Fichiers PDF et indexation par les moteurs de recherche

Note de JL Archimbaud (http://jl.archimbaud.free.fr/) avril 2014

Cette note donne quelques informations sur l'indexation des fichiers texte en format PDF par les moteurs de recherche

Contexte

Les fichiers texte sont généralement mis en ligne au format PDF. On désire qu'ils soient correctement indexés par les moteurs de recherche. Or il y a plusieurs formats PDF et de nombreux outils différents sont utilisés pour les générer. Le texte ci-dessous, compilation de messages suite à une question que j'ai posée dans une liste de diffusion, donne quelques informations à ce sujet.

Question initiale

J'aimerais que les fichiers PDF d'un site soient correctement indexés par les moteurs de recherche grand public et par SOLR le moteur de recherche interne.
Il me semble que pour un fichier texte ou présentation (transparents), le PDF généré peut être différent et ceci peut avoir la conséquence d'être bien ou mal indexé par les moteurs de recherche.

- 1. Est-ce que je me trompe ?
- 2. Quel est le comportement de LibreOffice ?

Réponse de Michel Bouloc (CINES) que j'ai résumée

- Pour 1 : je me trompe. Les outils de traitement de texte, à partir d'un document texte ou d'une présentation texte génèrent généralement un PDF dont le contenu texte est indexable par les moteurs de recherche.
- Pour 2 : donc LibreOffice génère un PDF indexable.

Compléments, principalement de Michel Bouloc (CINES), que j'ai résumés

- Lorsque l'on scanne un document texte on obtient un PDF 'image' donc le contenu texte n'est pas indexable, uniquement les méta-données le sont.
- Le format PDF est maintenant très vaste pouvant insérer des fichiers qui n'ont rien à voir avec le document initial par exemple.
- Il y a différentes versions du format PDF. Les versions normalisées sont 1.7 et A (comme archivage). Mais LibreOffice, de même que Office de Microsoft génèrent des PDF 1.4 ou 1.5, ce qui ne pose pas de réel problème. Le format A présente l'avantage de pouvoir être lu par le maximum de lecteurs PDF. Par contre il est restrictif, donc si on a inséré des effets (objet transparent...) ils sont supprimés.
- A partir de LibreOffice, lorsqu'on fait un Export PDF, on peut choisir des options PDF comme le format A, peut-être à recommander mais restrictif.
- Un éditeur de texte tel que NodePad sous windows permet de visualiser le contenu réel d'un fichier PDF (dump) mais il n'est pas facile de savoir si le fichier contient du texte ou des images... En revanche, on peut voir la version du format PDF du fichier. Remarque personnelle : si on peut faire un copier-coller ou une recherche de mot dans un éditeur de fichier PDF il me semble que ça veut dire que c'est un format texte.
- Le CINES propose un service en ligne gratuit et ouvert à tous pour vérifier qu'un fichier

PDF (et aussi d'autres formats) est correct http://facile.cines.fr/. Mais pour le PDF il n'indique pas si c'est du texte ou des images.

- Un document décrit les différentes versions du format PDF (109 pages) : http://www.huma-num.fr/ressources/guide-methodologique-le-format-de-fichiers-pdf/13-aout-2012
- La doc Adobe du format PDF 1.7 (1310 pages) : http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/pdf_reference_1-7.pdf

Compléments de Nicolas Limare (CMLA)

- Un PDF produit par LaTeX+dvips+ps2pdf ou LaTeX+dvipdf n'est pas correctement indexable, car le fichier dvi ne contient pas le texte sous forme de chaine de caractères, uniquement sa représentation visuelle. Un PDF produit par pdfLaTeX contient, lui, le texte, et est bien indexable.
- Le programme "pdftotext" permet d'extraire le texte d'un document PDF. S'il échoue à vous donner une texte correspondant à peu près au contenu attendu, alors il est probable que les moteurs de recherche ne feront pas mieux.
- Pour lire et modifier les meta-données, j'utilise "pdftk". Les metadonnées essentielles (titre, auteurs, résumé, mots-clé) sont pris en compte par les moteurs de recherche.