous les logiciels testés ont été retenus parce qu'ils étaient largement utilisés. Un autre critère de sélection a été aussi de mélanger des logiciels payants (Word de MicroSoft-Office, Adobe Acrobat Pro, Callas pdfaPilot) et open-source et gratuits (writer de LibreOffice et OpenOffice, PDFCreator).

Pour ces logiciels, seule la version la plus récente a été retenue afin de tester les possibilités de conversion de ces outils et de disposer notamment des dernières versions de PDF. Par exemple, le choix de tester Word 2003 n'est pas pertinent puisque le logiciel de conversion qu'il contient n'est pas capable de produire du Pdf/A (sorti en 2005). Une seule exception a été faite en retenant la version 2010 de Word de Microsoft-Office et non 2013, car celle-ci a semblé trop récente pour être largement utilisée. Elle pourra faire l'objet de tests futurs.

Les documents issus de scans ou d'OCRs ont également été volontairement écartés car ils soulèvent d'autres problématiques que celles associées à la conversion en PDF. Enfin, nous n'avons pas effectué de tests avec une application développée spécifiquement et utilisant une API pour la conversion PDF.

Il est à noter aussi le choix des logiciels d'Adobe et de Callas car ce sont deux fournisseurs qui sont « représentés » dans les instances de normalisation ISO de PDF.

| Nom | Version | Editeur du logiciel | Type d'outil |
|---|---|---|---|
| Word | 2010 | Microsoft | Type 1 |
| Writer | 3.6.5.2 | LibreOffice | Type 1 |
| Writer | 3.4.1 | OpenOffice | Type 1 |
| Acrobat Pro | 11 | Adobe | Type 1 |
| PdfaPilot | 4.0 | Callas | Type 2 |
| PDF Creator | 1.7.1 | Pdfforge | Type 3 |

**Tableau 1 :** Les logiciels testés dans cette étude

## Choix des formats en entrée

Afin d'orienter l'étude vers la conversion de document plutôt que vers la création, les formats de fichiers sélectionnés en entrée sont ceux directement produits par le logiciel d'origine (fichier au format « odt » pour le logiciel Writer-LibreOffice par exemple).

Seuls les formats de type « traitement de texte » ont été retenus afin de limiter l'étude. Cela correspond aux extensions de fichier : « .docx » et « .odt ».

Les fichiers au format « .doc », acceptés également par les dernières versions de Word, n'ont pas été retenus car lorsqu'un traitement de texte ouvre un fichier .doc ou .docx les deux sont convertis en mémoire dans un format binaire identique qui permet au logiciel de travailler et d'imprimer à partir de ce format. Le format du fichier en entrée n'est donc pas important.

Dans le cadre de cette étude, les fichiers de type « tableur » (Excel ou Classeur) et de type « présentation » (PowerPoint ou Présentation) ont été volontairement exclus dans un premier temps, car les logiciels qui les créent sont intégrés dans des suites bureautiques qui utilisent le même outil de conversion. Il serait toutefois intéressant de vérifier dans un second temps si les résultats obtenus en testant des fichiers de ces types sont identiques au résultat de cette étude.

Enfin, la qualité de la conversion en PDF des images a été testée au travers des fichiers bureautiques puisque plusieurs types d'images ont été insérés dans les fichiers de test (cf. partie Protocole de tests utilisé).

## Choix des formats en sortie

A titre de rappel, le schéma[2] ci-dessous présente l'articulation entre les différentes versions des formats PDF et PDF/A :

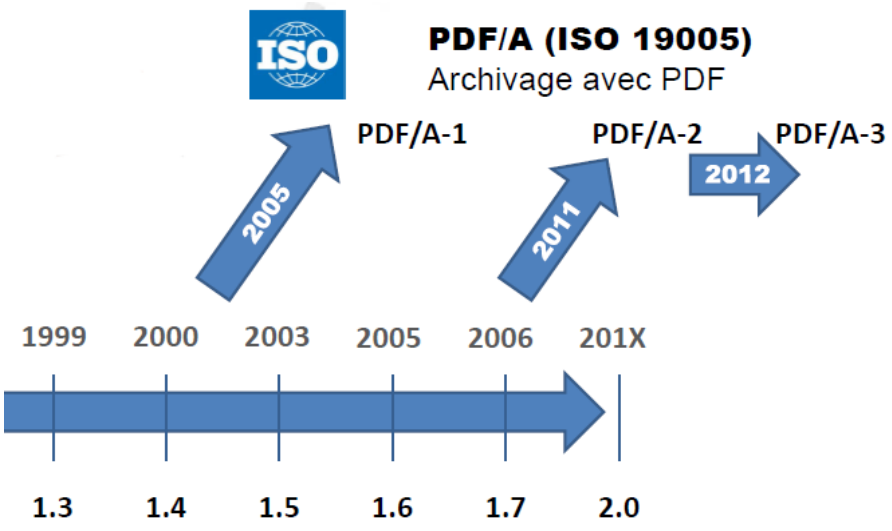



**Figure 5 :** Les différentes versions de PDF/A

[2] Source : Nick Parker « Etude sur le format de fichier PDF »
http://pin.association-aristote.fr/doku.php/public/reunion_pleniere/cr/cr_20130104

De manière générale, la stratégie retenue pour les tests a consisté à privilégier les versions de PDF suivantes lorsqu'elles étaient disponibles dans le logiciel testé :

- PDF 1.4 : car elle a servi de modèle pour la première version de PDF/A ;
- PDF 1.7 : qui est actuellement la version la plus aboutie au moment de la rédaction du présent rapport et qui a servi de base aux PDF/A-2 et PDF/A-3. ;
- PDF/A : qui est la référence en termes d'archivage. Parmi les versions 1, 2 et 3, le choix s'est porté vers le format le plus récent (PDF/A-3 lorsque cela était possible, sinon PDF/A-2 sinon PDF/A-1). Ensuite pour chaque version, il y a une lettre « **a** », « **u** » ou « **b** » pour définir trois niveaux de contrôle (« **a** » pour **a**vancé, « **u** » pour **u**nicode et « **b** » pour **b**asique). Nous avons choisi, suivant les possibilités du logiciel, le niveau de contrôle le plus exigeant : PDF/A-3a (ou PDF/A-2a ou PDF/A-1a), sinon PDF/A-(1u, 2u, 3u) ou PDF/A-(1b, 2b, 3b).

## Protocole de tests utilisé

Les tests ont été réalisés avec un fichier de test identique[3] qui a permis de vérifier la qualité de la conversion en PDF d'un certain nombre de propriétés, telles que :

- les images, vectorielles et matricielles (bitmap), en basse et haute définition, avec et sans transparence, et lorsque le logiciel le permettait, l'insertion d'une image PDF ;
- les couleurs ;
- les polices de caractère ;
- la structure logique avec l'utilisation des titres ;
- la transparence des objets ;
- les liens internes et externes ;
- les métadonnées.

Chaque scénario de test a fait l'objet d'une fiche qui identifie en entête la version et le logiciel ayant servi à la création du document à convertir, ainsi que le format du fichier converti et qui reprend ensuite l'ensemble des fonctionnalités testées.

L'ensemble des fiches de tests produites est jointe à cette étude dans un document[3].

*Remarque* : Pour produire des fichiers PDF identiques aux scénarii de tests, il faut se reporter aux fiches qui précisent la procédure utilisée dans le logiciel (paramétrage du logiciel…).

## Tableau synthétique des cas testés

| Logiciels utilisés | Versions de PDF testées dans cette étude | Versions de PDF supportées par le logiciel testé |
|---|---|---|
| **Writer LibreOffice 3.6.5.2** | PDF 1.4 (choix par défaut) | PDF par défaut (PDF 1.4) |
| | PDF/A-1a | PDF/A-1a |
| **Writer OpenOffice 3.4.1** | PDF 1.4 (défaut) | PDF par défaut (PDF 1.4) |
| | PDF/A-1a | PDF/A-1a |

[3] Ces fichiers sont disponibles en ligne :
https://alfresco.cines.fr/alfresco/faces/jsp/browse/browse.jsp;jsessionid=ED97671B9869BA13E6F24CE7D1921396
Connexion préalable : login=guest ; mdp=guest

| | | |
|---|---|---|
| **Word 2010** | PDF 1.5 (défaut) | PDF par défaut (PDF 1.5) |
| | PDF/A-1a | PDF/A-1a |
| **Acrobat Pro 11** | PDF 1.4 (défaut) | PDF 1.3; PDF 1.4; PDF 1.5; |
| | PDF 1.7 | PDF 1.6; PDF 1.7 |
| | PDF/A-3a | PDF/A-1a; PDF/A-1b;<br>PDF/A-2a; PDF/A-2b; PDF/A-2u;<br>PDF/A-3a; PDF/A-3b; PDF/A-3u |
| **pdfaPilot 4.0** | PDF 1.5 (défaut) | PDF 1.2; PDF 1.3; PDF 1.4; PDF 1.5; PDF 1.6 |
| | PDF/A-3a | PDF/A-1a; PDF/A-1b;<br>PDF/A-2a; PDF/A-2b; PDF/A-2u;<br>PDF/A-3a; PDF/A-3b; PDF/A-3u |
| **PDF Creator 1.7.1** | PDF 1.4 | PDF 1.2; PDF 1.3; PDF 1.4; PDF 1.5 |
| | PDF/A-2a | PDF/A-2b annoncé mais problème à réaliser<br>en réalité PDF/A-1b<br>(mais ne fonctionne pas sous Win7 – 64 bits) |

**Tableau 2 :** Les logiciels et versions de PDF testés

*Remarque* : le logiciel PDF Creator propose une conversion en PDF/A-2b, ce qui semble impossible puisqu'il ne propose pas de conversion en PDF 1.7 (le PDF/A-2 est à base de PDF 1.7). Par ailleurs, cette conversion n'a pas pu être effectuée lors des tests sous un environnement Windows 7 - 64 bits, car le fichier de tests utilisé comporte des parties qui perturbent le fonctionnement du logiciel et l'empêchent de terminer son traitement.

## Pré-requis pour les tests

Pour la réalisation des tests et l'analyse des résultats obtenus, les outils suivants ont été utilisés :

- **Adobe Acrobat Pro v11 :**

Cet outil est une référence pour le traitement de fichier PDF et permet d'afficher les fichiers de façon correcte. Il a été utilisé dans le cadre de l'étude comme outil de référence pour la validation du fichier PDF créé. Adobe est le créateur du format PDF et il est représenté dans les instances internationales ISO pour définir le PDF/A (le lecteur notera que nous aurions également pu prendre pdfaPilot de Callas qui participe à la normalisation du PDF/A).

- **Enfocus Pitstop Pro[4] :**

Cet outil est une extension d'Adobe Acrobat Pro qui permet d'afficher les propriétés de tous les objets d'une page PDF. Il permet d'analyser plus précisément les fichiers PDF créés et leur contenu.

- **Un éditeur d'images :**

L'utilisation d'un éditeur d'images (tels que « The GIMP » ou « Adobe Photoshop ») permet de créer et contrôler les images de test.

[4] http://www.enfocus.com/product.php?id=855

- **« Bloc-notes » de Windows :** pour vérifier si le « copier/coller » des textes est en Unicode.

# Définition des fonctionnalités testées

Comme expliqué précédemment, un outil de conversion qui génère un fichier PDF doit pouvoir créer un PDF de qualité, dont les informations sont identiques au document source et qu'il soit possible de récupérer le contenu textuel et les images pour pouvoir les réutiliser.

Les tests vont donc se concentrer sur les aspects suivants :

- contenu textuel — le texte du PDF peut être extrait ;
- structure logique du document — la structure du document originel (s'il en a) est gardée dans le fichier PDF ;
- images — les images du document original sont conservées sans aucune dégradation ;
- mise en forme — la mise en forme du document original est conservée ;

La méthodologie retenue pour les tests consiste donc à évaluer les outils de conversion sélectionnés d'après les fonctionnalités qu'offre le format PDF. Le contenu qui suit présente les fonctionnalités testées et la manière de valider les résultats. Pour plus d'explications sur ces fonctionnalités, se reporter au « *Guide méthodologique – le format de fichier PDF* ».

Dans certains contextes, le terme de « fonctionnalité » est parfois remplacé par l'expression « propriétés remarquables » (*significant properties*) afin de mettre davantage l'accent sur la notion d'identification de l'information essentielle à conserver contenue dans le document.

## Tableau récapitulatif des fonctionnalités

Pour les tests de l'environnement bureautique, un certain nombre de fonctionnalités a été retenu. Le tableau ci-dessous rappelle les principales fonctionnalités qui font la spécificité des différentes versions de PDF.

**Légende :**

| | |
|---|---|
| | Aucune fonctionnalité |
| | Fonctionnalité présente |
| | Amélioration significative de la fonctionnalité |

On a placé les abréviations des principales améliorations dans les cellules du tableau. Le lecteur est invité à consulter le « *Guide méthodologique – le format de fichier PDF* » pour avoir une liste plus détaillée des améliorations.

| **Fonctionnalités** | **PDF 1.0** | **PDF 1.1** | **PDF 1.2** | **PDF 1.3** | **PDF 1.4** | **PDF 1.5** | **PDF 1.6** | **PDF 1.7** | **PDF 2.0[5]** |
|---|---|---|---|---|---|---|---|---|---|
| Couleur | | Cal | Sép | ICC | | | | | |
| Polices | | | | | | | OT | | |
| Transparence | | | | | | | | | |
| Préférences de document | | | | | | | | | |
| Métadonnées | Dict | | | | XMP | | | | |
| Sommaire | | | | | | | | | |

[5] Le PDF 2.0 n'étant pas encore publié à l'ISO au moment de la rédaction de cette étude, cette liste est susceptible de changer.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Signatures numériques | | | | | | | | | |
| Articles | | | | | | | | | |
| Présentations | | | | | Alt. | | | | |
| Structure logique de document | | | | | | | | | |
| Contenu facultatif | | | | | | | | | |
| Annotations | | | | | | | | | |
| Liens hypertexte | Int | Ext | | | | | | | |
| Compression | | | Flate | | JB2 | JP2 | | | |
| Sécurité | | 40 | | | 128 | | | | 256 |
| Formulaires interactifs « AcroForms » | | | | | | | | | |
| Formulaires interactifs XFA | | | | | | | | | |
| Multimédia | | | Aud | | | Vid | | | |
| Illustrations 3D | | | | | | | U3D | | PRC |
| Programmation | | | | JS | | | | | |
| Fichiers embarqués | | | | | | | | | |
| OPI *(*Open Prepress Interface*)* | | | | | | | | | |
| CAO *(*Conception Assistée par Ordinateur*)* | | | | | | | | | |
| Informations géomatiques | | | | | | | | | |
| Capture de pages web en PDF | | | | | | | | | |

**Abréviations**
**Alt**=alternatives, **Aud**=audio, **Cal**=calibré, cryptage de **40**, **128**, **256** bits ; **Dict**=dictionnaire, **Ext**=externe ; **Flate** compression ; **ICC**=International Color Consortium ; **Int**=interne, **JB2**=JBIG2, **JP2**=JPEG2000, **JS**=JavaScript , **OT**=OpenType, PCR=les fichiers PCR peuvent contenir des données 3D encapsulées dans un fichier PDF ; **Sép**=séparations, **U3D**=Universal 3D ; **Vid**=vidéo ; **XMP**=eXtensible Metatdata Platform ;

**Tableau 3 :** Les fonctionnalités des différentes versions PDF

## Vérification du format du fichier converti

La version PDF déclarée par le fichier a été vérifiée dans le cadre de la présente partie de l'étude (« outils de conversion ») au moyen de l'outil Adobe Acrobat Pro version 11. Les tests spécifiques à la structure du format du fichier seront réalisés dans le cadre de la troisième partie de l'étude sur le format PDF (partie « outils de validation »).

Pour déterminer la version d'un fichier PDF, la procédure de test utilisée recherche la version dans le logiciel Adobe Acrobat Pro (Fichier->Propriétés) :

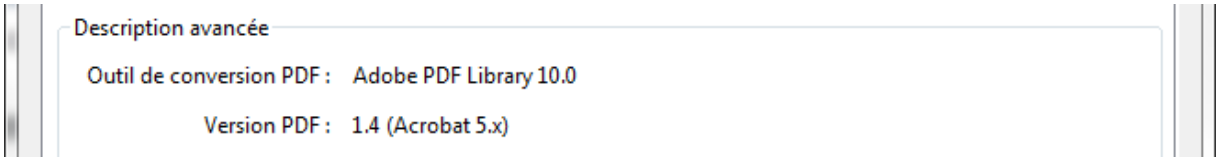



**Figure 6 :** Copie d'écran indiquant la version du PDF dans Adobe Acrobat Pro

Il n'y aura pas d'analyse complète sur la conformité PDF/A à ce stade. Seule la présence du bandeau bleu PDF/A est constatée lors de l'ouverture du fichier PDF dans Acrobat pour déterminer si le fichier se déclare comme étant du PDF/A :

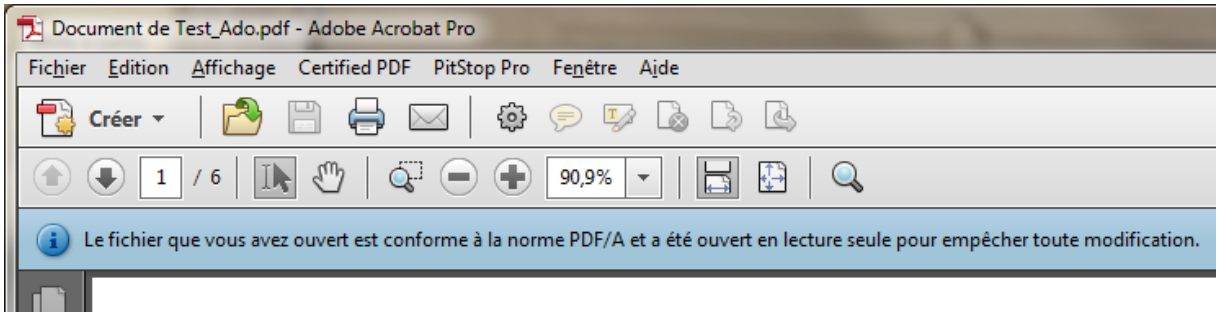



**Figure 7 :** Copie d'écran du bandeau bleu PDF/A

Le lecteur notera que l'apparition de ce bandeau requiert un paramétrage spécifique préalable dans Adobe Acrobat Pro (cf. dans les préférences Édition->Préférences->Général, sous l'onglet « Documents » il faut indiquer le mode d'affichage PDF/A pour ces documents) :

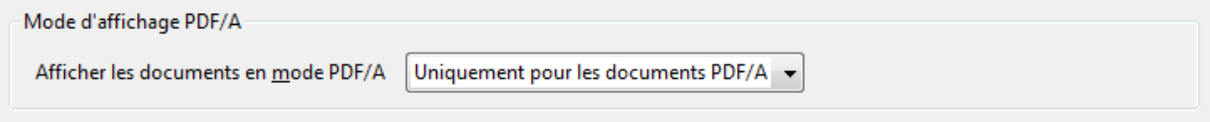



**Figure 8** : Copie d'écran du paramétrage dans Adobe Acrobat Pro

Les fiches détaillant les résultats des tests réalisés indiquent la version de PDF ou norme ISO qui a été demandée lors de la conversion (ou un tiret s'il n'était pas possible de préciser une version) ainsi que la version ou norme ISO signalée par le fichier PDF :

| | **Demandée** | **Constatée** |
|---|---|---|
| Version de PDF | | |
| Norme ISO | | |

## Modélisation de l'image

Pour l'archivage, comme pour toute autre utilisation, il est important que le rendu de la page soit identique au document original.

Pour ce test, des documents variés du type correspondant aux logiciels sources ont été retenus et ont été passés dans le convertisseur. Les résultats vus dans Adobe Acrobat ont été ensuite comparés avec le rendu dans le logiciel source.

Quelque soit le type d'image incluse, un premier test a donc consisté à regarder visuellement les pages PDF pour déceler d'éventuelles anomalies par rapport au document source. La rubrique « RENDU » des fiches constituant l'annexe 1 indique les anomalies constatées.

### Tests sur images matricielles incluses :

Les formats PNG et JPEG ont été privilégiés pour les tests car ils sont normalisés. Plusieurs types d'images ont été testés :

- image matricielle sans perte (PNG) sans transparence basse résolution
- image matricielle sans perte (PNG) sans transparence haute résolution
- image matricielle sans perte (PNG) avec transparence basse résolution
- image matricielle sans perte (PNG) avec transparence haute résolution
- image matricielle avec perte (JPEG) basse résolution
- image matricielle avec perte (JPEG) haute résolution

**Rendu à l'œil nu**

Pour ce test, il suffit de regarder l'aspect général du fichier obtenu après conversion. Est-ce que, sans zoom, l'image semble être correcte ?

Dans le formulaire des résultats : 1. correspond à un rendu correct, 2. à une dégradation visible de l'image.

| | | |
|---|---|---|
| Rendu à l'œil nu | 1. le rendu est correct<br>2. l'image a perdu en qualité | |

**Rendu en zoom**

La deuxième partie du test consiste à zoomer sur les détails de l'image (image dans le logiciel original et fichier PDF) et comparer les deux pour voir s'il y a des dégradations.

Voici des exemples de dégradations :

| Image source à 400% de zoom : | Conversion 1 PDF à 400% de zoom : | Conversion 2 PDF à 400% de zoom : |
|---|---|---|
|  |  |  |

Dans ce cas, l'image était un PNG avec transparence. Dans le troisième dessin, il est clair que le convertisseur PDF n'a pas su interpréter les pixels avec des taux de transparence et a tout converti en blanc lorsque le pixel n'était pas opaque. Le résultat est une forme avec une bordure irrégulière.

| Image source à 400% de zoom : | Conversion 1 PDF à 400% de zoom : | Conversion 2 PDF à 400% de zoom : |
|---|---|---|
|  |  |  |

Dans ce cas, il apparaît que les deux conversions ont subi un sous-échantillonnage, mais la conversion 2, bien qu'ayant la même définition d'image, a subi une forte dégradation dans la qualité de la couleur.
Ce genre d'effet est typique des compressions à perte où une couleur uniforme devient un mélange de pixels de couleurs légèrement différentes, surtout autour des bordures des objets.

Dans le formulaire des résultats : 1. correspond à un rendu correct, 2. à une perte évidente de qualité.

| | | |
|---|---|---|
| Rendu en zoom | 1. le rendu est correct<br>2. l'image a perdu en qualité | |

**Définition de l'image**

La définition des images mises dans le fichier de test est connue avant son insertion dans le document (cela correspond à sa définition en pixels). Les tests visuels auront probablement signalé dans un premier temps des problèmes de définition. Ce test permet de voir précisément les changements opérés sur la définition de l'image qui ont éventuellement été constatés lors de l'observation visuelle préalable.

Pour faire ce test, l'Inspecteur de PitStop est utilisé pour afficher la définition de l'image en pixels.

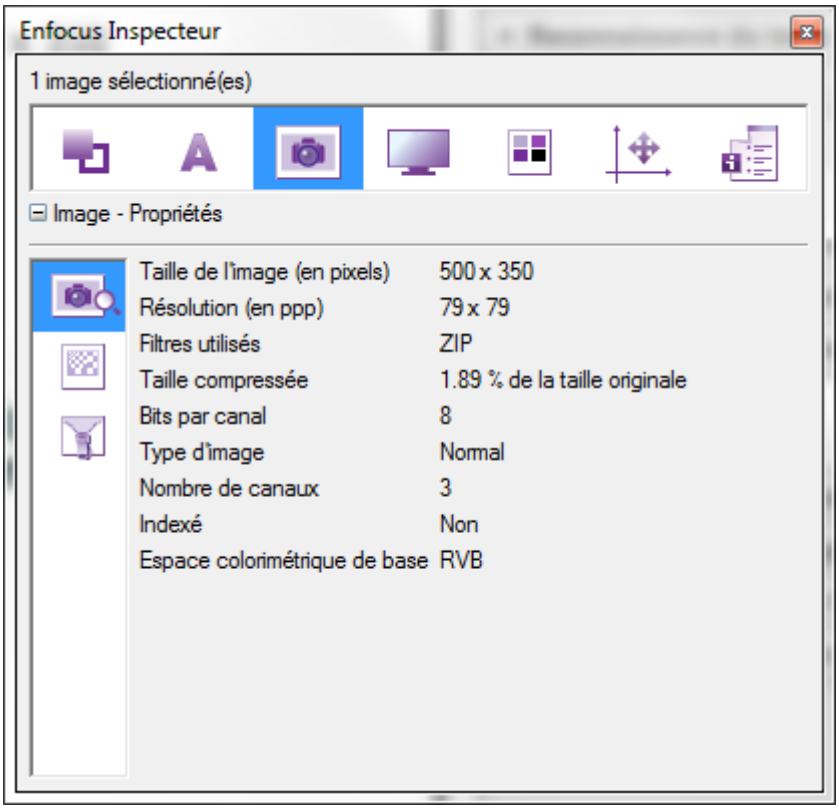



**Figure 9 :** Copie d'écran de l'inspecteur de PitStop

Dans cet exemple, la définition d'un fichier PNG à 500x350 a été conservée. Rien n'a donc été changé lors de la conversion en PDF. L'image pourrait donc être récupérée du fichier PDF sans problème.

Dans le formulaire des résultats : 1. correspond à une définition identique, 2. à une définition augmentée, 3. à une définition diminuée.

| | | |
|---|---|---|
| Définition de l'image | 1. la définition reste identique<br>2. la définition est augmentée<br>3. la définition est diminuée | |

**Compression de l'image**

Il existe deux types de compressions : une compression sans perte (la qualité de l'image reste identique) ou une compression avec perte (la qualité de l'image est dégradée de façon irrécupérable). La compression avec perte la plus connue est le JPEG.

Chaque fois qu'un fichier est réenregistré en JPEG, il y a perte de qualité.

Par ailleurs, le JPEG étant plutôt adapté aux photos, toute utilisation de JPEG sur une image vectorielle sera certainement néfaste sur sa qualité.

Pour ce test, l'inspecteur PitStop est utilisé pour connaître le type de compression appliqué à une image.

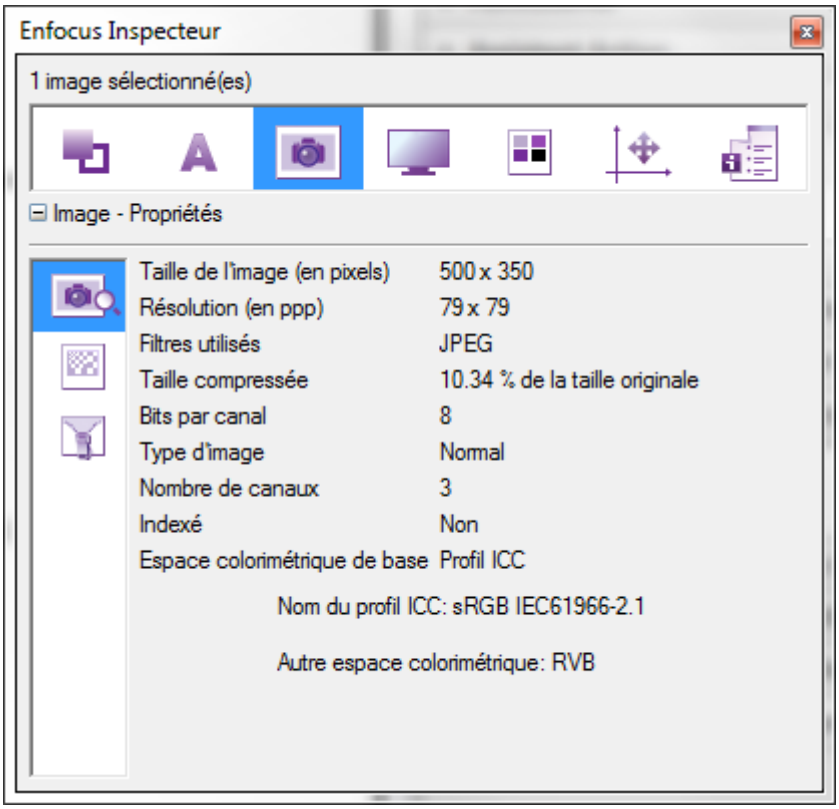



**Figure 10 :** Copie d'écran de l'inspecteur de PitStop

Dans cet exemple, l'image originellement en PNG a été compressée en JPEG, ce qui n'est pas conseillé pour l'archivage d'images vectorielles.
Dans l'exemple précédent (cf. Figure 9), l'image avait été compressée en ZIP qui est sans perte et donc préférable pour l'archivage.

Voici un exemple de perte de qualité avec JPEG :

Une image compressée en ZIP garde parfaitement les bords des objets rectangulaires (zoom 600%) :

Une image PNG compressée en JPG perd la netteté sur les bords des objets rectangulaires (zoom 600%) :

Dans le formulaire de résultats : 1. correspond à une compression sans perte, 2. à aucune compression, 3. à une compression avec perte.

| | | |
|---|---|---|
| Compression | 1. compression sans perte | |

| | | |
|---|---|---|
| | 2. aucune compression<br>3. compression avec perte | |

Tests sur images vectorielles incluses :

La problématique de contrôle de la qualité des images vectorielles est différente des images matricielles. Il ne doit pas y avoir de changement dans l'image puisque les formes en vecteur doivent être reproduites à l'identique dans le fichier PDF.

Trois types d'images vectorielles ont été fournis pour les tests :

- image vectorielle EMF ;
- image vectorielle EPSF ;
- image vectorielle PDF (lorsque le logiciel éditeur de texte le permet).

Pour chaque image, il faut vérifier que l'image a bien été conservée en image vectorielle (les objets de l'image restent des objets dans le fichier PDF) et non pas en image matricielle.

**Rendu à l'œil nu**

Pour ce test, il suffit de regarder le résultat dans la page. Est-ce que, sans zoom, l'image semble être correcte ?
Dans le formulaire de résultats, 1. correspond à rendu correct, 2. à une dégradation visible de l'image.

| | | |
|---|---|---|
| Rendu à l'œil nu | 1. le rendu est correct<br>2. l'image a perdu en qualité | |

**Rendu en zoom**

La deuxième partie du test consiste à zoomer sur les détails de l'image (image dans le logiciel original et fichier PDF) et comparer les deux pour voir s'il y a des dégradations.

Dans le formulaire des résultats, 1. correspond à rendu correct, 2. à une perte évidente de qualité.

| | | |
|---|---|---|
| Rendu en zoom | 1. le rendu est correct<br>2. l'image a perdu en qualité | |

**Format**

La vérification du format s'effectue au moyen de l'inspecteur PitStop.
Si l'image est vectorielle, chaque composant de l'image doit être un objet à part et il ne doit pas y avoir de dégradation avec un fort zoom.

Dans le formulaire de résultats : 1. correspond à une image vectorielle, 2. à une image convertie en raster (ou image matricielle), 3. à une absence d'image.

| | | |
|---|---|---|
| Format | 1. l'image reste vectorielle | |

| | | |
|---|---|---|
| | 2. l'image a été convertie en raster<br>3. l'image n'a pas été convertie | |

## Couleur

Il est important pour l'archivage que les définitions de couleurs des éléments présents dans le document source soient autant que possible identiques dans le fichier PDF créé.
Des éléments vectoriels et des éléments matriciels, aussi bien texte que graphiques/dessins, ont été testés avec les définitions de couleur suivantes :

- niveau de gris
- couleurs RVB

Pour chaque cas testé, la définition de couleur doit rester identique à celle du document d'origine.
Les couleurs RVB (affichage sur écran) ont été privilégiées par rapport aux couleurs CMJN, couramment utilisées par les logiciels de Publication Assistée par Ordinateur (PAO) pour l'impression papier. En effet, la présente étude a été réalisée dans une perspective d'archivage électronique et non de publication papier ou d'impression, qui nécessite la définition d'une imprimante de référence et de ses paramètres.

Pour visualiser une couleur, l'outil PitStop a été utilisé :

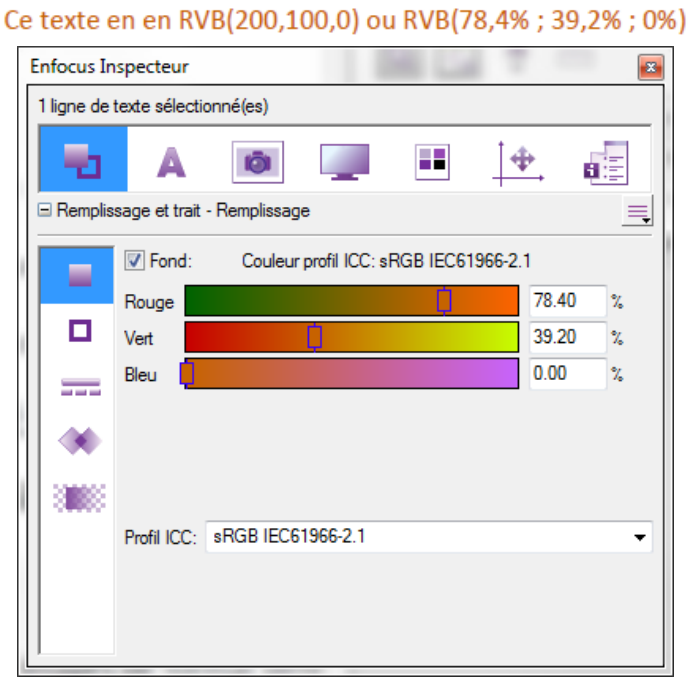



**Figure 11 :** Copie d'écran de l'inspecteur PitStop pour la vérification des couleurs

Cet exemple montre que les valeurs de couleurs ont été gardées et que le profil couleur RVB a été utilisé.
Dans le formulaire de résultats : 2. correspond à une couleur exacte, 1. à une couleur modifiée, 0. à un test impossible car le logiciel source ne supporte pas ce type de couleur.

| | Gris | RVB |
|---|---|---|
| Texte | | |
| Graphiques vectorielles | | |
| Graphiques raster (ou matriciels) | | |
| Autres commentaires | | |

Pour les images matricielles, seules les images PNG ont été testées. Les couleurs de JPEG sont forcément modifiées par le processus de compression.

*Remarque* : L'outil pipette de PitStop a été utilisé pour tester les valeurs RVG des pixels d'une image matricielle. Pour les autres types de couleurs, la procédure de tests s'est basée sur l'utilisation de l'outil flèche d'Acrobat pour copier l'image dans le presse-papiers puis la coller dans un outil comme GIMP pour valider les couleurs.

## Polices

Les tests sur les polices consistent à vérifier qu'aucune substitution de police n'est faite lors de la conversion et que le type de police est gardé autant que possible.
L'utilisation de jeux partiels de police est fortement recommandée pour éviter de charger un fichier avec des informations qui ne sont pas nécessaires.

Les tests effectués ont permis de vérifier l'utilisation des polices suivantes :

- police TrueType
- police OpenType (avec caractères cyrilliques, grecs, arabes, chinois et certains caractères casseau)

La police Type 1 n'a pas été testée car elle n'existe quasiment plus sous Windows.

Pour savoir si les polices sont incluses dans un fichier PDF, il faut l'ouvrir avec Acrobat Pro et regarder dans Fichier/Propriétés. L'onglet « Polices » indique les polices incorporées dans le fichier avec leurs caractéristiques (jeu complet, partiel, etc.).

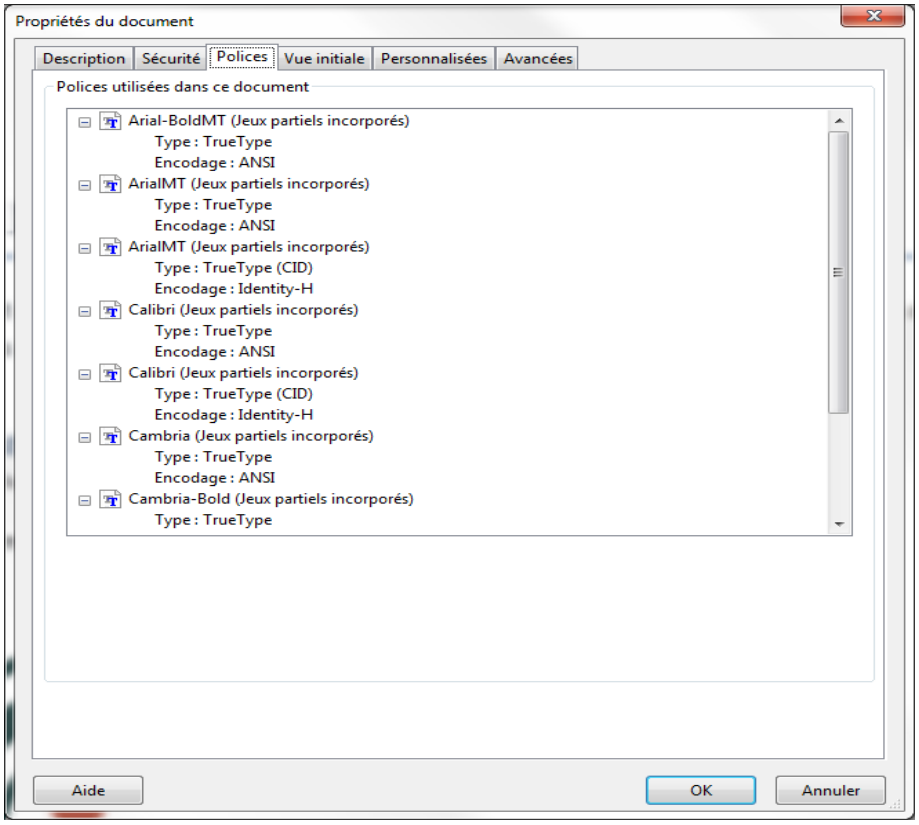



**Figure 12 :** Copie d'écran de l'onglet « Polices » dans Adobe Acrobat Pro

Dans le formulaire de résultats, pour chaque type de police : 2. correspond à une police incluse en jeu partiel, 1. à une police embarquée en entier, 0. à une police non incluse dans le fichier PDF.

## Transparence

Pour vérifier que la transparence d'une image est respectée, il suffit d'en construire une dans le fichier source et de vérifier dans Acrobat Pro si le fichier PDF créé donne le bon rendu.

**Rappel** : la transparence n'est pas permise dans le PDF/A-1.

Ce test n'est possible qu'avec des logiciels qui gèrent la transparence directement ou qui permettent l'inclusion de formats d'image vectorielle gérant la transparence (comme le PDF ou le SVG). Par exemple, Microsoft Word ne gère ni le PDF ni le SVG en image incluse, mais il permet de créer des dessins qui utilisent la transparence.

Dans le formulaire de résultats : 1. correspond à une bonne gestion de la transparence, 0. à une mauvaise gestion.

## Métadonnées

**Métadonnées utilisateur**

Il est important de garder les métadonnées du fichier d'origine dans le fichier PDF créé.

Dans Acrobat, « *Fichier->Propriétés->Métadonnées supplémentaires* » permet de voir quelles métadonnées ont été préservées. L'onglet *« Avancé »* montre la liste des types de métadonnées présents dans le document :

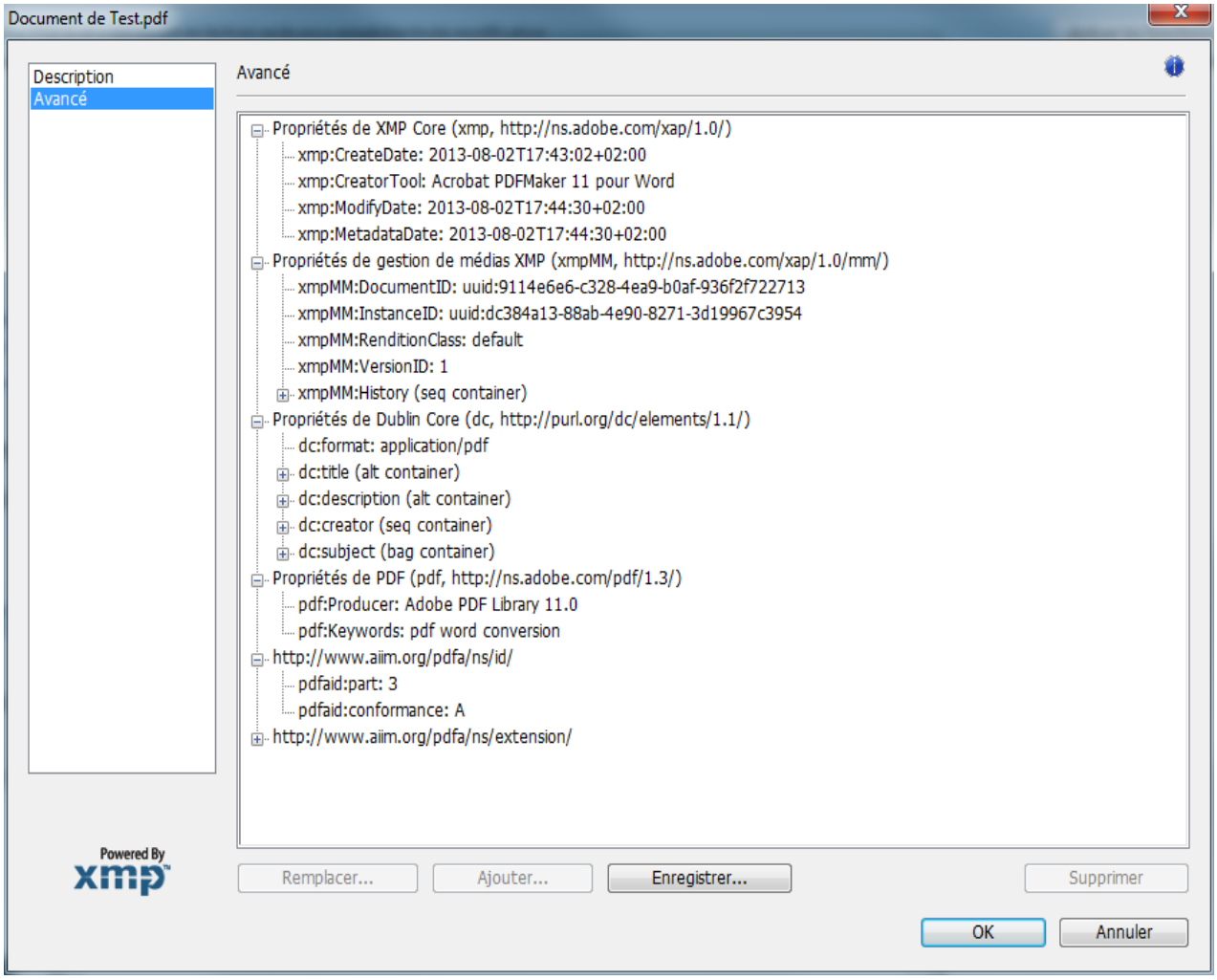



**Figure 13** : Copie d'écran de l'onglet « Avancé » dans Adobe Acrobat Pro

Les types de métadonnées les plus importants sont :

- les propriétés de PDF :
  - avec le nom du producteur « pdf:Producer »
  - et des mots clefs « pdf:Keywords »
- les propriétés de XMP Core :
  - date de création : « xmp:CreateDate »
  - outil de création : « xmp:Creator Tool »
  - date de dernière modification : « xmp:ModifyDate »
  - date de dernière modification des métadonnées : « xmp:MetadataDate »
- et les propriétés de Dublin Core :
  - Format : « dc:format »

- Titre : « dc:title »
- Description : « dc:description »
- Créateur : « dc:creator »
- Sujet : « dc:subject »

Ils comportent plusieurs niveaux qui indiquent les métadonnées conservées afin de comparer avec le document d'origine :

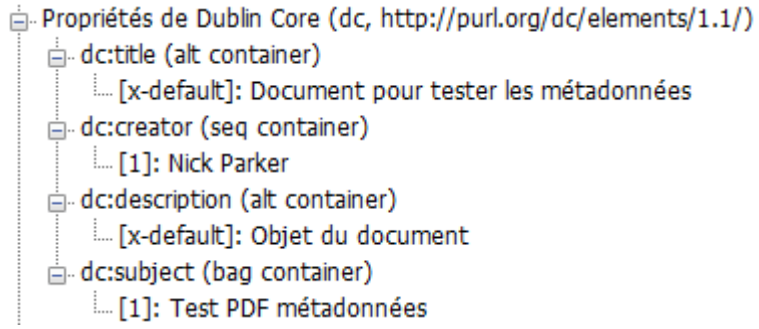



**Figure 14 :** Détail des métadonnées visibles dans l'onglet « Avancé »

Il n'y aura pas toujours une relation évidente « un pour un » entre toutes les métadonnées du fichier source et les métadonnées PDF, mais dans le cadre de la présente étude, si le titre et le nom d'auteur sont bien présents, le résultat peut être considéré comme acceptable. Le lecteur notera qu'il devra prêter une attention toute particulière au jeu de métadonnées dont la préservation est pertinente.

Dans le formulaire des résultats : 2. correspond à des métadonnées conservées de façon correcte dans le fichier PDF, 1. à un fichier source sans métadonnée, 0. à des métadonnées non reproduites dans le fichier PDF.

| | | |
|---|---|---|
| Métadonnées utilisateur | 0. métadonnées absentes dans le fichier PDF<br>1. le fichier source ne comporte pas de métadonnées<br>2. métadonnées correctes | |

**Métadonnées générées**

Il existe aussi des métadonnées générées directement par l'outil de conversion. Parmi celles-ci, les deux plus importantes sont :

- Application/Créateur (xmp:CreatorTool) — cette valeur correspond normalement au nom de l'application dans laquelle le document original a été créé (Microsoft Word par exemple) ;
- Outil de conversion/Producteur (pdf:Producer) — cette valeur correspond normalement au nom de l'outil qui a fait la conversion entre le fichier source et le fichier PDF.

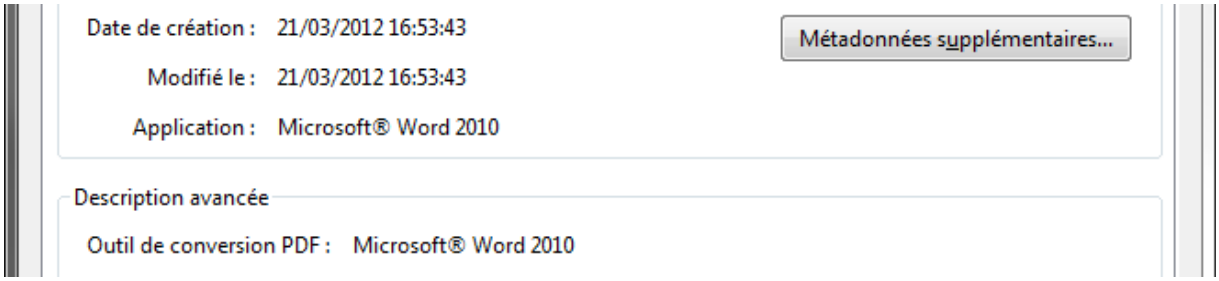



**Figure 15** : Copie d'écran de l'onglet « Description » dans Adobe Acrobat Pro

Dans le formulaire des résultats : 1. indique que les métadonnées générées sont correctes, 0. qu'elles ne le sont pas.

| | | |
|---|---|---|
| Métadonnées générés | 0. métadonnées absentes ou incorrectes<br>1. métadonnées correctes | |

## Sommaire du fichier PDF

La création d'un sommaire n'est possible que si l'outil de création a accès à la structure du document. Par exemple, dans Microsoft Word, les styles de type « Titre 1 », « Titre 2 » etc. peuvent être utilisés pour indiquer la hiérarchie des titres dans le document.
Cette hiérarchie peut alors être convertie en sommaire lors de la conversion en PDF.

La procédure de test repose sur l'ouverture du fichier à tester dans Acrobat et sur la sélection *« Affichage->Panneaux de navigation->Signets »*.
S'il y a un sommaire dans le document, il est alors affiché à gauche de la fenêtre comme suit :

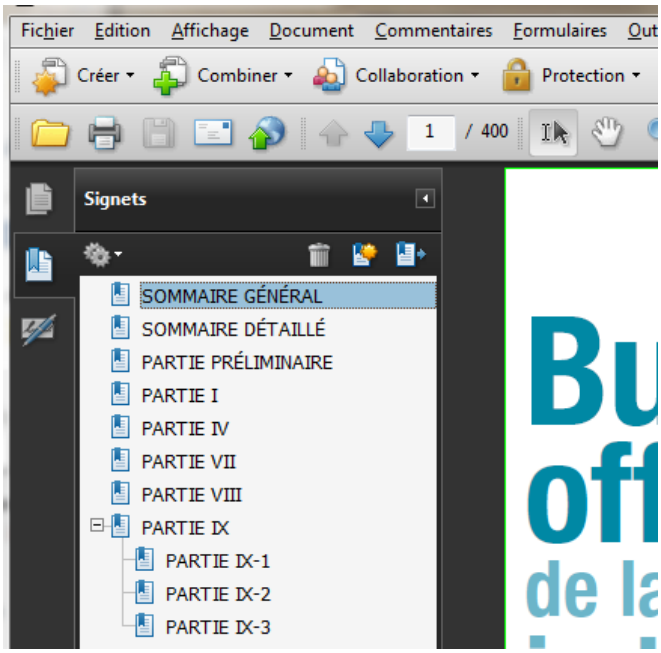



**Figure 16 :** Copie d'écran du sommaire PDF

Il faut vérifier qu'il est possible de cliquer sur chaque entrée du sommaire pour afficher la page concernée par l'entrée dans la fenêtre principale de l'application.

Dans le formulaire de résultats : 2. correspond à un sommaire créé à partir de la structure du document source, 1. à un format de document source ne permettant pas de décrire une structure de document, 0. à une structure du document source non reproduite dans le fichier PDF.

| | | |
|---|---|---|
| Création de sommaire | 0. la structure n'est pas reproduite<br>1. le format source ne décrit pas la structure<br>2. un sommaire est créé correctement | |

Les tests ont également permis d'évaluer les possibilités de configuration du sommaire pour chaque outil (i.e. est-ce qu'on peut indiquer le nombre de niveaux à mettre dans le sommaire ou la façon de faire correspondre le contenu du document source avec le sommaire généré ?)

Dans le formulaire de résultats : 0. correspond à aucune configuration possible, 1. à un paramétrage possible.

| | | |
|---|---|---|
| Configuration de sommaire | 0. il n'est pas possible de définir les niveaux<br>1. il est possible de définir les niveaux à inclure | |

## Structure logique de document

### Structuration :

La structure logique du document est importante pour l'archivage. C'est la présence de cette structure qui permet de différencier une simple image de page d'un document avec un contenu qui peut être extrait.

On peut reconnaître quatre niveaux de structuration dans un document PDF :

1) Aucune structure — le fichier PDF contient simplement des images de pages ;

2) Flux Unicode — le fichier PDF contient en plus des images de page, le contenu textuel en Unicode, mais sans fournir des informations utiles sur la structure du document ;

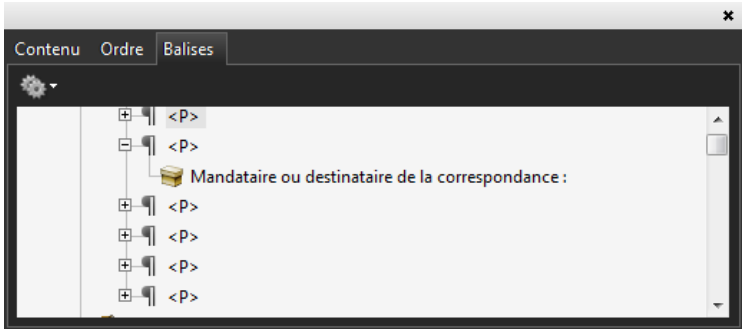
3) Structure PDF de base — le fichier PDF contient en plus du flux textuel en Unicode, des indications de structure utilisant uniquement les balises définies dans la norme PDF ;

4) Structure PDF riche — la structure du document utilise des balises présentes dans le document source pour indiquer la nature de chaque élément. Ces balises ont aussi une correspondance avec les balises de base PDF pour permettre une interprétation par un outil que ne comprennent pas les balises du document.

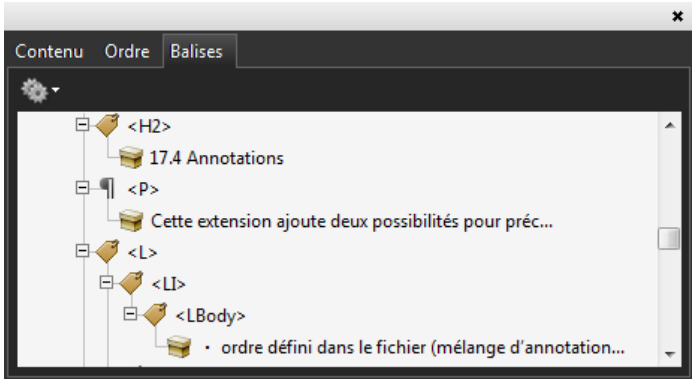
Le niveau de structuration a été déterminé avec Acrobat en sélectionnant le menu Affichage/Panneaux de navigation/Balises.
Une fenêtre s'ouvre et permet de naviguer dans les balises du document.

1) Dans ce cas il n'existe pas de structure dans le fichier PDF :

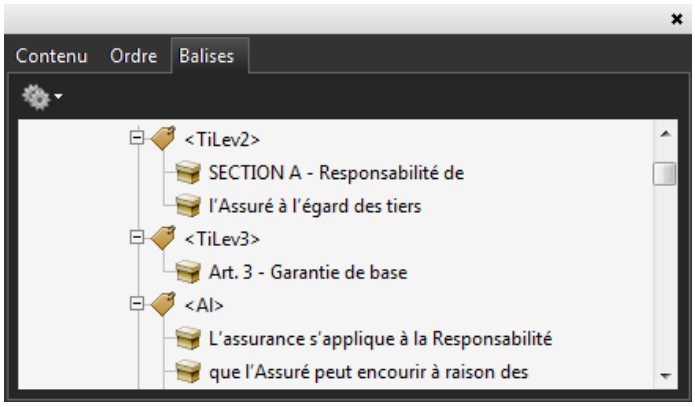

2) Dans ce cas il y a une structure, mais elle est plate. Par exemple, tous les paragraphes utilisent la balise <P> même les titres.




3) Dans ce cas, les balises sont les balises PDF de base comme décrit dans la norme ISO 32000. C'est donc un document PDF avec une structure PDF de base :



4) Dans ce cas, les noms des styles utilisés dans le document source ont été préservés en tant que balises dans le document PDF :




Le test de cette fonctionnalité se fait avec un document source bien structuré. Dans un traitement de texte, par exemple, les titres utilisent les styles prévus comme des titres de façon hiérarchique et les paragraphes utilisent des styles qui indiquent leur fonction.

Lors de la conversion du fichier source, il faut généralement choisir l'option PDF balisé :

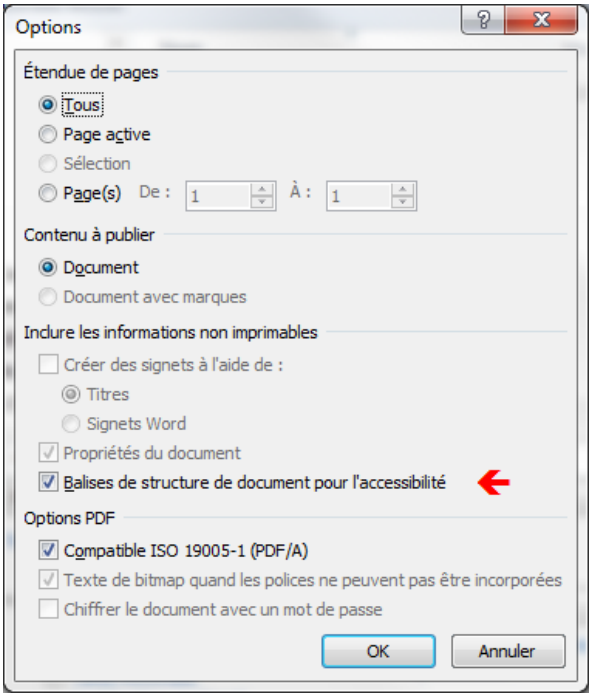



**Figure 17** : Copie d'écran de l'onglet « Options » d'Adobe Acrobat Pro

Après la création du fichier PDF, il faut comparer les balises créées avec la structure du document source pour déterminer le niveau de structuration.

Dans le formulaire des résultats, on a indiqué le niveau de structuration constaté :

| | | |
|---|---|---|
| Niveau de structuration | 1. aucune structure<br>2. flux Unicode<br>3. structure PDF de base<br>4. structure PDF riche | |

## Gestion Unicode :

Outre la structure, il est important que le document contienne des informations Unicode sur les caractères qu'il contient. Des validations PDF/A peuvent détecter si les polices se déclarent conforme à Unicode, mais il est aussi intéressant de vérifier si les caractères peuvent être récupérés par copier/coller du texte.

Par exemple, pour tester la récupération des caractères non latins (d'autres alphabets ou des caractères casseau), il faut les copier depuis le fichier PDF vers un autre éditeur de texte qui reconnaît Unicode (le « Bloc-Notes » de Windows par exemple),

Par exemple si le document PDF affiche :

Des caractères arabes : يَسْطَعُ النُّورُ فِي الظُّلْمَةِ، وَالظُّلْمَةُ لَمْ تَهزِمْهُ

mais le Bloc-notes reçoit :

Des caractères arabes : يَسْطَ عُ ٱلنُّو رُ فُ ِي ٱلظُّلْمَةِ، وَٱلظُّلْمَ ةُ لَ ُمْ ْ تَ ْهزِمْ هُ

tandis qu'un autre convertisseur est capable de produire :

Des caractères arabes : يَسْطَعُ النُّورُ فِي الظُّلْمَةِ، وَالظُّلْمَةُ لَمْ تَهزِمْهُ

Il est constaté que la conversion occasionne des problèmes/erreurs. Pour l'arabe, par exemple, il faut bien prendre en compte la direction du texte — « Bloc-Notes » est capable de retourner le texte pour le lire de droite à gauche et donc l'afficher correctement (le déplacement du curseur dans le texte permet de vérifier l'ordre).

Il peut y avoir également des problèmes avec les caractères casseau. Ils se transforment en lettres de l'alphabet, disparaissent du flux de texte, ou des blancs inattendus apparaissent dans les mots. Il faut alors vérifier que les polices utilisées dans le document source comportent des informations Unicode, et que l'éditeur de texte utilisé permet bien de visualiser des caractères Unicode. L'utilisation dans « Bloc-Notes » de la police définie par défaut (*Lucida Sans Unicode*) permet de visualiser la plupart des caractères Unicode.

Par exemple, si le fichier source affiche :

Des caractères casseau : ▲▶▼◀♠ ♡ ◇ ♣

le « Bloc-Notes » doit recevoir :

Des caractères casseau : ▲▶▼◀♠♡◇♣

et non :

Des caractères casseau :

Ici les caractères casseau ont complètement disparu du flux de texte.

Dans le formulaire des résultats : 1. indique que les caractères sont bien conservés, 0. qu'il y a des problèmes de conservation des caractères.

| | | |
|---|---|---|
| Gestion Unicode | 0. il y a des problèmes dans la conservation des caractères<br>1. les caractères sont conservés correctement | |

## Liens hypertexte

Il peut être utile de garder des liens hypertexte actifs dans un document PDF pour faciliter son utilisation. Trois types de liens ont été testés :

- les liens vers des adresses web (liens externes) ;
- les liens à l'intérieur du document ;
- les tables des matières (qui sont des liens internes mais générés de façon différente).

En testant les liens dans le document PDF créé, il est important de vérifier qu'Acrobat Pro n'a pas désactivé les liens. Dans certaines versions d'Acrobat Pro, un fichier signalé comme PDF/A peut être ouvert dans un mode spécifique où les liens ne peuvent pas être utilisés. Il est donc nécessaire de désactiver ce mode. Pour cela, il suffit de :

- sélectionner l'onglet *« Documents »* du menu *« Édition/Préférences »*
- on y trouve une option pour indiquer s'il faut utiliser ce mode d'affichage. En choisissant « Jamais » les liens seront activés dans le document :

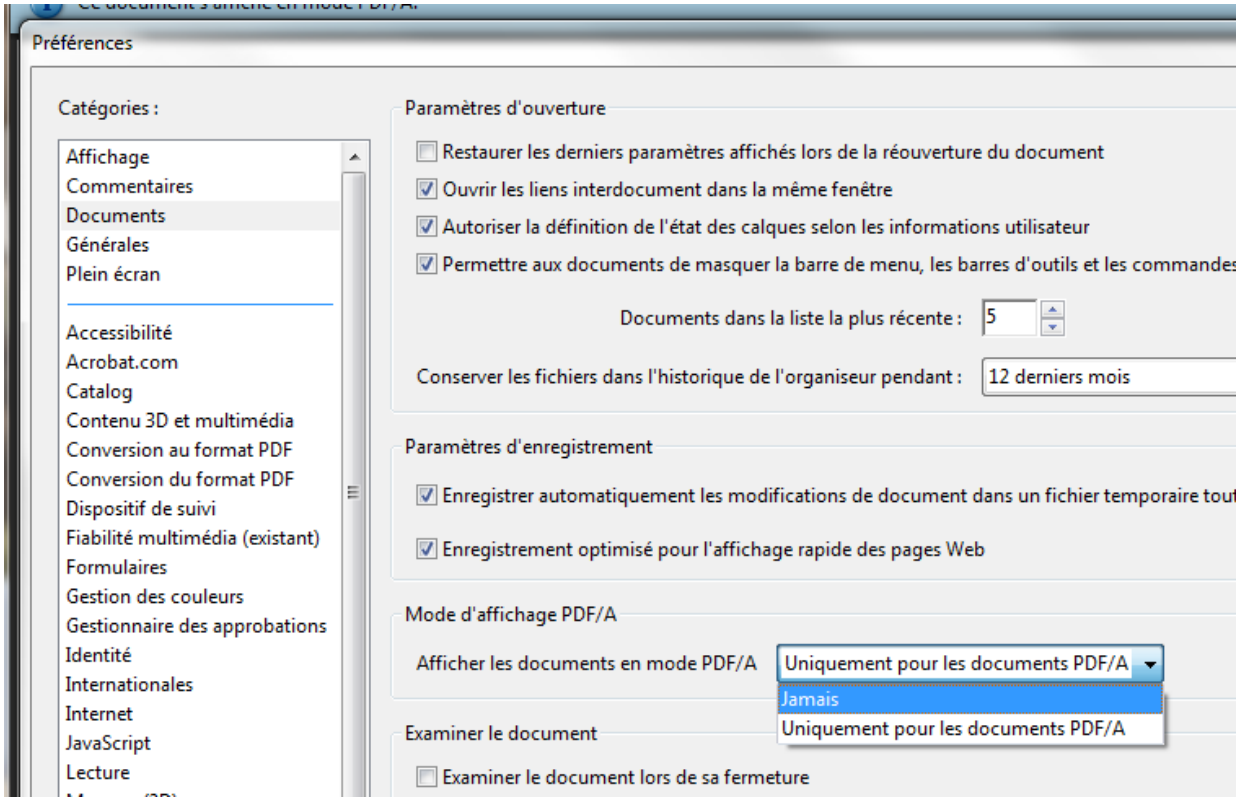



**Figure 18** : Copie d'écran des préférences de l'onglet « Documents » d'Adobe Acrobat Pro

Après avoir vérifié que les liens sont bien actifs, il suffit de cliquer sur chaque texte qui est censé être un lien pour voir si le lien fonctionne comme prévu.

Dans le formulaire des résultats : 1. indique que le lien fonctionne, 0. qu'il ne fonctionne pas.

| | | |
|---|---|---|
| Lien vers un URL | 0. ne fonctionne pas<br>1. fonctionne correctement | |
| Lien vers une autre partie du document | 0. ne fonctionne pas<br>1. fonctionne correctement | |
| Table des matières | 0. ne fonctionne pas<br>1. fonctionne correctement | |

## Fonctionnalités non testées

Certaines fonctionnalités n'ont pas été testées car elles ont été jugées comme mineures ou non pertinentes dans le cadre de l'archivage.

Il s'agit notamment des fonctionnalités d'impression, de préférences de documents, d'articles, de contenu facultatif, d'annotations (car ajoutées a posteriori sur le fichier PDF), de compression (en dehors de la compression sur les images matricielles), de formulaires interactifs (Acroforms, XFA), de scripts, de fichiers embarqués (utile uniquement en PDF/A-3), d'impressions professionnelles d'images de haute résolution (OPI), et de CAO.

Par ailleurs, les signatures numériques ne faisant pas partie des fichiers d'origine, les convertisseurs ne sont pas capables de les inclure dans le fichier créé sans une intervention manuelle. Aussi cette fonctionnalité n'a pas été testée, car dans une perspective d'archivage, les signatures numériques posent des problèmes quant à la lisibilité et à l'accès ultérieur au document.

Pour l'archivage, il est important qu'il n'y ait pas de limites d'accès sur un fichier. Les fonctionnalités de mot de passe ne sont donc pas compatibles avec l'objectif de cette étude.

Même si le format PDF peut contenir des objets multimédias, cette fonctionnalité n'a pas été testée car les normes PDF/A ne les autorisent pas et il existe des formats plus appropriés pour stocker ce genre de contenu[6].

Enfin, les fonctionnalités de capture de pages web permettent surtout d'identifier la source des informations.

Etant donné que le format HTML n'a pas été retenu comme format source pour les tests, aucun test de cette fonctionnalité n'a été effectué.

[6] Voir à ce sujet le « *Guide méthodologique pour le choix de formats numériques pérennes dans un contexte de données orales et visuelles* », publié par le SIAF/TGE Adonis/CINES, v2, 2011, accessible en ligne : http://www.archivesdefrance.culture.gouv.fr/static/4923

# Bilan des tests par fonctionnalité du PDF

## Conversion de l'image

Dans tous les cas testés, au premier abord, le rendu visuel de l'image est correct. Toutefois, une analyse plus approfondie permet de noter des différences quant à la qualité des fichiers convertis grâce aux différents logiciels (pertes de qualité fréquentes).

Images matricielles :

| | | PNG sans transparence | | PNG avec transparence | | JPEG | |
|---|---|---|---|---|---|---|---|
| **Logiciel utilisé** | **Format sortie** | **Basse résolution (500px)** | **Haute résolution (2000px)** | **Basse résolution (500px)** | **Haute résolution (2000px)** | **Basse résolution** | **Haute résolution** |
| **Writer LibreOffice** | PDF 1.4 | Perte de qualité | Diminution de la résolution | Perte de qualité | OK | Perte de qualité + compression avec perte | Diminution de la résolution + compression avec perte |
| | PDF/A-1a | OK | Diminution de la résolution | Perte de qualité | OK | Perte de qualité + compression avec perte | Diminution de la résolution + compression avec perte |
| **Writer OpenOffice** | PDF 1.4 | Perte de qualité | Diminution de la résolution | Perte de qualité | OK | Perte de qualité + compression avec perte | Diminution de la résolution + compression avec perte |
| | PDF/A-1a | Perte de qualité | Diminution de la résolution | Perte de qualité | OK | Perte de qualité + compression avec perte | Diminution de la résolution + compression avec perte |
| **Word** | PDF 1.5 | Perte de qualité | Diminution de la résolution | Perte de qualité | Diminution de la résolution | Perte de qualité + compression avec perte | Diminution de la résolution + compression avec perte |
| | PDF/A-1a | Perte de qualité | Diminution de la résolution | Perte de qualité | Diminution de la résolution | Perte de qualité + compression avec perte | Diminution de la résolution + compression |

| | | PNG sans transparence | | PNG avec transparence | | JPEG | |
|---|---|---|---|---|---|---|---|
| **Logiciel utilisé** | **Format sortie** | **Basse résolution (500px)** | **Haute résolution (2000px)** | **Basse résolution (500px)** | **Haute résolution (2000px)** | **Basse résolution** | **Haute résolution** |
| | | | | | | | avec perte |
| **Acrobat Pro** | PDF 1.4 | Perte de qualité | Diminution de la résolution | Perte de qualité | Perte de qualité + diminution de la résolution | Perte de qualité + compression avec perte | Diminution de la résolution + compression avec perte |
| | PDF 1.7 | Perte de qualité | Diminution de la résolution | Perte de qualité | Perte de qualité + diminution de la résolution | Perte de qualité + compression avec perte | Diminution de la résolution + compression avec perte |
| | PDF/A-3a | OK | Diminution de la résolution | Perte de qualité Diminution de la résolution | Perte de qualité + diminution de la résolution + compression avec perte | Perte de qualité + compression avec perte | Diminution de la résolution + compression avec perte |
| **pdfaPilot** | PDF 1.5 | Perte de qualité + compression avec perte | Perte de qualité + diminution de la résolution + compression avec perte | Perte de qualité | Perte de qualité + diminution de la résolution + compression avec perte | Perte de qualité + compression avec perte | Perte de qualité + diminution de la résolution + compression avec perte |
| | PDF/A-3a | Perte de qualité + compression avec perte | Perte de qualité + diminution de la résolution + compression avec perte | Perte de qualité | Perte de qualité + diminution de la résolution + compression avec perte | Perte de qualité + compression avec perte | Perte de qualité + diminution de la résolution + compression avec perte |
| **PDF Creator** | PDF 1.4 | Perte de qualité + | Diminution de la résolution | Perte de qualité + | Diminution de la résolution | Perte de qualité + | Diminution de la résolution |

| | | PNG sans transparence | | PNG avec transparence | | JPEG | |
|---|---|---|---|---|---|---|---|
| **Logiciel utilisé** | **Format sortie** | **Basse résolution (500px)** | **Haute résolution (2000px)** | **Basse résolution (500px)** | **Haute résolution (2000px)** | **Basse résolution** | **Haute résolution** |
| | | diminution de la résolution | | diminution de la résolution | | diminution de la résolution | |
| | PDF/A-2a | Impossible à obtenir avec Windows 7 - 64 bits | | | | | |

Les tests effectués montrent que le format PNG est plus fiable que le format JPEG lors de la conversion PDF, notamment si on utilise les outils de LibreOffice et d'OpenOffice.

L'utilisation en amont d'images en haute résolution peut être un bon moyen de compenser la diminution de la résolution, pendant la conversion PDF, afin d'obtenir un résultat qui reste correct. Le lecteur notera que les paramètres sont généralement accessibles dans l'outil de conversion.

Images vectorielles :

| Logiciel utilisé | Format sortie | Images EMF | Images EPSF | Images PDF |
|---|---|---|---|---|
| **Writer LibreOffice** | PDF 1.4 | OK | OK | Impossible |
| | PDF/A-1a | OK | OK | Impossible |
| **Writer OpenOffice** | PDF 1.4 | OK | OK | Impossible |
| | PDF/A-1a | OK | OK | Impossible |
| **Word** | PDF 1.5 | OK | OK | Perte de qualité + image Bitmap |
| | PDF/A-1a | OK | OK | Perte de qualité + image Bitmap |
| **Acrobat Pro** | PDF 1.4 | OK | OK | Perte de qualité + image Bitmap |
| | PDF 1.7 | OK | OK | Perte de qualité + image Bitmap |
| | PDF/A-3a | Perte de qualité + image convertie en bitmap | OK | Perte de qualité + image Bitmap |

| Logiciel utilisé | Format sortie | Images EMF | Images EPSF | Images PDF |
|---|---|---|---|---|
| **pdfaPilot** | PDF 1.5 | OK | OK | Perte de qualité + image Bitmap |
| | PDF/A-3a | OK | OK | Perte de qualité + image Bitmap |
| **PDF Creator** | PDF 1.4 | OK | OK | Perte de qualité + image Bitmap |
| | PDF/A-2a | Impossible à obtenir avec Windows 7 - 64 bits | | |

Au vu des tests réalisés, la conversion en PDF d'images vectorielles entraîne moins d'erreurs que pour les images rasters. De manière générale, les logiciels testés gèrent correctement la conversion. Étonnamment, Acrobat Pro est le seul logiciel qui ne convertit pas convenablement une image EMF en PDF/A.

Concernant le cas particulier de l'insertion d'une image PDF dans un fichier source, les logiciels de traitement de texte Open Source (Writer de LibreOffice et OpenOffice) ne sont pas performants et il est donc ensuite impossible de convertir correctement le document source en PDF. Cette fonctionnalité existe pourtant dans LibreOffice et OpenOffice, en ajoutant un « add-in » mais dans le cas de test, l'image obtenue après insertion n'était pas utilisable. Il semble plausible de trouver des cas particuliers pouvant vérifier parfaitement cette fonctionnalité. Ainsi l'image obtenue après insertion peut être correctement visible et donc il serait alors possible de convertir le document obtenu ensuite. Cela témoigne de la difficulté d'obtenir un ou des fichiers de tests pertinents, permettant de vérifier toutes les fonctionnalités à tester.
Cela indique aussi l'importance du contenu du fichier à convertir. Chaque fichier à convertir est un cas particulier qui demande une adaptation spécifique pour arriver à un résultat correct (surveiller le paramétrage du logiciel qui est différent suivant la conversion en PDF 1.7 ou PDF/A-3a par exemple).

## Couleur

| Logiciel utilisé | Format sortie | Niveaux de gris | RVB |
|---|---|---|---|
| **Writer LibreOffice** | PDF 1.4 | OK | OK |
| | PDF/A-1a | OK | OK |
| **Writer OpenOffice** | PDF 1.4 | OK | OK |
| | PDF/A-1a | OK | OK |
| **Word** | PDF 1.5 | OK | OK |
| | PDF/A-1a | OK | OK |
| **Acrobat Pro** | PDF 1.4 | OK | OK |
| | PDF 1.7 | OK | OK |

| | | | |
|---|---|---|---|
| | PDF/A-3a | OK | OK |
| **pdfaPilot** | PDF 1.5 | OK | OK |
| | PDF/A-3a | OK | OK |
| **PDF Creator** | PDF 1.4 | OK | OK |
| | PDF/A-2a | Impossible à obtenir avec Windows 7 - 64 bits | |

Quel que soit le logiciel, les tests réalisés montrent que les couleurs sont bien restituées après la conversion PDF, que ce soit pour le texte ou les images.

## Polices

| Logiciel utilisé | Format sortie | Police TrueType | Police Open Type | Police cyrillique | Police grecque | Police arabe | Police chinoise | Caractères Casseau |
|---|---|---|---|---|---|---|---|---|
| **Writer LibreOffice** | PDF 1.4 | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel |
| | PDF/A-1a | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel |
| **Writer OpenOffice** | PDF 1.4 | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel |
| | PDF/A-1a | Changement de police pour certaines | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel |

| Logiciel utilisé | Format sortie | Police TrueType | Police Open Type | Police cyrillique | Police grecque | Police arabe | Police chinoise | Caractères Casseau |
|---|---|---|---|---|---|---|---|---|
| | | polices + polices non embarquées pour d'autres polices | | | | | | |
| **Word** | PDF 1.5 | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel |
| | PDF/A-1a | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Police non embarquée | Jeu partiel |
| **Acrobat Pro** | PDF 1.4 | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | Change-ment de police + police non embarquée | Jeu partiel | Jeu partiel |
| | PDF 1.7 | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | Change-ment de police + police non embarquée | Jeu partiel | Jeu partiel |
| | PDF/A-3a | Changement de police pour | Jeu partiel | Jeu partiel | Jeu partiel | Change-ment de police + | Jeu partiel | Jeu partiel |

| Logiciel utilisé | Format sortie | Police TrueType | Police Open Type | Police cyrillique | Police grecque | Police arabe | Police chinoise | Caractères Casseau |
|---|---|---|---|---|---|---|---|---|
| | | certaines polices + polices non embarquées pour d'autres polices | | | | police non embarquée | | |
| **pdfaPilot** | PDF 1.5 | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | OK (Police embarquée) | Jeu partiel | Jeu partiel |
| | PDF/A-3a | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | OK (Police embarquée) | Jeu partiel | Jeu partiel |
| **PDF Creator** | PDF 1.4 | Changement de police pour certaines polices + polices non embarquées pour d'autres polices | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel | Jeu partiel |
| | PDF/A-2a | Impossible à obtenir avec Windows 7 - 64 bits | | | | | | |

Les résultats obtenus montrent que pour la plupart des polices testées (hormis les polices TrueType), seul un jeu partiel est embarqué, ce qui est fortement recommandé du fait du gain de place obtenu.
Toutefois lors d'une conversion en PDF/A, il n'est pas aisé de savoir si la police d'origine est bien incorporée. Il faut faire une comparaison entre la liste des polices dans le fichier en entrée et la liste des polices concernées dans le fichier converti.

Il est à noter également que le **PDF/A-1** n'autorise pas les polices OpenType, ce qui explique leur conversion en Type 1 ou TrueType, sans toutefois que cela n'entraîne d'importants changements (pour rappel : les polices OpenType sont un dérivé des deux autres).

Enfin, lors de l'ouverture d'un document avec Word ou Writer de LibreOffice, il faut accorder une vigilance particulière sur la présence effective des polices sur l'ordinateur. En effet, ces logiciels affichent généralement le nom de la police enregistrée dans le document même si celle-ci n'est pas présente sur le poste et que le logiciel l'a remplacée par une autre.

## Structure logique du document

| Logiciel utilisé | Format sortie | Sommaire PDF | Niveau de structuration | Gestion Unicode | Liens internes et externes |
|---|---|---|---|---|---|
| **Writer LibreOffice** | PDF 1.4 | OK | Aucune Structure PDF | Pb de conservation des caractères | OK |
| | PDF/A-1a | OK | Structure PDF de base | Pb de conservation des caractères | OK |
| **Writer OpenOffice** | PDF 1.4 | OK | Aucune Structure PDF | Pb de conservation des caractères | OK |
| | PDF/A-1a | OK | Structure PDF de base | Pb de conservation des caractères | OK |
| **Word** | PDF 1.5 | Structure non reproduite | Structure PDF de base | Pb de conservation des caractères | OK |
| | PDF/A-1a | Structure non reproduite | Flux unicode | Pb de conservation des caractères | OK |
| **Acrobat Pro** | PDF 1.4 | OK | Structure PDF de base | Pb de conservation des caractères | OK |
| | PDF 1.7 | OK | Structure PDF de base | Pb de conservation des caractères | OK |
| | PDF/A-3a | OK | Structure PDF de base | Pb de conservation des caractères | OK |
| **pdfaPilot** | PDF 1.5 | OK | Structure PDF de base | OK | OK |
| | PDF/A-3a | OK | Structure PDF de base | OK | OK |
| **PDF Creator** | PDF 1.4 | Aucune structure | Aucune structure PDF | Les caractères sont conservés correctement sauf pour les caractères arabes et chinois | OK pour les liens externes.<br>KO pour les liens internes |
| | PDF/A-2a | Impossible à obtenir avec Windows 7 - 64 bits | | | |

Les logiciels testés génèrent correctement un sommaire PDF lors de la conversion (excepté Word), ce qui facilite la navigation dans le document.
La plupart des logiciels proposent une structure de PDF de base, ce qui permet une meilleure accessibilité au document. A noter que Word convertit la structure en flux Unicode pour le PDF/A-1a.
La gestion Unicode des textes est très bien exploitée par l'outil **pdfaPilot de Callas.** Les tests des autres logiciels montrent que les langues « européennes » sont mieux traitées que les langues « orientales ».
Dans tous les cas testés, la conversion PDF préserve correctement l'ensemble des liens présents dans le document.

## Autres fonctionnalités

| Logiciel utilisé | Format sortie | Gestion de la transparence | Métadonnées |
|---|---|---|---|
| **Writer LibreOffice** | PDF 1.4 | OK | OK |
| | PDF/A-1a | Fausse transparence | OK |
| **Writer OpenOffice** | PDF 1.4 | OK | OK |
| | PDF/A-1a | Fausse transparence | OK |
| **Word** | PDF 1.5 | OK | OK |
| | PDF/A-1a | Pas de gestion de la transparence | OK |
| **Acrobat Pro** | PDF 1.4 | OK mais fausse transparence | Absence de métadonnées utilisateurs et de métadonnées générées |
| | PDF 1.7 | OK mais fausse transparence | OK |
| | PDF/A-3a | OK mais fausse transparence | OK |
| **pdfaPilot** | PDF 1.5 | OK mais fausse transparence | OK |
| | PDF/A-3a | OK mais fausse transparence | OK |
| **PDF Creator** | PDF 1.4 | OK | OK |
| | PDF/A-2a | Impossible à obtenir avec Windows 7 - 64 bits | |

La gestion de la transparence est révélatrice de ce que propose la norme pour le PDF/A pour des images qui se superposent. Le visuel semble parfait mais il n'est pas possible de faire de copier/coller dans un autre éditeur et récupérer séparément les deux objets images.

Souvent il y a transformation en une seule image qui permet de voir la superposition des deux objets images mais l'export des deux objets séparément n'est pas possible (= « fausse transparence »). Seuls les « Writer » pour du PDF 1.4 et Word pour du PDF 1.5 respectent la transparence des deux dessins.
On peut noter que bien que la norme PDF/A-1a n'autorise pas la transparence, elle est toutefois visible dans les convertisseurs testés avec ce format de sortie mais il ne s'agit que d'une apparence visuelle (elle est convertie sous forme d'image Bitmap).

Les métadonnées sont bien gérées lors de la conversion en PDF. Seul Acrobat Pro en convertissant en PDF 1.4 ne respecte pas ce principe.

# Problèmes rencontrés et conseils aux utilisateurs

La difficulté lors de la réalisation de tests est de construire un jeu de tests pertinent. Chaque fichier étant particulier, leur conversion est aussi spécifique. De plus, chaque logiciel n'a pas les mêmes possibilités de paramétrages. Le paramétrage par défaut permet d'optimiser, en principe, la rapidité d'exécution des programmes et l'optimisation de la place disque pour le fichier converti. Mais ce n'est pas nécessairement le bon profil pour obtenir un fichier PDF qui respecte les critères importants pour l'utilisateur :

- l'intégration des polices (même partielle) ;
- la compression des images sans pertes ;
- la conservation des liens ;
- la présence des métadonnées ;
- l'Unicode.
- Etc.

Les tests effectués ont permis de constater que ces types de paramétrage sont plus accessibles avec des logiciels « professionnels » payants comme Acrobat Pro d'Adobe ou pdfaPilot de Callas qu'avec les logiciels « gratuits » ou open source.
Toutefois, ce propos peut être nuancé s'il s'agit de convertir peu de fichiers, dont le contenu est relativement simple, uniforme et standard (dans le choix des polices par exemple) ou que le niveau d'exigences sur le respect du contenu (tant sur la forme que sur le fond) n'est pas trop élevé. A ce moment-là, l'utilisation de logiciels « gratuits » peut répondre aux besoins.
Il est à noter que l'installation de nombreux logiciels de conversion sur le même poste de travail peut perturber l'environnement de travail, en raison notamment des macros ajoutées par les logiciels de conversion dans les logiciels de traitement de texte.

La qualité de la conversion dépend aussi du format souhaité en sortie. Un fichier PDF 1.4 sera plus facile à convertir qu'un fichier PDF/A-3a par exemple : les contraintes sur le format étant moins fortes, le risque d'erreurs générées par la conversion est réduit.

Pour la conversion de contenu textuel, la gestion des polices peut poser des problèmes. En effet, la police utilisée dans le fichier peut ne pas exister sur le poste de travail sur lequel est effectuée la conversion.
Il est donc préférable d'effectuer les conversions directement sur le poste de travail utilisé lors de la création du fichier. Lorsque ce n'est pas le cas, il faut porter une attention particulière à la police déclarée par le logiciel et vérifier que celle-ci est bien installée sur le poste de travail. En effet, il arrive que le logiciel de traitement de texte affiche le nom de la police d'origine tout en la remplaçant par une autre police si celle-ci n'est pas présente sur le poste de travail.

Pour les images, il faut s'assurer que les images originales, avant intégration dans le document à convertir, aient un format normalisé ou ouvert. Par exemple pour les images matricielles (ou bitmap), les formats PNG, JPEG conviennent parfaitement (ce qui n'est pas le cas de BMP, propriété de Microsoft par exemple). Pour les images vectorielles, il faudrait choisir EMF, EPSF, PDF (image) ou SVG plutôt que SWF (Flash) propriété d'Adobe ou PICT (Apple).
Lorsqu'un paramétrage est possible, il faut également opter pour une compression sans perte : utilisation de la compression ZIP au lieu de la compression JPEG.

Le choix des formats PDF/A-3 et PDF/A-2 est fortement recommandé pour l'archivage mais sans utiliser la fonctionnalité de conteneur proposée par le PDF/A-3. En effet, l'inclusion de fichiers dans le fichier PDF pose des problèmes supplémentaires notamment quant à la validation du format de ces fichiers lors de l'archivage.

Ainsi, dans l'absolu, on pourrait dire que la liste des formats à privilégier pour l'archivage serait, par ordre de préférence :

1. PDF/A-3a (ou PDF/A-2a) : repère mnémonique avec « a » pour avancé
2. PDF/A-3u (ou PDF/A-2u) : « u » pour Unicode ;
3. PDF/A-3b (ou PDF/A-2b) : « b » pour basic
4. PDF/A-1a
5. PDF/A-1b
6. PDF 1.7
7. PDF 1.6
8. PDF 1.5
9. PDF 1.4

Les PDF 1.0, PDF 1.1, PDF 1.2 et PDF 1.3 ne sont pas conseillés parce qu'il s'agit de versions trop anciennes et non normalisées.

Enfin, une dernière difficulté a porté sur le choix d'un validateur PDF de référence pour s'assurer que le format des fichiers convertis était valide et bien formé. En effet, il n'existe pas de validateur de PDF normalisé, y compris pour le PDF/A, et les tests réalisés ont permis de constater que les validateurs intégrés dans les logiciels de conversion n'indiquaient pas forcément les mêmes résultats : Acrobat Pro d'Adobe et pdfaPilot de Callas par exemple, ou même les différentes versions d'Acrobat Pro (un fichier PDF/A a été validé par les versions 9 et 11 d'Acrobat Pro mais pas par la version 10).
Dans le cadre de cette partie de l'étude, c'est le logiciel Acrobat Pro version 11 d'Adobe qui a été retenu pour la validation des fichiers. En effet ce logiciel permet de valider tous les formats PDF et le fournisseur Adobe est une référence internationale reconnue dans le monde car il est l'inventeur de ce format. Mais ce n'est pas le seul validateur de format PDF. La troisième partie de cette étude permettra de comparer des validateurs différents et de vérifier si le choix fait ici, était pertinent.

# Conclusion

La conversion de fichiers sources vers les versions « standards » de PDF produit de meilleurs résultats que la conversion vers le PDF/A, plus adapté pour l'archivage mais plus contraignant. Les évolutions du format PDF permettent l'utilisation de nouvelles fonctionnalités au fur et à mesure des nouvelles versions, mais il n'est pas certain que celles-ci existent encore dans le futur. Tandis qu'une version PDF normalisée garantit davantage de stabilité pour l'utilisation du fichier converti et de son contenu.

La question du traitement par lot des fichiers n'est pas pertinente dans le cadre de cette étude qui visait à tester à partir d'un fichier de test unique les outils de conversion existants, notamment ceux intégrés dans les logiciels de traitement de texte ; et non le traitement de conversion d'un ensemble de fichiers de traitement de texte comme jeu d'essai en entrée de processus.

**Dans tous les cas, il n'existe pas de logiciel « miracle ». Tout dépend du niveau d'exigences et de l'utilisation future du fichier. Est-ce un fichier dont on veut conserver le rendu visuel du document initial, un fichier qui pourrait être utilisé pour des recherches documentaires ou pour des traitements bureautiques ultérieurs ?**

**En règle générale, la conversion au format PDF donne un rendu convenable qui est fidèle visuellement au document initial. C'est la qualité essentielle qui est demandée par l'utilisateur de fichier PDF et c'est la qualité minimale obtenue par tous les logiciels testés dans cette étude. On pourrait s'attendre à plus, en particulier avec le format PDF/A créé spécifiquement pour les besoins de l'archivage.**

Un niveau d'exigences est demandé pour que le fichier PDF contienne par exemple :

- des caractères en Unicode pour permettre une réutilisation des textes dans différentes langues ;

- des images sans compression ni déformation par rapport à l'image originelle avant conversion pour ne pas avoir de perte d'information ;
- la conservation de toutes les métadonnées préexistantes ;
- la structure du document avec un balisage ;
- des liens internes et externes opérationnels ;
- la police utilisée qui soit totalement incorporée dans le fichier converti ;
- le respect de la transparence de deux images superposées (c'est une fonctionnalité qui demande des calculs importants) non seulement au niveau visuel mais aussi lors de l'extraction de ces deux objets numériques ;
- etc …

Cela montre l'importance du contenu du document avant conversion car un même logiciel peut avoir une réaction différente si le fichier contient une police spécifique avec des caractères arabes ou chinois par exemple. Ce fichier sera peut-être converti en PDF/A mais le texte ne pourra peut-être pas être récupéré par un « copier/coller ».
Il y a toutefois une amélioration des logiciels de conversion au fur et à mesure de la progression de leur version.
Pour expliquer cela, il faut aussi prendre en compte l'évolution du format PDF : à partir de la version 1.7, les formats PDF sont normalisés.

Cette différence de traitement selon les logiciels témoigne des diverses interprétations possibles de la norme PDF par les fournisseurs de logiciel.
Le format PDF 2.0 devrait reprendre les définitions des fonctionnalités du PDF plus clairement pour permettre une homogénéité du format du fichier converti quel que soit le logiciel de conversion utilisé. Ceci devrait faire aussi évoluer le format PDF/A puisqu'une version PDF/A est définie à partir d'une version de PDF : PDF/A-1 créé à partir de PDF 1.4, PDF/A-2 à partir de PDF 1.7 et PDF/A-3 à partir de PDF 1.7 avec la prise en compte de la fonctionnalité de conteneur de fichiers.
Mais ce travail de clarification de la norme PDF prend plus de temps prévu. La version PDF 2.0, qui devait initialement être publiée fin 2012, ne sera vraisemblablement effective qu'en 2014 au plus tôt. Il va donc falloir attendre encore quelques années pour voir probablement des fichiers PDF valides et bien formés quel que soit le logiciel utilisé pour la conversion et la validation.

# Tests des fichiers TeX et LaTeX (édition scientifique)

## Présentation du format TeX

Largement utilisé dans le contexte de l'enseignement supérieur et de la recherche, le format TeX[7] est à la fois un langage de création de document et un format de fichiers, pour lequel il existe un grand nombre d'éditeurs. Riche et variable d'une version à une autre, sa particularité est qu'il est compilable. Le résultat de sa compilation génère un type de fichier différent selon l'outil utilisé.

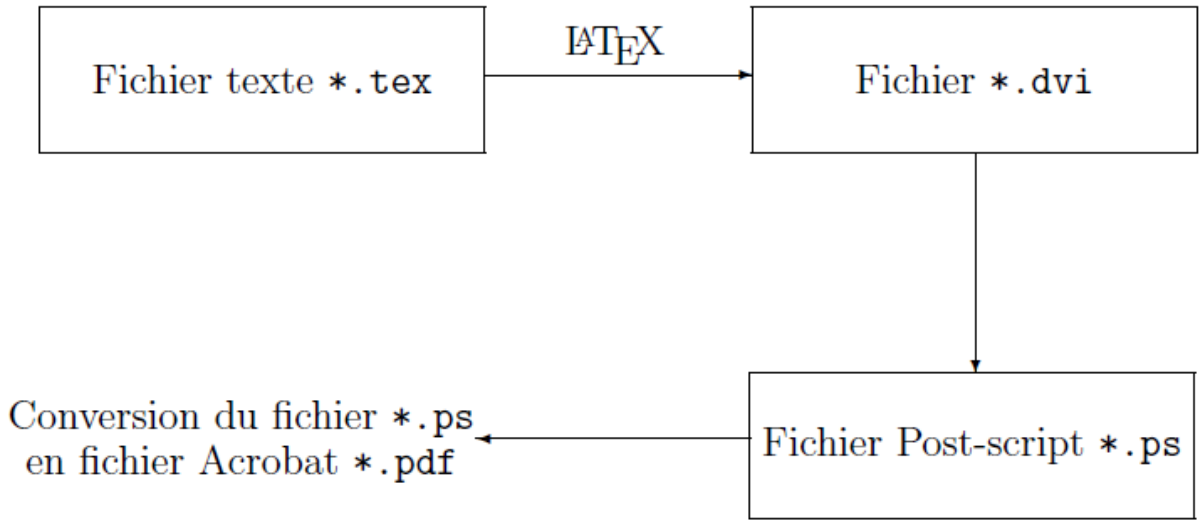



**Figure 19 :** Génération d'un document à partir d'un document .TeX

Deux méthodes permettent d'obtenir un fichier PDF à partir d'un fichier TeX. La méthode classique que la plupart des distributions TeX utilisent est illustrée ci-dessus par la Figure 19. Elle se fait en trois étapes : du TeX au DVI, du DVI au PS et du PS au PDF. L'autre méthode, quant-à-elle, consiste à passer directement du TeX ou DVI au PDF.

Le format DVI présente l'avantage d'être facilement maniable. En effet, les fichiers DVI contiennent des données binaires qui correspondent à la transcription visuelle du code provenant du fichier TeX source. Les distributions TeX intègrent par défaut des logiciels permettant de visualiser des fichiers dans ce format et il existe de nombreux pilotes et logiciels d'impression permettant d'obtenir du PS ou du PDF à partir du DVI.

Dans le cadre de cette étude, il est également important d'introduire la différence entre TeX et LaTeX. Comme défini ci-dessus, le TeX est un langage de programmation qui a aussi donné son nom au processeur de texte générant des fichiers de cette extension (.tex). LaTeX[8] est une bibliothèque de macros au même titre que ConTeXt[9] que l'on utilise dans le langage TeX.

[7] http://www.tug.org/begin.html
[8] http://www.latex-project.org/
[9] http://wiki.contextgarden.net/Main_Page

ConTeXt, en comparaison avec LaTeX, a la particularité d'être plus souple grâce à un plus grand nombre de macros prédéfinies et moins conflictuelles. De plus, elle est basée sur la notion de *plain text* (texte brut), du fait de la pluralité des langues qu'elle intègre. Les problèmes de caractères spéciaux y trouvent rapidement solution.

Cette étude concerne essentiellement le rendu en PDF des fichiers TeX contenant du LaTeX.
Les éditeurs TeX sont pour la plupart intégrés à des distributions qui contiennent aussi des compilateurs, des outils de conversion et des outils de visualisation. Même si cette étude concerne les seuls outils de conversion, les distributions dans lesquelles ils ont été testés sont néanmoins énumérées.

# Définition de la stratégie de tests

## Les outils testés

Le périmètre de cette étude a volontairement été restreint aux seuls outils capables de générer des PDF directement à partir du format TeX, sans passer par le format DVI, ce qui correspond à une minorité de logiciels. Cette restriction est due au fait que le format DVI présente les mêmes problèmes qu'un pilote d'impression qui, n'a pas accès aux métadonnées et à la structure du fichier.

| Outil de conversion | Distribution |
|---|---|
| pdfLatex | Winshell, TexWorks, TexnichCenter, TexMaker |
| pdfTex | TexWorks, TexLive |

**Tableau 4 :** Outils de conversion testés pour le TeX

Les tests effectués sur la génération de fichiers PDF avec ces outils montrent qu'ils produisent des fichiers valides indépendamment de la version du PDF choisie. La définition de la version du PDF qu'on souhaite obtenir se fait via la modification du fichier de configuration de l'outil concerné. Les dernières versions de ces outils produisent généralement du 1.4 par défaut.

## L'échantillon de fichiers de tests

Les fichiers utilisés dans le cadre de cette étude ont été fournis par l'ABES[10]. Il s'agit des fichiers sources de thèses réellement archivées au CINES en PDF.

| Type de contenu du document | Nombre de fichiers | Résultat obtenu |
|---|---|---|
| Mathématiques | 20 | PDF, PDF/A |
| Littérature | 3 | PDF, PDF/A |
| Fichier de test avec images et texte | 1 | PDF, PDF/A |
| Autre (sans image) | 4 | PDF, PDF/A |

**Tableau 5 :** Principales caractéristiques des fichiers TeX testés

[10] Agence Bibliographique de l'Enseignement Supérieur.

### Les fonctionnalités testées

Pour cette partie de l'étude, seul un sous-ensemble des fonctionnalités du PDF, jugées pertinentes dans le cas de fichiers TeX, ont été testées :

- les images ;
- la couleur ;
- les polices ;
- la transparence ;
- les métadonnées ;
- le sommaire ;
- la structure du document.

Pour plus de détails sur ces fonctionnalités, se reporter à la description donnée dans la partie « Tests des fichiers bureautiques ».

## La conversion en PDF/A

Il n'existe pas de solution standard pour obtenir du PDF/A à partir des outils testés. Plusieurs fichiers TeX avec des contenus différents ont été testés au cours de cette étude. Il en ressort qu'il n'existe pas encore de procédure permettant d'obtenir du PDF/A directement à partir du LaTeX. Les raisons varient généralement en fonction du contenu du document.
Voici un exemple d'erreurs récurrentes lorsqu'un PDF généré à partir du LaTeX est analysé avec l'outil Prépresse d'Adobe Acrobat Pro XI :

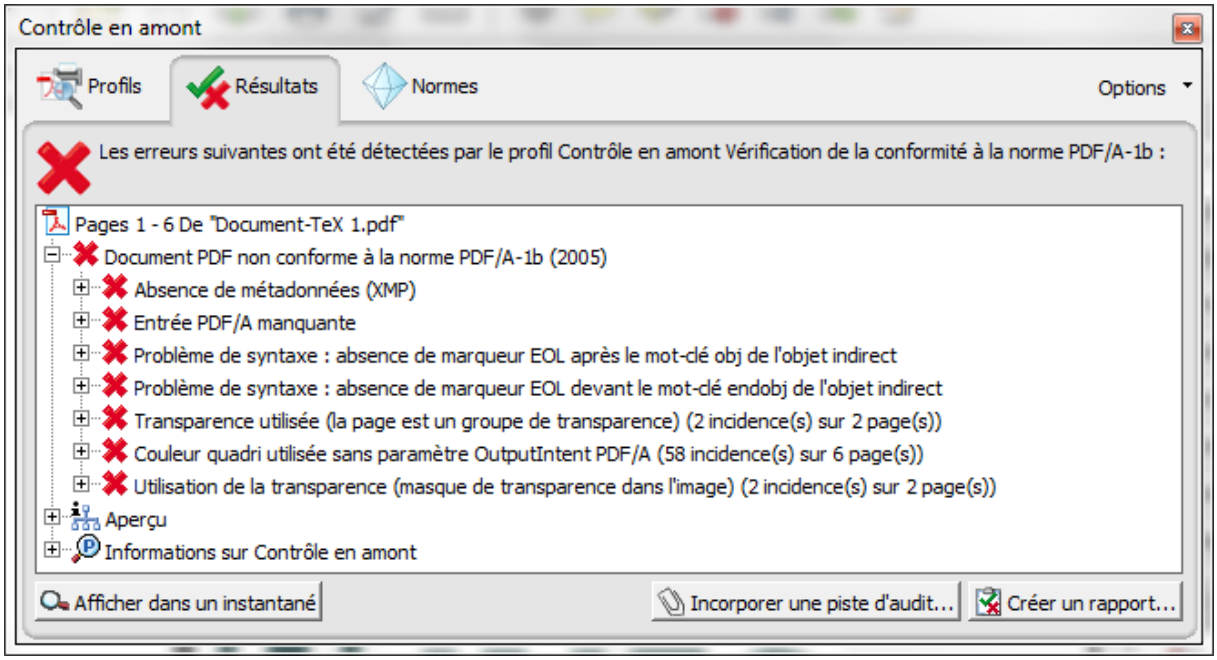



**Figure 20** : Liste des erreurs détectées par Adobe Acrobat Pro XI

- Métadonnées XMP : c'est le plus important de tous les problèmes rencontrés lors de la conversion TeX vers PDF. Soit les métadonnées ne correspondent pas au catalogue PDFInfo, soit il n'y en a pas assez.
- Problèmes de police : les polices utilisées en LaTeX n'ont pas toujours leur équivalent dans le PDF.

- Problèmes de syntaxe : en majeure partie dû au fait que le formatage des outils de conversion n'est pas (encore) accepté par PDF.
- Lorsque le fichier contient des images, ces images devraient respecter les spécifications du PDF/A pour prétendre satisfaire la norme.

Deux solutions ont pu être testées pour résoudre le problème majeur lié aux métadonnées XMP. L'une consiste à rajouter des informations dans l'entête du document. Le fichier TeX est compilé avec un fichier XMP contenant un certain nombre de tags, qui est partagé sur le site d'une communauté TeX, et un fichier ICM qui correspond au profil de couleurs que l'on va utiliser pour générer le fichier PDF. L'ensemble des fichiers testés avec cette méthode a permis de produire des PDF/A.
L'autre solution consiste à inclure la bibliothèque **pdfx** et à définir les métadonnées XMP, soit dans l'entête du fichier TeX, soit dans un fichier portant le même nom que le fichier source avec l'extension **xmpdata**. Exemple : *\usepackage [a-1b] {pdfx}*.

Les sources peuvent être récupérées à cette adresse :
http://support.river-valley.com/wiki/images/3/3f/Pdfa-supp.zip

Générer du PDF/A à partir du LaTeX n'est donc pas une question de logiciel ou de configuration. Si le contenu, après compilation, produit un fichier PDF qui respecte les spécifications basiques et contient des métadonnées XMP, il peut être reconnu comme du PDF/A.

# Bilan des tests par fonctionnalité du PDF

## Images

Le rendu des images incluses dans les documents LaTeX dépend surtout de la qualité initiale de l'image. La compilation d'un fichier source TeX avec les outils pdfLatex et pdfTex accepte plusieurs types d'images : JPEG, PNG, TIFF, etc. ; ce qui est un avantage pour cette méthode de conversion, contrairement à la méthode qui utilise un fichier DVI et qui présente plus de contrainte. Néanmoins, les images créées à l'aide de bibliothèques propres au LaTeX telles que **tikz** ont un meilleur rendu.

| Type image | | Qualité initiale | Qualité du rendu | |
|---|---|---|---|---|
| | | | œil nu | zoom |
| PDF | | Correcte | Correcte | Correcte |
| PNG | avec transparence | Basse (500x350) | Perte de qualité | Perte de qualité |
| | sans transparence | Haute (2000x1400) | Correcte | Correcte |
| JPEG | | Basse (500x350) | Perte de qualité | Perte de qualité |
| | | Haute (2000x1400) | Correcte | Correcte |

**Tableau 6** : Résultat des tests effectués sur la conversion des images

## Couleur

La restitution des couleurs est de bonne qualité. Les proportions de rouge, vert et bleu (RVB) sont conservées dans le fichier PDF telles qu'indiquées dans le fichier source TeX.

## Polices

La gestion des polices entre LaTeX et PDF reste encore un problème et un obstacle lorsqu'on souhaite obtenir du PDF/A. Acrobat Pro ne reconnaît qu'un seul type de police dans les PDF générés avec pdfTex ou pdfLatex : ***Type 1***.

## Transparence

LaTeX possède des bibliothèques permettant de créer des images. Le test à ce niveau consiste donc à superposer deux images de telle sorte que celle en dessous puisse être visible. Le rendu en PDF conserve bien cette superposition et chaque image est bien distincte.

## Métadonnées

Il s'agit de l'un des critères les plus importants pour la structure du fichier. Néanmoins les informations concernant l'auteur, le titre et autres ne sont pas restituées lors de la conversion. Les seules informations présentes sont la date de création du fichier pdf et le logiciel responsable de la conversion de TeX vers PDF.
Cependant, une solution consiste à utiliser le package **hyperref** qui permet de créer des hyperliens internes ou externes, des tables de matières automatiques et définir les métadonnées du document. Voici le bloc à insérer dans le document LaTeX :

```
\hypersetup{
plainpages=false,
colorlinks=true, linkcolor=black, anchorcolor=black,
citecolor=black, menucolor=black, urlcolor=black,
bookmarks=true, bookmarksopen=true, bookmarksnumbered=true,
pdftitle={Etude sur LaTeX},
pdfauthor={Franklin BOUMDA},
pdfsubject={Etude sur LaTeX},
pdfcreator={TeX}, pdfproducer={pdfTeX, pdfLateX },
pdfkeywords={LaTeX, cours}
}
```

Une autre solution utilisant le package **pdfx** consiste à rajouter les métadonnées dans l'entête du fichier comme suit :

```
\def\Title{LaTeX to PDF/A}
\def\Author{boumda@cines.fr}
\def\Subject{Un exemple de document}
\def\Keywords{LaTeX,Exemple,Document, PDF/A, PDF}
```

Et elles deviennent accessibles lorsqu'on rajoute le bloc **pdfinfo** suivant :

```
\pdfinfo{%
    /Title    (\Title)
    /Author   (\Author)
    /Subject  (\Subject)
    /Keywords (\Keywords)
    /Trapped  /False
}
```

Résultat :

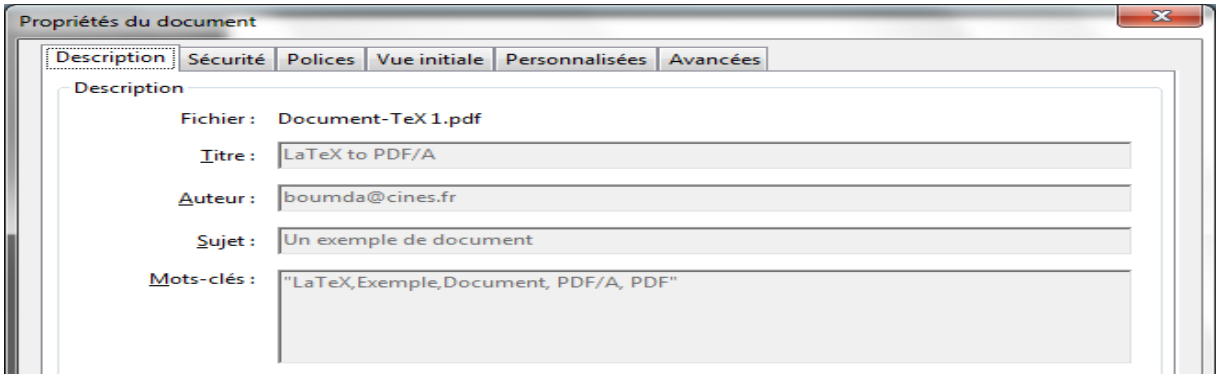



**Figure 21** : Métadonnées présentes dans le PDF généré

### Sommaire

Le sommaire est créé automatiquement suivant la structure du texte. S'il est constitué de liens cliquables, le rendu en PDF l'intégrera également.

### Structure du document

Il s'agit dans cette partie de vérifier que le fichier PDF généré restitue bien la forme du document. Le format PDF utilise un système de balisage, comme le HTML, de telle sorte que la structure du document peut être retrouvée. Le balisage distingue titres, paragraphes, listes, etc. Il n'a jamais été possible de retrouver ce balisage dans les fichiers PDF générés à partir du LaTeX. La structure du document est alors perdue.

Bien que l'étude ait porté sur les moyens d'obtenir du PDF valide à partir de LaTeX seulement, il est important de signaler qu'il existe d'autres méthodes qui permettent d'aboutir aux mêmes résultats. Il s'agit notamment, de passer par le format DVI à partir duquel on pourrait, soit générer directement du PDF, soit du PS qui sera ensuite transformé en PDF. En matière de qualité de rendu, l'étude réalisée ne permet pas de donner des résultats concrets de comparaison, mais ce sont des pistes qui pourront faire l'objet d'autres études.

# Conclusion

Cette étude a permis de dévoiler les différents moyens d'obtenir du PDF archivable à partir du TeX. **Le constat est que les solutions ne sont pas encore intégrées aux environnements de développement TeX.** Il y a là une opportunité à voir pour la communauté scientifique de réfléchir à l'évolution des outils de création afin qu'ils intègrent de manière performante ces aspects compte-tenu du volume important de données exprimées en format TeX.

# Tests des fichiers DWG (dessins techniques)

## Présentation du format DWG

Le format DWG est probablement le format le plus largement utilisé pour les dessins de Création Assistée par Ordinateur (CAO). Bien qu'étant le format natif des fichiers produits par le logiciel Autocad de la société Autodesk, ce format est également supporté par d'autres applications de Dessin Assisté par Ordinateur (DAO) ; ce qui en fait un format pivot pour les plans et dessins.
Par ailleurs, le consortium Open Design Alliance (ODA) réunit de nombreux éditeurs de logiciels de CAO pour faire la promotion d'un format DWG ouvert.
L'objectif de cette étude est de mettre en évidence les problématiques et les avantages de la conversion des fichiers DWG vers un format PDF valide et bien formé en vue d'un archivage pérenne.

Le format DWG ne comporte pas moins de 18 versions. Il permet de stocker des données et des métadonnées.
Un fichier DWG est constitué d'un « objet » complexe incluant le chemin absolu de fichiers externes appelés Xref, qui sont eux-mêmes des fichiers DWG. Cet « objet » constitue l'espace de travail. C'est à partir de cet objet (ou Model en anglais), de paramètres définis par l'utilisateur dans le logiciel (échelle calques, etc.) et de références externes que va être construit le plan à imprimer, appelé aussi pour les versions supérieures à AutoCAD 2002 des « présentations » (Layout en anglais).
Dans un visualiseur, les onglets représentent l'objet (unique) et ses présentations.

- L'onglet « objet » :

Il est constitué d'un dessin ou d'un cartouche. On peut définir dans l'objet une zone d'impression. C'est en principe le dessin original sur la base duquel vont être créées les présentations. Certains objets n'ont pas de zone d'impression définie et ne peuvent être imprimés sans intervention manuelle. Il peut aussi y avoir une zone par défaut et dans ce cas précis, seule la partie du dessin se trouvant dans la zone va être imprimée.

- L'onglet « présentation » :

Il est constitué d'un dessin ou d'un cartouche et, comme il est en principe destiné à l'impression, d'une zone définie pour l'impression.

- Le dessin :

Il est créé avec les outils propres du logiciel de DAO : création de plans, de calques, inclusion d'images .png, de texte .pdf, .dwf (format pdf d'AutoCAD), etc. Le dessin est une superposition de plusieurs couches.

- Les calques :

Un dessin peut être considéré comme une association de plusieurs couches. Ces couches ou calques sont des composants spécifiques, comme l'ensemble des murs extérieurs d'un bâtiment, le réseau électrique, les cotes, etc.

- Le fichier Xref :

C'est un fichier .dwg. Il s'agit souvent de fichiers communs à un ensemble de plans afin d'éviter la redondance d'informations. Il est référencé dans le fichier DWG « maitre » par son chemin absolu.
Plusieurs problèmes se posent quant à ces fichiers de références. Outre leur possible perte, un autre problème subsiste s'ils existent : il est lié au chemin absolu des répertoires inscrit « en dur » dans le .dwg. Une solution est par exemple d'inclure les références externes dans les « objets » ou dans les « présentations » avant de faire les conversions ou de les déplacer dans le même répertoire.

# Définition de la stratégie de tests

## Les outils testés

Il existe sur le marché plusieurs logiciels qui convertissent les fichiers DWG en PDF. Dans le cadre de cette étude, le choix a été fait de tester le logiciel propriétaire AutoCAD et son visualiseur gratuit DWG TrueView et trois logiciels open source : AutoDWG, Any DWG to PDF, et TotalCAD.
AutoCad a été choisi car c'est le logiciel de référence du format dwg, les trois autres car ce sont des logiciels libres parmi les plus utilisés.

| Nom | Version | Editeur du logiciel | PDF généré |
|---|---|---|---|
| **AutoCAD/traceur** | 2013 | Autodesk | 1.5 |
| **/Export** | | | 1.6 |
| **Any DWG to PDF Converter** | 2013 | Any DWG software | 1.3 |
| **AutoDWG Convertisseur PDF 2013** | 4.60 2003~2013 | AutoDWG DWG2PDF Converter 2013 | 1.4 |
| **TotalCAD Converter** | 4.1.22 | CoolUtils | 1.4 |

**Tableau 7 :** Outils de conversion testés pour le DWG

Sur aucun de ces logiciels, il n'est possible de choisir une version PDF autre que celle imposée par le logiciel.

*Remarque* : Dans le cadre de ces tests, les trois logiciels open source ont été achetés car les versions téléchargeables gratuitement incluaient un ensemble de logiciels additionnels « indésirables » et parfois très difficiles à supprimer du poste de travail (Hola search par exemple).

Lors de cette étude, un autre logiciel de conversion pertinent a pu être identifié. Il s'agit de la suite logicielle propriétaire d'Incitius[11] : iFILES, iVIEW, iPDF.
Ces logiciels n'ont cependant pu être testés en raison de leur coût d'acquisition (supérieur à 15 000 euros) mais une présentation a permis de voir qu'une conversion de masse des fichiers DWG peut être faite. Les fichiers en sortie sont valides et bien formés et les possibilités de paramétrage très puissantes. Le format de sortie est de l'iPDF, un format propriétaire mais répondant à tous les critères de la norme PDF.

## L'échantillon de fichiers de tests

Les tests ont été effectués sur 5 versions différentes de DWG :

- AC1009 correspond à des dessins issus d'un export généré à partir du logiciel Allplan ;
- AC1012 et 1014 correspondent aux plans du lycée le Corbusier à Aubervilliers. (Pour ces fichiers il est probable qu'ils aient été remaniés car la date de création des fichiers semble antérieure à la version constatée).
- AC1018 correspond à des plans trouvés sur internet.
- AC1027 correspond à des plans du bâtiment du CINES produits par le logiciel AutoCAD.

Certains des fichiers DWG testés ont été créés il y a longtemps. Ils correspondent à des versions anciennes, à des fichiers plusieurs fois remaniés ou encore à des fichiers dont l'historique a été perdu ; ce qui pose un certain nombre de problèmes pour l'archivage.

[11] www.incitius.com

Pour ces fichiers, il n'y a pas de garantie qu'ils aient été produits par le logiciel Autocad puisque le DWG est supporté par plusieurs autres logiciels de DAO.
De plus, les références externes associées aux fichiers ont parfois été perdues. Il en est de même pour les polices, les échelles, etc.

| Format d'entrée | Version du format | Versions d'AutoCAD correspondantes |
|---|---|---|
| **DWG R11/12** | AC1009 | AutoCAD Release 11, AutoCAD Release 12 |
| **DWG R13** | AC1012 | AutoCAD Release 13 |
| **DWG R14** | AC1014 | AutoCAD Release 14 |
| **DWG 2004** | AC1018 | AutoCAD 2004, AutoCAD 2005, AutoCAD 2006 |
| **DWG 2013** | AC1027 | AutoCAD 2013 |

**Tableau 8 :** Versions des fichiers DWG testés

## Les fonctionnalités testées

Les fichiers DWG possèdent des caractéristiques propres à ce format. Par conséquent, les fonctionnalités qui ont été testées lors de cette étude sont différentes de celles définies pour les fichiers de type bureautique. Elles sont présentées ci-dessous.

### Gestion des fichiers de références

Les fichiers DWG utilisent des références externes (pas de redondance de l'information, mise à jour des plans facilitées). Les logiciels peuvent lister ou inclure ces fichiers au moment de l'ouverture d'un .dwg. L'intégration des Xref peut être faite « manuellement » ou « automatiquement », si ces derniers sont dans les mêmes répertoires.
Pour satisfaire à un archivage pérenne, il est important que tous les fichiers de références puissent être archivés, soit dans leur « présentation », soit dans une bibliothèque dédiée.

### Gestion des objets et présentations

La sortie PDF peut inclure tous les dessins ou se faire par « objet » et/ou « présentations ».

### Options de sortie

Lors d'une conversion de DWG en PDF, il peut être nécessaire de régler certains paramètres en fonction des besoins :

- taille de la zone d'impression : si le fichier a été convenablement configuré à sa création, les zones d'impression des dessins sont automatiques, sinon une opération de cadrage manuelle est requise ;
- résolution ;
- couleur ;
- polices : inclusion de fichiers de polices ou personnalisation
- options de tracé : importation de fichiers de tracé (CTB) ou défaut.

### Gestion des calques

Les fichiers dwg sont constitués (en partie) de calques. La gestion de ces calques est importante puisqu'elle permet une « dissection horizontale » fine des plans. Selon la finalité de l'archivage, les calques sont indispensables. Dans le cas où une « image » globale du plan est suffisante il n'est pas indispensable de les gérer.
Les différents paramétrages possibles pour la gestion des calques sont :

- l'affichage des calques avant conversion ;

- la possibilité de retirer des calques avant la conversion ;
- ou encore la présence des calques dans l'onglet de la sortie PDF.

### Inclusion de métadonnées

Il est important dans le cadre de l'archivage qu'il soit possible d'enregistrer des métadonnées dans le PDF avant l'impression.

### Versions de DWG supportées

Les versions de DWG prises en charge varient selon les logiciels de conversion testés. Il est donc important de connaître pour chaque logiciel testé, les versions de DWG supportées.

### Gestion des onglets PDF

Le format PDF permet l'affichage d'un certain nombre d'onglets pour peu que les informations soient récupérées au moment de la conversion. L'étude réalisée vérifie la présence de signets, de vignettes, de calques et de métadonnées dans les onglets PDF.

### Visualisation du fichier DWG

Les logiciels possèdent-ils un visualiseur qui permet la visualisation des fichiers lors de l'ouverture dans le logiciel ?

### Aperçu avant impression

Pour faciliter le paramétrage de la conversion, l'existence d'une visualisation des fichiers avant impression est un atout, surtout quand les zones d'impression ne sont pas définies.

### Traitement par lot

La fonctionnalité de traitement par lot des fichiers a également été testée dans cette étude.

# Bilan des tests par logiciel

## Any DWG to PDF

Ce logiciel génère un fichier PDF v1.3 en sortie.

**Gestion des fichiers de références**
Le logiciel inclut automatiquement les fichiers de références s'ils sont présents dans le même répertoire mais il ne les liste pas. L'absence de mention des fichiers de références ne permet pas de distinguer le dessin « maître » du dessin affiché. La seule façon de savoir s'il y a des fichiers de références est de sortir le fichier de son contexte (son répertoire) et de l'imprimer.

**Gestion des objets et présentations**
La sortie PDF peut inclure tous les dessins ou se faire par « objet » et/ou « présentations ».

**Options de sortie**

- Taille de la zone d'impression **:** ajustement automatique du format de sortie ou paramétrage possible par tâtonnement !
- Choix des polices : inclusion de fichiers de police ou personnalisation
- Importation de fichiers de tracé (CTB) possible en plus des fichiers proposés par défaut par le logiciel

**Gestion des calques**

- Pas d'affichage des calques avant conversion
- Pas de possibilité de retirer des calques avant conversion
- Pas de présence des calques dans l'onglet de la sortie PDF

**Inclusion de métadonnées** : oui (auteur, titre, sujet, mots-clés)

**Versions de DWG supportées**
Versions R2.5/2.6, R9, R10, R12, R13, R14, R2000/2002, R2004/2005/2006, 2007/2008/2009, 2010/2011/2012, 2013.

**Gestion des onglets PDF**

- Présence de signets qui portent le nom de l'objet ou présentation imprimé (pas le nom du fichier)
- Présence de vignettes

**Visualisation du fichier DWG**
Pas de visualiseur

**Aperçu avant impression**
Pas d'aperçu avant impression

**Traitement par lot**
Le logiciel permet le traitement par lot, seulement en version pro

*Remarque :* Le logiciel est capable de détecter les versions « imprimables ». Par exemple, pour certains fichiers DWG, il ne détecte que l'objet ou que la présentation si celle-ci à une zone d'impression définie. Dans la mesure où il n'y a pas de visualiseur, on ne sait pas à l'avance ce qui est imprimable. Le logiciel gère cet aspect automatiquement et les onglets PDF Signet et Vignettes sont justes.

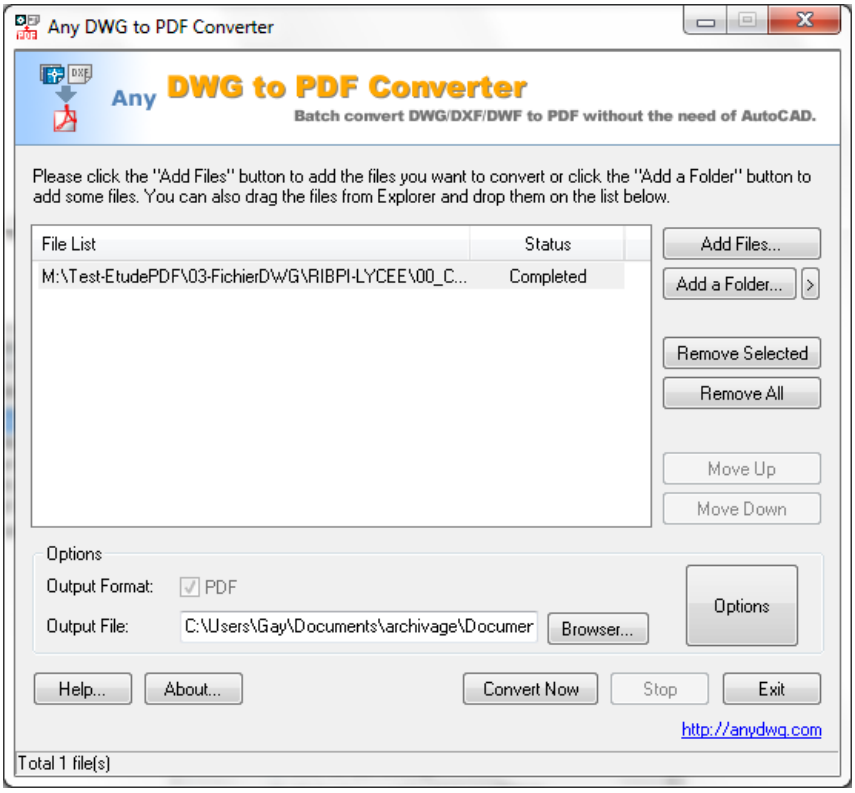



**Figure 22** : Page d'accueil du logiciel Any DWG to PDF

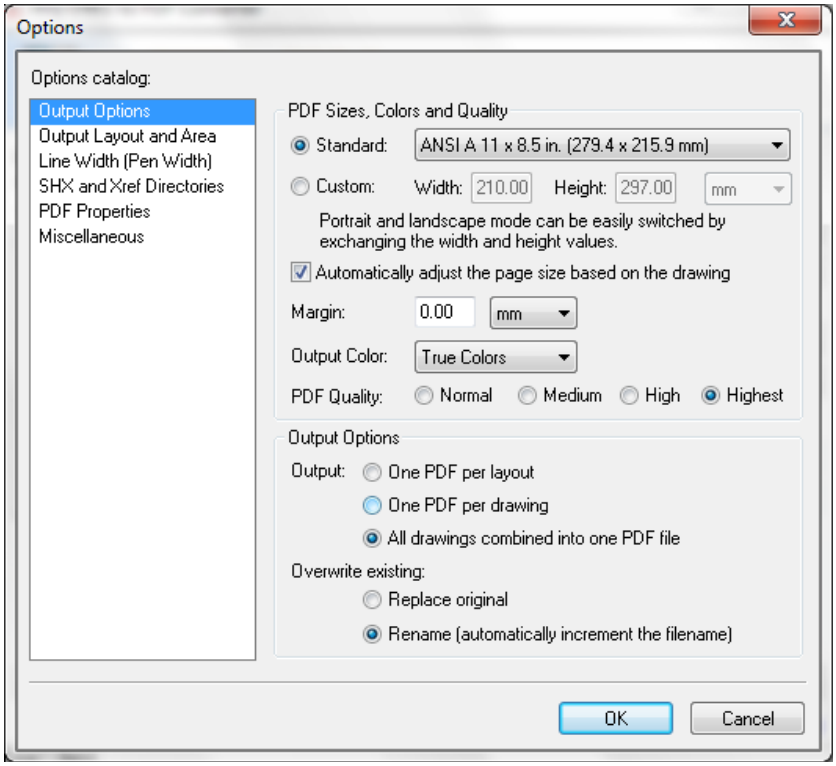



**Figure 23 :** Options de sortie du logiciel Any DWG to PDF

## AutoDWG DWG2PDF

Ce logiciel génère un fichier PDF v1.4 en sortie, valide et bien formé.

**Gestion des fichiers de références**
Le logiciel inclut automatiquement les fichiers de références s'ils sont présents dans le même répertoire.
Avant la conversion, il affiche un message qui liste les références manquantes :

> Warning:============================================================
> The Xref Drawing ax_log.dwg can not be found; maybe some information will be lost. or copy the dwg file and its xref files in a same folder.
> The Xref Drawing site.dwg can not be found; maybe some information will be lost. or copy the dwg file and its xref files in a same folder.
> The Xref Drawing axesb.dwg can not be found; maybe some information will be lost. or copy the dwg file and its xref files in a same folder.
>
> ============================================================

**Gestion des objets et présentations**
La sortie PDF peut inclure tous les dessins ou se faire par « objet » et/ou « présentations ».
Par défaut, le logiciel détecte les dessins imprimables.

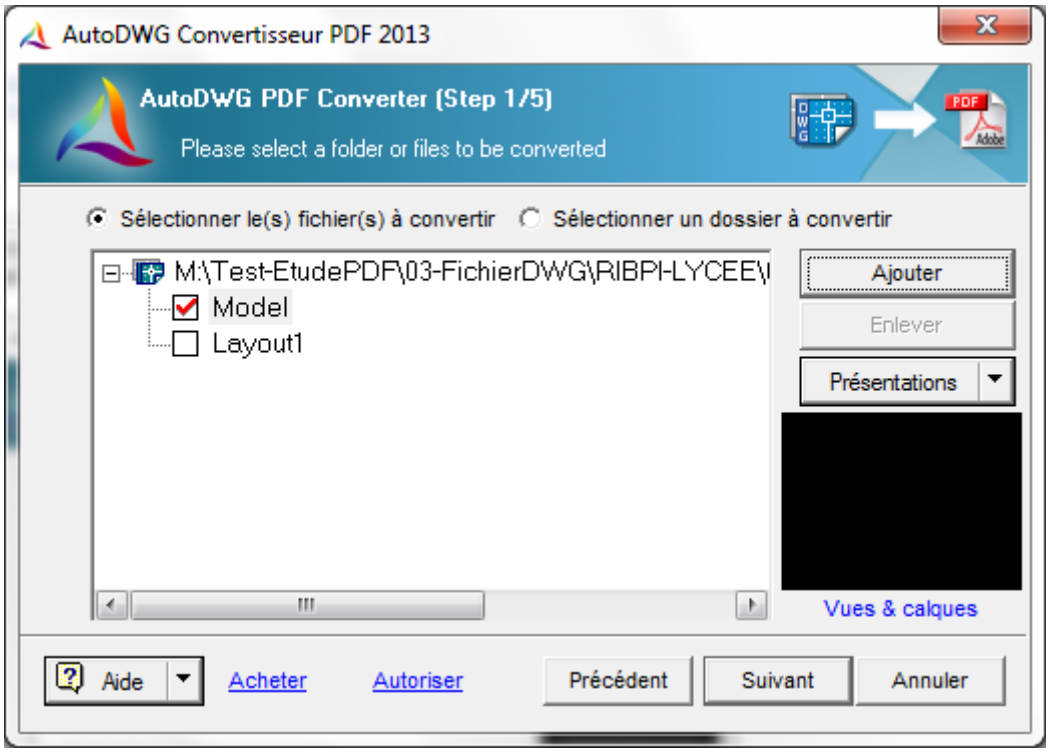



**Figure 24 :** Visualisation des objets et présentations contenues dans le fichier

**Options de sortie**

- Taille : ajustement automatique du format de sortie (lineweight scale : =1) ou choix possible
- Résolution couleur : ajustement automatique du format de sortie ou choix possible
- Choix des polices : inclusion de fichiers de police ou personnalisation
  Importation de fichiers de tracé (CTB) possible en plus des fichiers proposés par défaut par le logiciel

**Gestion des calques**

- Affichage des calques avant conversion
- Possibilité de retirer des calques avant conversion
- Pas de présence des calques dans l'onglet de la sortie PDF

**Inclusion de métadonnées** : non

**Version de DWG supportées**
Versions R2.5/2.6, R9, R10, R12, R13, R14, R2000/2002, R2004/2005/2006, 2007/2008/2009, 2010/2011/2012, 2013.

**Gestion des onglets PDF**

- Présence de signets qui portent le nom du fichier et le nom de l'objet ou présentation imprimé
- Présence de vignettes

**Visualisation du fichier DWG**
Visualisation complète de tous les dessins (objet et présentations) si les zones d'impression sont définies

**Aperçu avant impression**
Pas d'aperçu avant impression

**Traitement par lot**

Le logiciel permet le traitement par lot, seulement en version pro

*Remarque* :
Le logiciel est capable de détecter les versions imprimables. Par exemple pour certains fichiers dwg il ne détecte que l'objet ou que la présentation. Dans la mesure où il y a un visualiseur, on sait à l'avance ce qui est imprimable, le logiciel gère cet aspect automatiquement et les signets sont justes.

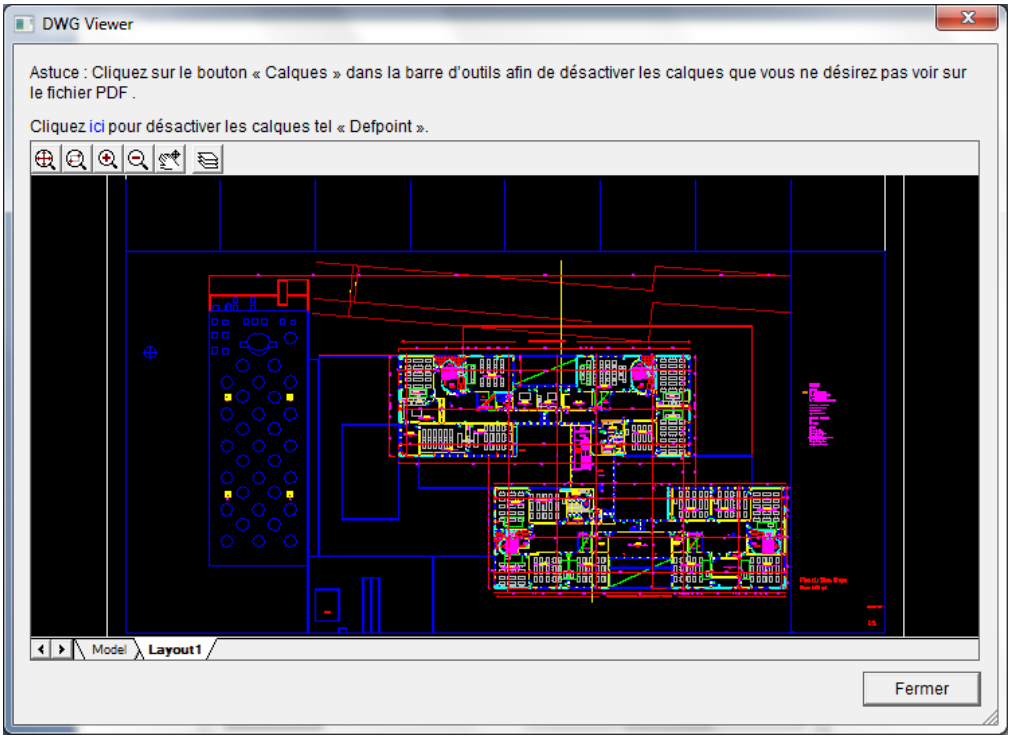



**Figure 25 :** Visualiseur

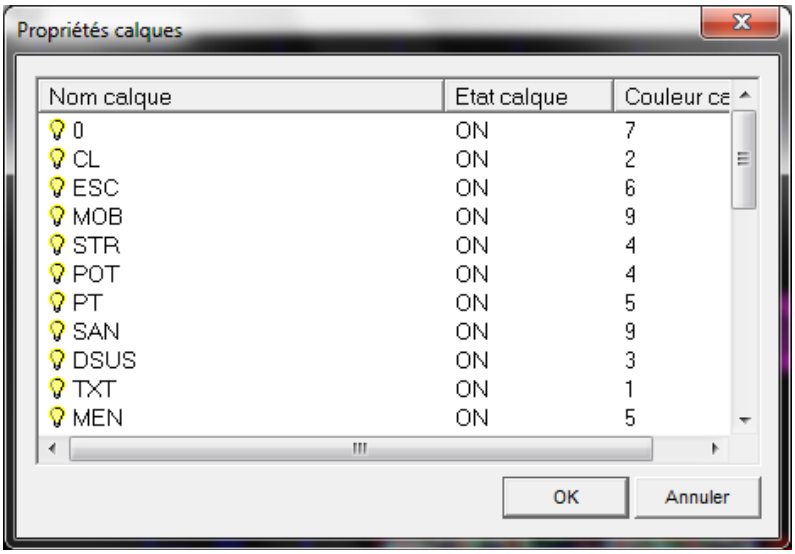



**Figure 26 :** Option du visualiseur pour afficher la liste des calques

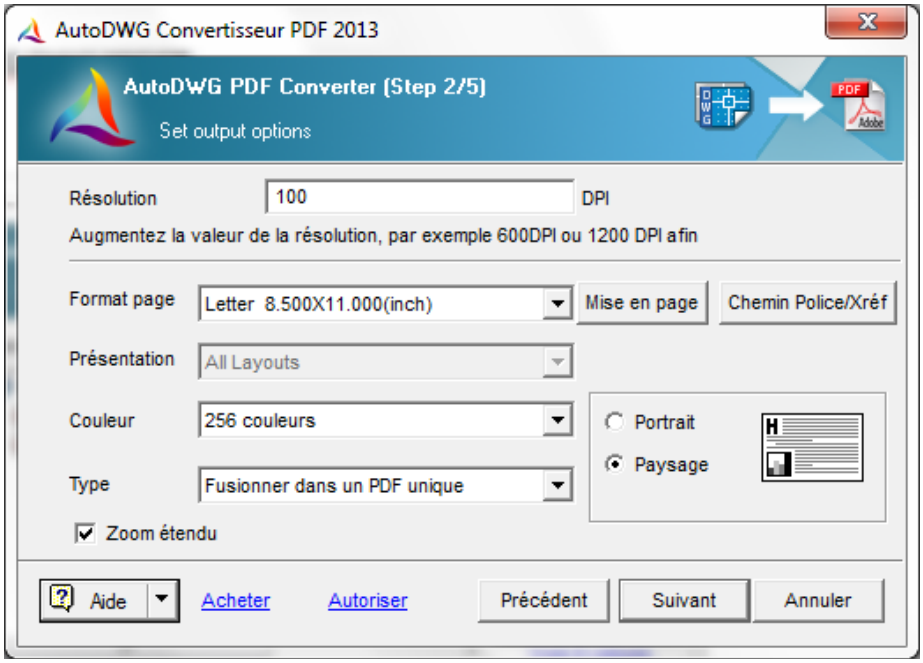



**Figure 27 :** Options de sortie

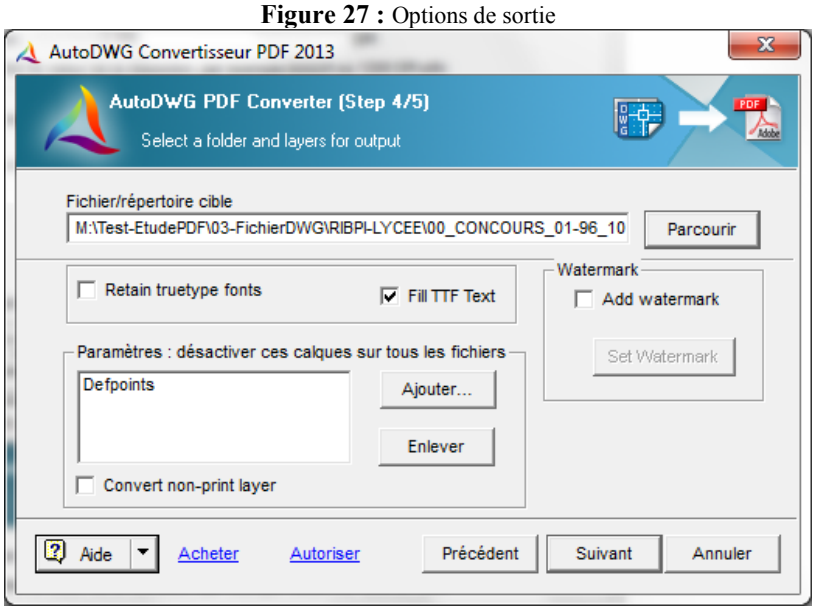
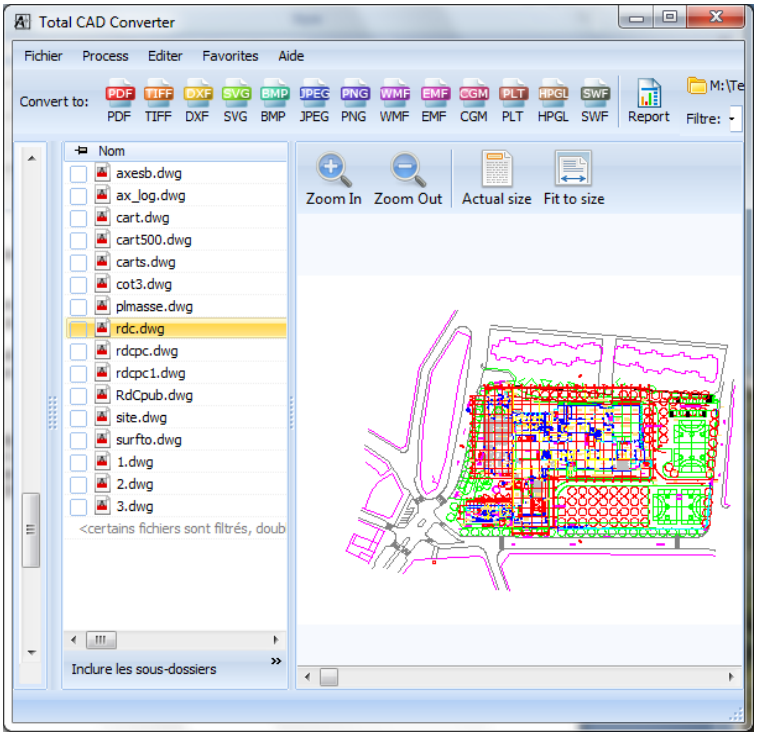



**Figure 28 :** Options de sortie pour l'emplacement d'enregistrement du fichier

## TotalCAD Converter

Ce logiciel génère un fichier PDF v1.4 en sortie, valide et bien formé.

**Gestion des fichiers de références**
Le logiciel détecte automatiquement les fichiers de références s'ils sont présents dans le même répertoire et les affiche en visualisation, mais il ne les imprime pas et ne les liste pas.




Ici le fichier rdc.dwg et ses fichiers de références ont été sélectionné. Le visualiseur à droite montre une image complète (axesb.dwg, ax_log.dwg et site.dwg sont les trois fichiers Xref présents dans le répertoire).

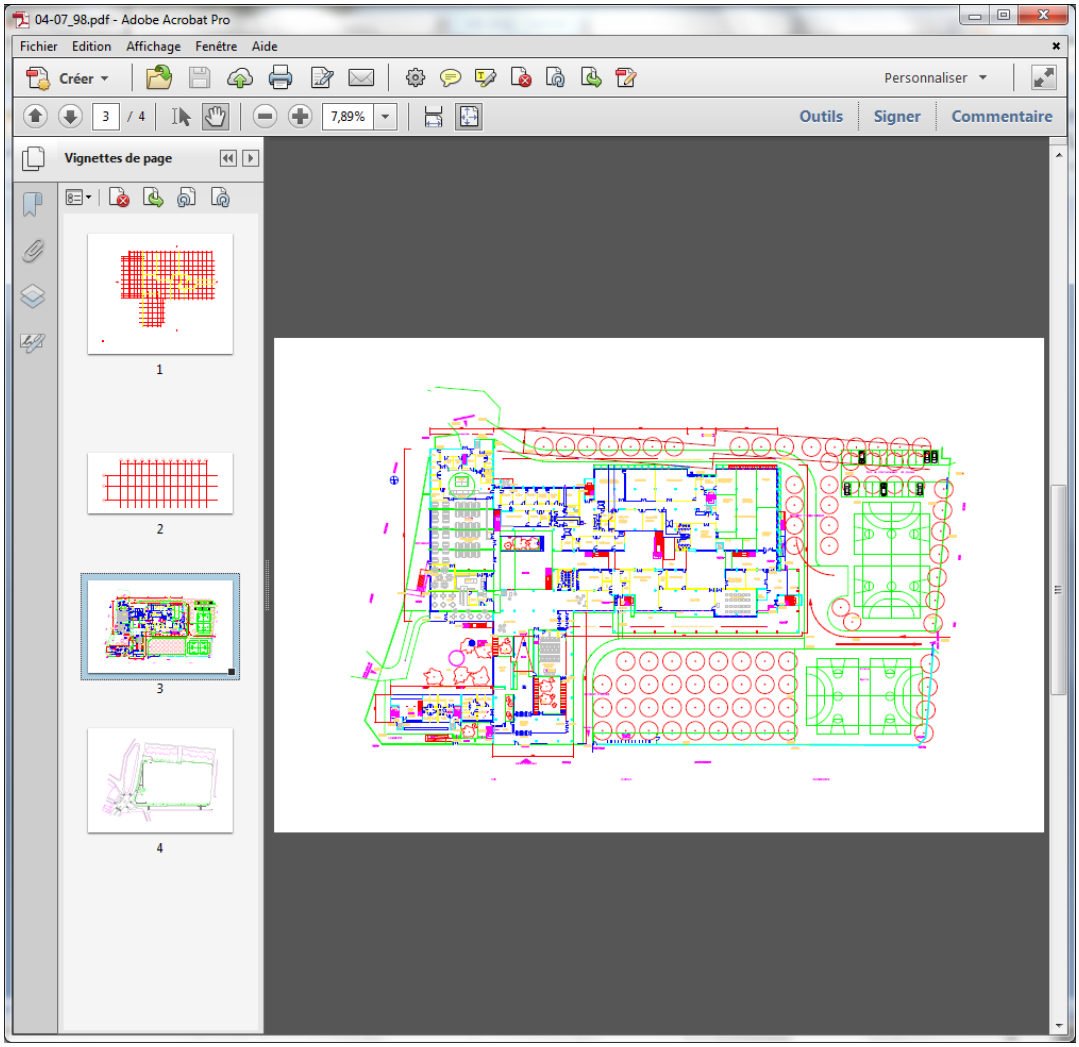
**Gestion des objets et présentations**

Bug dans le logiciel pour les fichiers DWG avec présentations multiples

A l'impression, un seul dessin est exporté : le dernier dessin ! Le visualiseur affiche une image différente de ce qui va être imprimé.

Si le nom des fichiers de références est connu, il est possible de les cocher et de faire une impression de plusieurs images dans un même PDF. Ils sont alors distingués dans le PDF de sortie et non fusionnés dans une seule « image ».

La sortie PDF peut inclure tous les fichiers DWG d'un répertoire. Mais elle n'imprime qu'un seul dessin par fichier (le dernier).




Ici le fichier converti en PDF montre les 3 fichiers de références et le rdc.dwg initial.

Si l'on sélectionne un répertoire complet, tous les objets et toutes les représentions sont imprimés, mais il y a une perte totale des signets et des noms des dessins.

**Options de sortie**

- Taille : ajustement automatique du format de sortie (autosize) ou choix possible
- Résolution couleur : ajustement automatique du format de sortie ou personnalisation (minimale)
- Choix des polices : personnalisation
- Pas d'importation de fichiers de tracé (CTB)

**Gestion des calques**

- Pas d'affichage des calques avant conversion
- Pas de possibilité de retirer des calques avant conversion
- Pas de présence des calques dans l'onglet de la sortie PDF

**Inclusion de métadonnées** : oui (auteur, créateur, objet, titre, mots-clés)

**Versions de DWG supportées**
Versions R2.5/2.6, R9, R10, R12, R13, R14, R2000/2002, R2004/2005/2006, 2007/2008/2009, 2010/2011/2012, 2013

**Gestion des onglets PDF**

- Présence de signets pour la conversion d'un fichier .dwg unique (attention : il affiche tous les signets mais il n'y a qu'une vignette) ; pas de signets en cas d'impression multiple de fichiers
- Présence d'une seule vignette par fichier

**Visualisation du fichier DWG**
Visualisation de l'objet uniquement. Attention la visualisation n'est pas forcément l'impression !

**Aperçu avant impression**
Pas d'aperçu avant impression

**Traitement par lot**
Le logiciel permet le traitement par lot, seulement en version pro

*Remarque :* Les PDF sont « bruts » mais l'avantage est que l'on peut archiver un répertoire complet avec l'ensemble des dessins en un seul clic. Par contre, il n'y a qu'un dessin imprimé par fichier. Les présentations multiples sont perdues.

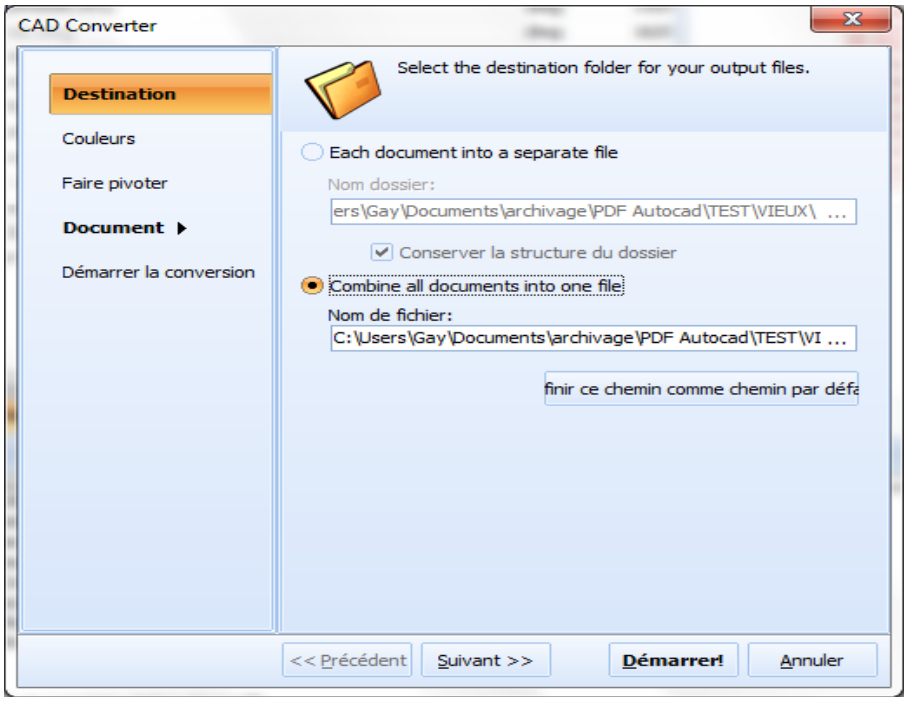



**Figure 29 :** Choix de l'impression dans un ou plusieurs fichiers

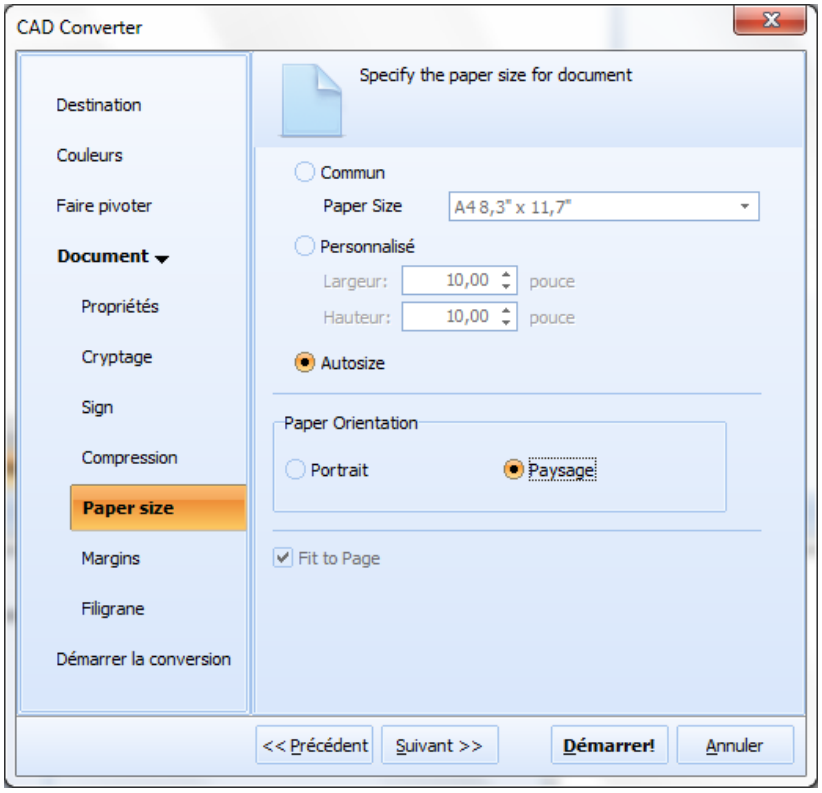



**Figure 30** : Options de taille de sortie

## AutoCAD 2013

Les onglets de sortie du logiciel AutoCAD permettent deux générations distinctes de fichiers PDF. Cette étude a privilégié l'option Export car l'option Traceur est moins performante.

### Onglet « Traceur » :

La sortie « DWG to PDF pc3 » génère un PDF v1.6, valide et bien formé avec calques et nom de la présentation dans les signets identiques à l'onglet export PDF que nous présentons ci-dessous :

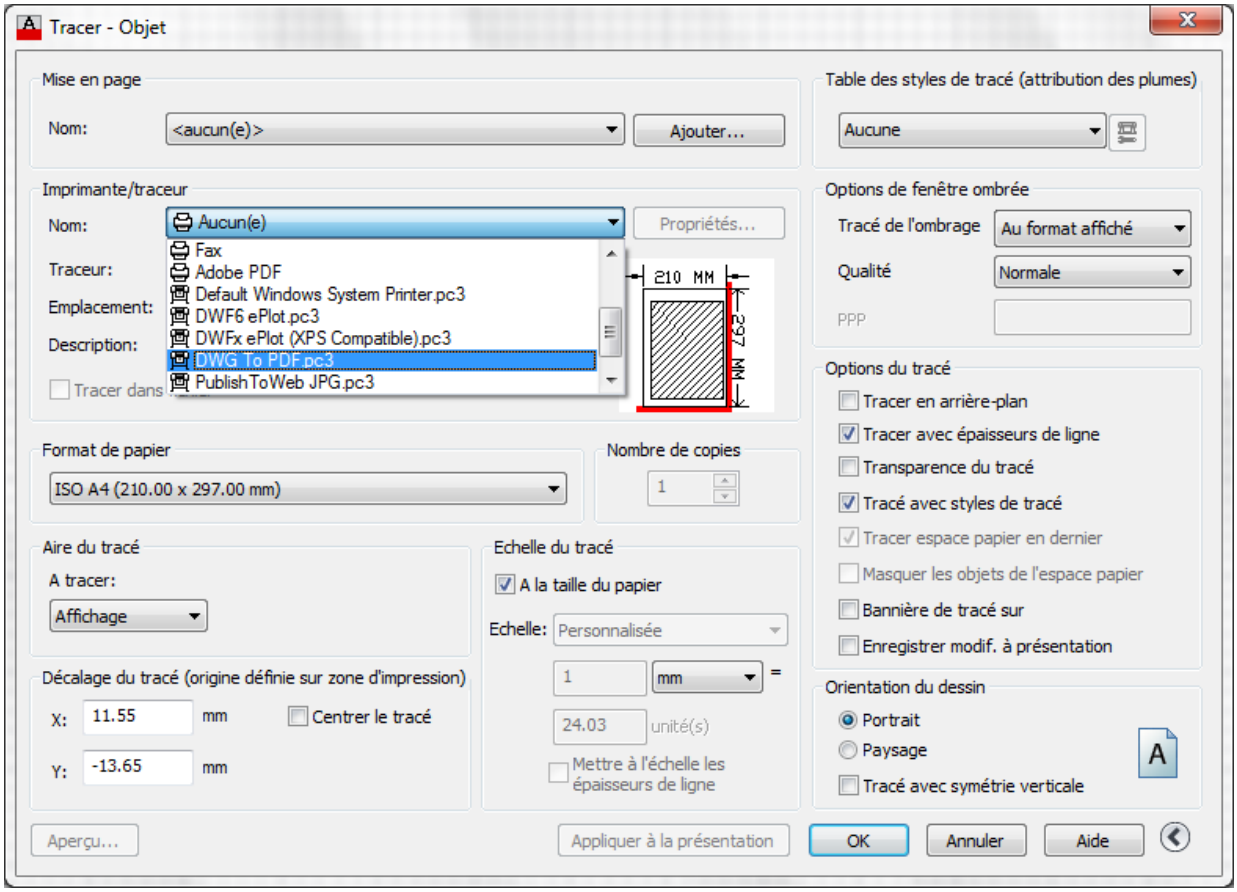



**Figure 31** : Imprimantes

## Onglet « Export PDF » :

Le logiciel génère un PDF v1.6 valide et bien formé avec inclusion de tous les calques des dessins.

**Gestion des fichiers de références**
Le logiciel inclut automatiquement les fichiers de références s'ils sont présents dans le même répertoire. S'ils ne sont pas présents, il en donne une liste et permet s'ils sont connus de les intégrer.

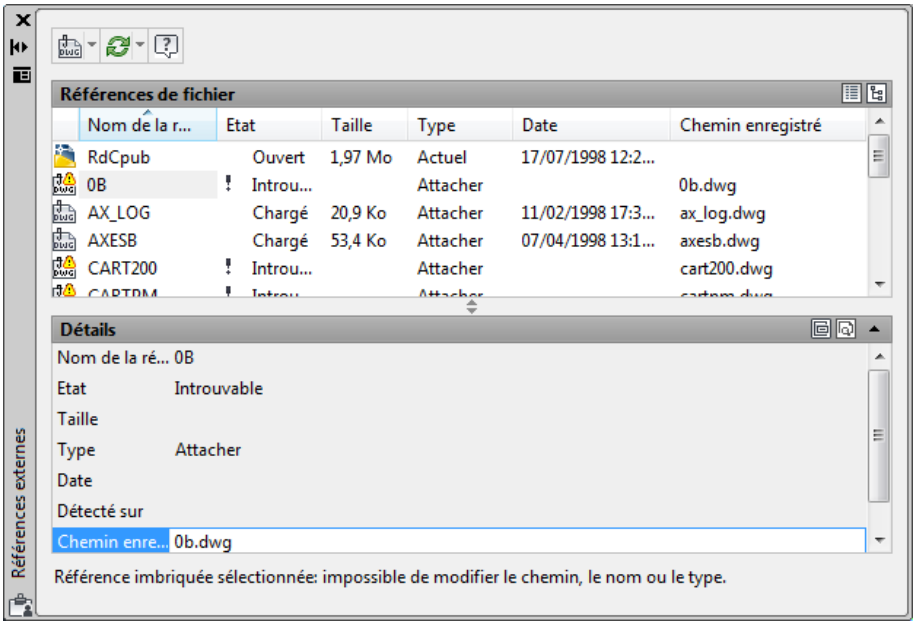



**Figure 32 :** Liste des fichiers de références

**Gestion des objets et présentations**
La sortie PDF peut inclure toutes les présentations ou chaque dessin individuellement.

**Options de sortie**
- Taille : ajustement automatique du format de sortie ou choix possible
- Choix des polices : automatique
- Fichiers de tracé (CTB) donnés par défaut

**Gestion des calques**
- Affichage des calques avant conversion
- Possibilité de retirer des calques avant conversion

**Inclusion de métadonnées**
Les métadonnées Titre, Sujet, Auteur, Mots clés et Commentaires se retrouvent le PDF (Fichier/Propriétés).

**Version de DWG supportées**
Toutes

**Gestion des onglets PDF**
- Pas de signets
- Présence de vignettes
- Présence des calques dans l'onglet de la sortie PDF

**Visualisation du fichier DWG**
Présence d'un visualiseur

**Aperçu avant impression**
Aperçu avant impression

**Traitement par lot**
Fonctionnalité non testée pour ce logiciel dans le cadre de cette étude.

*Remarque :* L'export de l'objet ou des présentations nécessite un paramétrage manuel si les zones d'impression n'ont pas été correctement définies afin que l'objet apparaisse dans sa totalité.
Autocad est le seul logiciel à fournir l'onglet des calques dans le fichier PDF et à intégrer aussi ceux des fichiers de références.

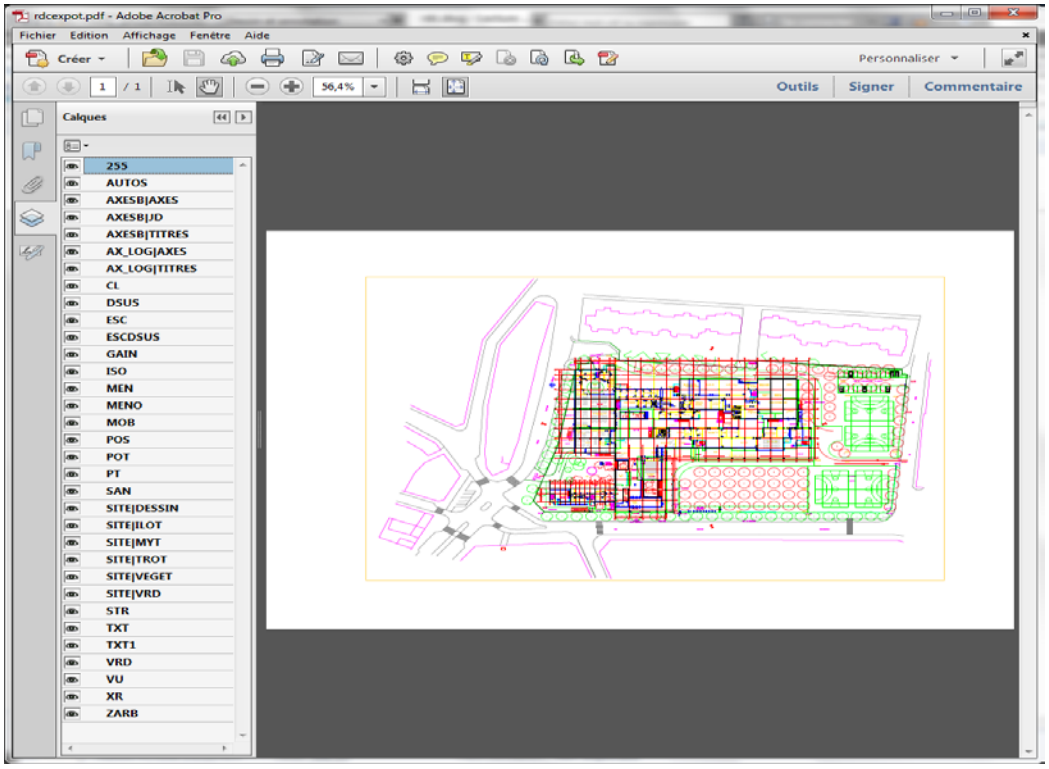



**Figure 33 :** Exemple de PDF v1.6 avec la liste des calques ( xxx|yyy correspond aux calques des fichiers de références)

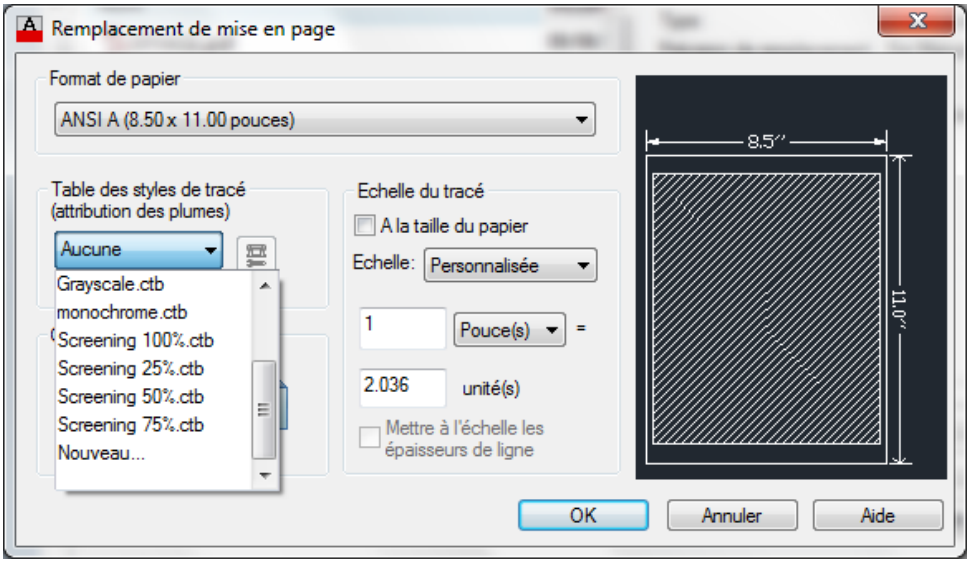



**Figure 34 :** Options de taille de sortie

# Tableau récapitulatif par fonctionnalité

| | | AutoCAD | AnyDWG | AutoDWG | TotalCAD |
|---|---|---|---|---|---|
| **Visualisation à l'ouverture** | **Dessin** | ✓ | ✗ | ✓ | ✓ |
| | **Calques** | ✓ | ✗ | ✓ | ✗ |
| **Impression des fichiers de références** | **Si présents dans répertoire** | ✓ | ✓ | ✓ | ✗ |
| | **Ajout manuel** | ✓ | ✗ | ✗ | ✗ |
| **Métadonnées PDF** | | ✓ | ✓ | ✗ | ✓ |
| **Visualisation avant impression** | | ✓ | ✗ | ✗ | ✗ |
| **Onglets PDF** | **Vignettes** | ✓ | ✓ | ✓ | ✓ |
| | **Signets** | ✗ | ✓ | ✓ | ✓ |
| | **Calques** | ✓ | ✗ | ✗ | ✗ |
| **Options de sortie** | **Polices** | ✓ | ✓ | ✓ | ✓ |
| | **Tracés** | ✓ | ✓ | ✓ | ✓ |
| **Version de PDF générée** | | 1.6 | 1.3 | 1.4 | 1.4 |

# Conclusion

La complexité du format DWG nécessite, dans le cadre d'une conversion, un travail en amont très important.
Plusieurs questions se posent :

- Quels sont les formats DWG ?
- Que veut-on réellement conserver : une image ou un plan à l'échelle avec ses calques ?
- Comment ont été créés les fichiers et avec quel logiciel ?
- A-t-on les échelles sur les dessins ?
- Les fichiers de références existent-ils encore ? Si oui où sont-ils ?
- Doit-on fusionner tous les fichiers dans une même librairie afin d'avoir accès à tous les fichiers de références ? Attention toutefois, des fichiers Xref portant le même nom et placés dans des bibliothèques différentes sont parfois des fichiers différents.
- Doit-on faire des doublons pour que chaque librairie ait la totalité les fichiers de références qui lui sont nécessaires (quand on les connaît) ?
- Les zones d'impression sont-elles correctes ? Impossible de le vérifier si l'on n'ouvre pas les fichiers avec un logiciel adéquat.
- Que doit-on privilégier : les épaisseurs du trait ou la lisibilité du dessin ? Pendant les tests, il s'est avéré que la lisibilité du dessin se faisait au dépend de la qualité des épaisseurs des traits (zoom catastrophique, tâches de couleur indistinctes).

La perte des références externes pose un premier problème. Si des fichiers externes existent, un autre problème subsiste lié au chemin des répertoires inscrits « en dur » dans le DWG. Une solution est d'inclure les références externes dans les « objets » ou dans les « présentations » avant de faire les conversions.

Les formats de sortie ne sont pas toujours automatiques et la mise à l'échelle peut s'avérer complexe.
Le choix des paramètres peut varier d'un fichier à l'autre. Le manque de visualisation avant impression accentue la difficulté d'affinage du paramétrage.
Dans le cas d'impression de masse dans un seul fichier PDF, le paramétrage « commun » entraîne des différences d'échelle entre les diverses présentations et peut tronquer un dessin, voire le faire disparaitre.
La complexité d'affichage liée aux fichiers de références est difficile à gérer. Chaque dessin est caractérisé par sa zone d'impression qui n'englobe pas forcement la totalité du dessin. Si on ne refait pas une zone d'impression, on peut perdre le dessin.

Les fichiers AC1009 (AllPlan) et 1018 issus d'export de logiciels de CAO différents d'Autocad ont posé problème. AutoCAD n'imprimait que l'objet alors que les logiciels Any DWG to PDF et AutoDWG affichaient l'objet et sa présentation (deux dessins identiques).

AutoCAD crée un fichier PDF dans lequel tous les calques du fichier DWG sont distincts y compris les calques des fichiers de références. Il est alors possible de cocher et décocher les calques dans le fichier PDF.
Cette option qui semble incontournable, ne se trouve que dans l'Export AutoCAD.

**Au vu de ces tests, l'option de la conversion en PDF des fichiers DWG semble indiquée uniquement pour garder une image de dessins de plans aboutis où toutes les références seraient incluses, les zones d'impression proprement définies, les cartouches et échelles inclus dans les dessins et les calques conservés.**

La conversion au format PDF demande un travail en amont assez considérable et une bonne documentation sur les fichiers. Le travail doit être confié à quelqu'un maîtrisant les logiciels de DAO et si possible ayant participé à l'élaboration des plans ou du moins à leur suivi. Les outils sur le marché peuvent faire un bon travail au cas par cas, mais une conversion de masse semble plus qu'hasardeuse, au risque de perdre de l'information, voire le dessin dans sa totalité.

# Conclusion générale de l'étude

La réalisation de cette étude sur les outils de conversion du format PDF s'est avérée relativement complexe compte-tenu de la problématique posée. Le format PDF est largement utilisé afin de recevoir le contenu produit par des logiciels nombreux et hétérogènes (notamment de par leurs formats de fichiers). L'objectif premier a été de définir un périmètre de tests répondant aux besoins de la communauté à l'origine de cette étude.

Contrairement à ce que l'on aurait pu penser au premier abord, la question de la conversion de fichiers vers le format PDF n'est pas triviale. Quels que soient les types de documents testés (fichiers bureautiques, édition scientifique, dessins techniques), les traitements préalables à la conversion (paramétrages) ne sont pas négligeables et peuvent difficilement être effectués pour un lot de fichiers.

Dans le cas des fichiers de traitement de texte, la démarche à suivre pour la conversion PDF est très liée au type de contenu présent dans le fichier et à l'usage que l'on souhaite faire du PDF (fonctionnalités à préserver).
En effet, bien que les logiciels de conversion puissent générer des PDF valides et bien formés avec un rendu visuel correct, aucun des outils testés ne gère, sans perte d'informations, l'ensemble des fonctionnalités proposées par le format PDF. Il est donc important de déterminer en amont de la conversion, quelles sont les fonctionnalités essentielles à préserver en fonction des objectifs visés et ensuite de choisir l'outil de conversion et la version de PDF la plus appropriée. Si malgré tout, l'objectif principal de l'archivage est d'obtenir un fichier PDF/A, cette conversion étant plus exigeante, le choix de convertir en PDF/A-1, PDF/A-2 ou PDF/A-3 peut déterminer le choix du logiciel. En effet, les tests ont révélé que les logiciels payants sont les plus adaptés pour produire les formats récents de PDF/A.

Dans le cas des fichiers aux formats TeX et DWG, le traitement par lot des fichiers est tout simplement inenvisageable, du fait même de la nature des fichiers. En effet, un document au format TeX regroupe un ou plusieurs fichiers qui doivent être compilés ensemble dans le logiciel pour produire un seul fichier PDF. Cette opération ne peut être réalisée pour plusieurs documents à la fois.

Pour ce qui est des fichiers DWG, les paramètres d'impression à définir dépendent tellement des fichiers et de l'objectif visé lors de la conversion (plan entier, quelques vues choisies, etc.) que l'opération requiert une intervention manuelle conséquente ainsi que la maîtrise du format DWG et de l'outil utilisé. Au-delà d'un traitement par lot, c'est l'idée même d'une conversion en PDF qui peut être rediscutée pour l'archivage des dessins techniques (bien qu'elle puisse faire sens dans un contexte de diffusion) ne serait-ce que parce qu'elle supprime toute possibilité d'exploitation du fichier dans le futur.

Dans tous les cas, il est important de penser à la conversion des fichiers le plus en amont possible du cycle de vie du document et d'associer le producteur à cette démarche. Il est en effet le plus à même de connaître les fonctionnalités importantes à préserver dans ses fichiers.

Outre les résultats obtenus, cette étude propose également une méthodologie de tests qui peut être facilement adaptée et réutilisée par quiconque veut poursuivre les tests, sur d'autres logiciels de conversion, d'autres versions des logiciels déjà utilisés dans l'étude ou d'autres formats de fichiers PDF, afin d'enrichir la réflexion de manière collaborative.

# Annexe : Liste de convertisseurs

Dans le cadre de cette étude, une liste de convertisseurs a été identifiée. Elle s'est basée sur une étude des outils utilisés pour la création des 60 000 fichiers PDF présents dans la plateforme d'archivage PAC du CINES, ainsi que sur des produits particulièrement connus ou pour lesquels l'éditeur participe aux travaux de normalisation PDF, ce qui peut laisser envisager un meilleur respect de la norme.

La liste ci-dessous les présente globalement des plus répandus aux moins répandus, avec quelques aménagements toutefois liés à l'organisation en catégories.

Un prix arrondi est mentionné à titre indicatif, sachant que certains ont des licences en fonction de l'utilisation (gratuit en utilisation personnelle ou en période de test, payant en utilisation commerciale).

Un astérisque est positionné devant les noms de logiciels produits par des entreprises qui participent activement aux travaux de normalisation PDF.

## 1. Convertisseurs de Type 1 : Plug-in du logiciel original

- Fichiers PDF issus de suites de logiciels bureautiques (traitement de texte, tableur, présentation) :
  - **OpenOffice.org** (http://www.openoffice.org/) — version de StarOffice maintenant maintenue par Apache. On peut aussi mentionner LibreOffice (http://fr.libreoffice.org/) du Document Foundation qui est une autre branche de développement d'OpenOffice.
  - ***Microsoft Office** (http://office.microsoft.com/) — suite bureautique de Microsoft.
  - **StarOffice** — suite bureautique développée par Sun (maintenant Oracle) et maintenant arrêtée en faveur de OpenOffice.org.
- Fichiers PDF issus de logiciels de conversion :
  - ***Adobe PDF Maker** — ajoute des fonctionnalités aux produits Microsoft Office pour la création de fichiers PDF. [une partie d'Acrobat 400€]
  - **ScanSoft PDF Create** (http://www.nuance.fr/for-business/by-product/pdf/pdf-create7/) — logiciel de conversion de fichiers en PDF (avec fonctionnalités spécifiques Microsoft Office) [49€].
- Fichiers TeX convertis :
  - **pdfTeX** (http://www.tug.org/applications/pdftex/) — extension de TeX qui génère du PDF directement au lieu du fichier DVI.
  - **dvipdfm** (http://gaspra.kettering.edu/dvipdfm/) — convertisseur de DVI vers PDF.
  - **dvips** (http://www.radicaleye.com/dvips.html) — convertisseur de DVI vers PostScript (qui peut être converti en PDF après)
  - **MiKTeX** (http://miktex.org/) — version moderne de TeX pour Microsoft Windows avec une sortie PDF miktex-pdftex.
  - **pdfeTeX** — version de e-Tex (des extensions à TEX) qui peut produire des fichiers PDF en plus des fichiers DVI
  - **PDFLaTeX** — couche sur pdfTeX pour utiliser LaTeX.
- Fichiers PDF générés par des logiciels de PAO :
  - ***InDesign** (http://www.adobe.com/products/indesign.html) [1000€]
  - **QuarkXPress** (http://www.quark.com/) [1400€ ]

### 2. Convertisseurs de Type 2 : Conversion depuis le fichier source

- Conversion bureautique :
  - ***Callas pdfaPilot** (http://www.callassoftware.com/) — conversion de documents Microsoft Office et OpenOffice directement en PDF/A. [400€]
  - ***Foxit PhantomPDF** (http://www.foxitsoftware.com/products/phantomPDF/) — conversion de documents Microsoft Office en PDF. [30€]
  - **Moyea PPT to PDF convertor** (http://www.dvd-ppt-slideshow.com/ppt-pdf-converter/) — conversion de fichiers PowerPoint en PDF. [gratuit]
  - **3-Heights Document Converter** (http://www.pdf-tools.com/pdf/document-converter-service-office-pdf-pdfa.aspx) — solution d'entreprise pour la conversion de documents bureautiques en PDF. [6000€]
  - **All File to PDF** (http://www.officeconvert.com/advanced-word-to-pdf-converter.htm) — convertisseur de documents bureautiques en PDF. [$50]
- Conversion PAO :
  - **Enfocus InstantPDF** (http://www.enfocus.com/fr) — conversion de documents Adobe CS ou QuarkXPress en PDF [6000€]
- Conversion CAO :
  - **DWG to PDF Converter** (http://anydwg.com/dxf-to-pdf-ex.html) [$83]
  - **Convert DXF to PDF** http://www.coolutils.com/dxf-to-pdf [$100]
  - **AutoDWG** http://www.autodwg.com/dxf-to-pdf.htm [$100]
  - **DWG to PDF Converter MX** http://www.dwgtool.com/dwg-to-pdf.htm [$100]

### 3. Convertisseurs de Type 3 : Pilote d'impression

- **Mac OS Quartz PDFContext** — Quartz 2D est le moteur graphique de Mac OS X. [une partie de MacOS]
- ***Adobe PDF** (http://www.adobe.com/products/acrobatstandard.html) — composant d'Acrobat qui s'installe comme pilote d'impression pour la création de PDF. [une partie d'Acrobat 400€]
- **pdfFactory** et **pdfFactory Pro** (http://fineprint.com/) — pilote d'impression pour la création de fichiers PDF (40€ ou 80€)
- **BCL easyPDF** (http://www.pdfonline.com/) — pilote d'impression Windows qui génère du PDF ($95).
- **Jaws PDF Creator** (http://www.jawspdf.com/) — pilote d'impression Windows qui génère du PDF (49€). Ce produit ne marche pas sur Windows 7 et il est proposé en remplacement des solutions pour entreprise de Global Graphics Software (http://www.globalgraphics.com/).
- **PDF Converter** — Plusieurs logiciels répondent à ce nom. Par exemple, le logiciel PDF Converter Enterprise distribué par la société Nuance (http://www.nuance.fr/) [40€]
- **PrimoPDF** (http://www.primopdf.com/) — pilote d'impression Windows qui génère du PDF. [gratuit]
- **PDFCreator** (http://www.pdfforge.org/pdfcreator) — pilote d'impression Windows qui génère du PDF. [gratuit]
- **Agfa Apogee Create Normalizer** — système de workflow pour la génération de PDF normalisé à partir d'un pilote d'impression. [sous licence]
- **deskPDF Professional** (http://www.docudesk.com) — pilote d'impression Windows pour la création de PDF. [$30]
- **OX PDF Creator** (http://www.oxpdf.com/pdf-creator.html) — pilote d'impression Windows pour la création de PDF. [$26]

### 4. Convertisseurs de Type 4 : Convertisseur PostScript

- **Ghostscript** (http://www.ghostscript.com/) avec Gsview, MacGSview, muPDF — interpréteur pour PostScript et PDF. [gratuit]
- **pstopdf** — outil intégré dans MacOS pour la conversion de PostScript en PDF (/usr/bin/pstpdf). [une partie de MacOS]
- ***Adobe Distiller** (http://www.adobe.com/products/acrobatstandard.html) — composant d'Acrobat qui permet de convertir des fichiers PostScript en PDF. [une partie d'Acrobat 400€]
- **PStill** (http://www.wizards.de/~frank/pstill.html) — convertisseur PostScript pour Windows/MacOS/Linux pour la création de PDF. [shareware, 20€, 40€]

### 5. Autres moyens de conversion :

- Fichiers PDF construits à partir d'applications qui utilisent des bibliothèques :
  - **PDFlib** (http://www.pdflib.com/) — bibliothèque écrite en C, disponible en versions 32-bit et 64-bit et qui peut être utilisée avec PHP, Java, .NET, C et C++ etc. [800€]
  - **FPDF** (http://www.fpdf.org/) — bibliothèque PHP pour la création de fichiers PDF. [gratuit]
  - **ClipPDF** (http://www.fastio.com/) — bibliothèque en ANSI C qui permet de créer des fichiers PDF. Il comporte une interface PHP. [gratuit pour utilisation personnelle, sinon 1000€]
  - **iText** (http://itextpdf.com/) — bibliothèque de fonctions pour la création de fichiers PDF en Java ou C#. [gratuit ou sous licence]
  - ***Adobe PDF Library** (http://www.adobe.com/devnet/pdf/library.html) — bibliothèque de fonctions C/C++ pour la création et la manipulation de fichiers PDF. [sous licence]
- Fichiers PDF créés depuis le format XSL-FO :
  - **Apache FOP** (http://xmlgraphics.apache.org/fop/) *Formatting Objects Processor* — outil pour mettre en page des documents décrits par XSL-FO. [gratuit]
- Fichiers PDF issus de logiciels dans les imprimante/scanners (matériel) :
  - **Xerox WorkCentre Pro** — gamme d'imprimantes/scanners capables de numériser un document et en créer un fichier PDF.
- Fichiers PDF issus de logiciels OCR :
  - **Omnipage** (http://www.nuance.fr/for-individuals/by-product/omnipage/) [99€]
  - **ABBYY FineReader** (http://france.abbyy.com/) [99€]
- Des sites web proposant la création de PDF :
  - http://www.conv2pdf.com/
  - http://www.freepdfconvert.com/
  - http://www.ps2pdf.com/ (GhostScript)

L'étude a aussi identifié d'autres outils qui ne sont pas des créateurs de PDF, mais qui permettent de manipuler des fichiers PDF :

- **StampPDF** (http://www.appligent.com/stamppdf-batch) — outil qui permet d'ajouter du texte ou des images à un fichier PDF
- **PSNormaliser** — il n'a pas été possible d'identifier le fournisseur de ce logiciel.
- **Appligent AppendPro** (http://docs.appligent.com/) — outil pour fusionner et manipuler des fichiers PDF (comme ajouter du texte par exemple).
- **Apex PDFWriter** — il n'a pas été possible d'identifier ce logiciel.