

Les formats de documents ouverts

François Pelletier

25 novembre 2017

Les différents formats de documents ouverts

OpenDocument

- ▶ Format ouvert de données pour la bureautique.
- ▶ Basé sur la notation XML, assez proche du HTML
- ▶ Plusieurs types de documents:
 - ▶ Traitement de texte
 - ▶ Tableur
 - ▶ Présentation
 - ▶ Diagramme
 - ▶ Base de données
- ▶ Norme publiée par OASIS

OpenDocument: Avantages

- ▶ Interopérabilité: OpenOffice.org, LibreOffice, KOffice, Google Documents, IBM Notes
- ▶ Microsoft Office ne gère pas bien le format OpenDocument afin de lui donner une impression de mauvaise qualité.
- ▶ Abri contre la péremption des données
- ▶ Permet un formatage riche du contenu
- ▶ Permet l'automatisation de la production de documents

OpenDocument: Inconvénients

- ▶ Les fichiers sont lourds
- ▶ Il peut être difficile d'éditer le code XML directement sans corrompre le document
- ▶ L'utilisation des logiciels d'édition nécessite une période d'apprentissage et beaucoup de pratique.
- ▶ Intégration limitée de contenu scientifique tel que des équations ou des diagrammes

OpenDocument: Éditeurs libres

- ▶ OpenOffice.org
- ▶ LibreOffice
- ▶ NeoOffice
- ▶ KOffice
- ▶ AbiWord

Installation de Abiword:

```
sudo apt-get install abiword
```

OpenDocument: Références

- ▶ OASIS Open Document Format for Office Applications (OpenDocument) TC
- ▶ Document Freedom Day

L^AT_EX

- ▶ L^AT_EX est un système de préparation de document.
- ▶ Constitue un ensemble de macros développées par Leslie Lamport servant à faciliter l'utilisation de T_EX, le langage créé par Donald Knuth.
- ▶ Conçu pour la production de documents techniques et scientifiques.

L^AT_EX : Avantages

- ▶ Standard reconnu par la plupart des publications scientifiques.
- ▶ Permet de ne pas avoir à se soucier de l'apparence finale du document et de se concentrer sur le contenu.
- ▶ Permet de structurer efficacement de très grands documents
 - ▶ Table des matières
 - ▶ Création d'index et de bibliographie
 - ▶ Numérotation automatique des chapitres, sections et sous-sections
- ▶ Permet d'inclure des formules mathématiques et scientifiques très complexes
- ▶ Permet de générer des figures et des graphiques très complexes
- ▶ Bien intégré avec la plupart des langages de programmation pour la génération automatique de contenu et de documentation.

L^AT_EX : Inconvénients

- ▶ Apprentissage difficile.
- ▶ C'est un langage de programmation compilé.
- ▶ Déboguer un document qui ne compile pas peut être ardu.
- ▶ Le document produit est à la base dans un format imprimable, ce qui peut être contraignant (surtout sur la largeur du document).

Distributions de \LaTeX

- ▶ MikTeX, la distribution recommandée pour Windows. Aussi disponible sur macOS via Homebrew.
- ▶ TeX Live, la distribution la plus courante et celle par défaut sur les distributions majeures de GNU/Linux.
- ▶ MacTeX, distribution TeX Live adaptée pour macOS

Installation de TeX Live (attention, prévoir environ 3 Go de données et 1h !):

```
sudo apt-get install texlive-full
```

\LaTeX : Éditeurs libres

- ▶ AucTeX, extension de GNU Emacs pour \TeX , Multi-plateforme
- ▶ TeXMaker, interface graphique Multi-plateforme
- ▶ LyX, Éditeur de type WYSIWYM utilisant \LaTeX en arrière plan. Multi-plateforme
- ▶ LaTeXila, interface graphique pour GNOME, GNU/Linux seulement
- ▶ TeXnicCenter, interface graphique avancée pour Windows seulement.

\LaTeX : Références

- ▶ The \LaTeX Project
- ▶ \TeX Users Group web site
- ▶ \LaTeX Wikibook

Markdown

- ▶ Markdown est un langage de balisage léger
- ▶ Inspiré du courriel en mode texte
- ▶ Il en existe plusieurs variantes, dont le GitHub Flavored Markdown, le Pandoc Markdown et le R Markdown
- ▶ Très populaire pour produire de la documentation
- ▶ Presque toujours converti vers un format HTML

Markdown: Éditeurs libres

- ▶ Remarkable, Windows et GNU/Linux
- ▶ Visual Studio Code, Multi-plateforme
- ▶ Stackedit, Multi-plateforme, application web

Markdown: Références

- ▶ Site officiel
- ▶ Un guide pour bien commencer avec markdown

DocBook

- ▶ DocBook est un schéma XML très large qui permet de définir sémantiquement le contenu d'un livre ou d'un article.
- ▶ Il ne définit pas comment l'affichage se fait. On utilisera alors un fichier XSLT qui permet de transformer le XML dans un autre format, souvent le HTML.

DocBook: Références

- ▶ What is DocBook?

Pandoc

Pandoc est un outil et une librairie écrit en Haskell qui permet de convertir des documents entre plusieurs formats ouverts.

Il y a aussi une librairie très utilisée par plusieurs autres logiciels. Si votre éditeur permet d'exporter votre travail dans une multitude de formats, c'est probablement du à Pandoc.

Pandoc est multi-plateformes, mais certaines fonctionnalités peuvent être limitées. Par exemple, il faut avoir Microsoft Word ou LibreOffice pour produire des DOCX.

Installation

```
sudo apt-get install pandoc
```

Installation avec Cabal, pour avoir la version la plus récente:

```
cabal install pandoc
```

Pandoc: Formats d'entrée

Les formats d'entrée incluent:

- ▶ markdown
- ▶ DocBook
- ▶ LaTeX
- ▶ OpenDocument
- ▶ Epub
- ▶ ...

Pandoc: Formats de sortie

Les formats de sortie incluent:

- ▶ HTML
- ▶ OpenDocument
- ▶ EPub
- ▶ DocBook
- ▶ LaTeX
- ▶ PDF
- ▶ markdown
- ▶ MediaWiki
- ▶ DokuWiki

Pandoc Markdown

Le Pandoc Markdown est un des formats d'entrée les plus vertatiles. Il s'agit d'une variante de Markdown supportant plusieurs fonctionnalités additionnelles.

Référence: [Anchoring Pandoc Markdown](#)

Pandoc Markdown: Blocs de lignes

Les **blocs** de lignes permettent de séparer le texte manuellement.
Par exemple, pour les adresses:

Code:

```
| Centre de loisirs St-Louis-de-France  
| 1560 Route De L'Église  
| Québec, G1W 3P5
```

Résultat:

Centre de loisirs St-Louis-de-France
1560 Route De L'Église
Québec, G1W 3P5

Pandoc Markdown: Tables simples

Il faut spécifier l'extension `+simple_tables`

Code:

Droite	Gauche	Centre	Defaut
12	12	12	12
123	123	123	123

Table: Demonstration de la syntaxe des tables simples.

Résultat:

Table 1: Demonstration de la syntaxe des tables simples.

Droite	Gauche	Centre	Defaut
12	12	12	12
123	123	123	123

Pandoc Markdown: métadonnées

Les métadonnées permettent d'identifier le titre, l'auteur et la date dans l'entête du document. On utilise l'extension

`+pandoc_title_block`

Code:

```
% Les formats de documents ouverts  
% François Pelletier  
% 25 novembre 2017
```

Pandoc Markdown: mathématiques

Pandoc permet de traiter les équations mathématiques saisies avec la syntaxe LaTeX à l'aide de différents outils tout dépendamment du format de document en sortie.

Les principaux outils utilisés sont MathML et AMS \LaTeX

Pandoc Markdown: mathématiques

On saisit les contenus mathématiques entre signes \$ ou \$\$ pour les expressions multilignes.

Code:

\$\$

```
{\begin{aligned}/home/francois/nextCloud/LinuQ/20171025_par  
&\gamma_{ij}(x,t) \\\n=&\sum_{k=1}^N \sigma_{ik}(x,t) \sigma_{jk}(x,t) \\\n\end{aligned}}
```

\$\$

Résultat:

$$\begin{aligned} &\gamma_{ij}(x,t) \\ &= \sum_{k=1}^N \sigma_{ik}(x,t) \sigma_{jk}(x,t) \end{aligned}$$

Produire des présentations facilement

Un exemple: cette présentation a été écrite avec le langage Markdown et exportée en PDF et en HTML avec Pandoc.

```
#!/bin/bash
```

```
pandoc -f markdown+simple_tables+pandoc_title_block \  
-t beamer -s presentation.md -o presentation.pdf  
pandoc -f markdown+simple_tables+pandoc_title_block \  
-t slidy -s presentation.md -o presentation.html
```

Publier sur un wiki ou un site web

Il est aussi possible d'exporter le contenu d'un document vers la syntaxe dokuwiki, ce qui peut être très utile !

```
pandoc -f markdown+simple_tables+pandoc_title_block \  
-t dokuwiki -s presentation.md -o presentation.dokuwiki
```

Numériser un document papier: Simple Scan

Simple Scan est un logiciel de numérisation de documents qui s'utilise avec un numériseur à plat ou avec alimentation automatique. Il permet de créer des documents PDF facilement. Cette application est développée par le projet GNOME.
Source: GitHub: [GNOME/simple-scan](https://github.com/GNOME/simple-scan)

Numériser un document papier: Tesseract

Tesseract est un logiciel libre de reconnaissance optique de caractères. Il est très utile pour extraire le contenu d'un document numérisé.

Installation:

```
sudo apt-get install tesseract-ocr tesseract-ocr-fra
```

Il faut au préalable préparer le document à la reconnaissance des caractères.

Exemple de document

Les pâtes Catelli Smart® sont une source très élevée de fibres alimentaires. Voici une manière délicieuse d'augmenter votre apport en fibres et d'améliorer votre alimentation. Il suffit de préparer les pâtes Catelli Smart® de la même manière que les pâtes ordinaires et de profiter ensuite d'un goût exceptionnellement savoureux. Voilà le moyen d'obtenir plus de fibres pour toute la famille.

Figure 1

Exemple de document (suite 1)

En utilisant ImageMagick, on peut aligner le document. On utilise ensuite Tesseract pour extraire le texte. On peut combiner les deux commandes en utilisant un pipe.

Code:

```
convert DOC-20171024-215135.jpg -deskew 40% jpg:- | \  
tesseract stdin -l fra -psm 1 DOC-20171024-215135
```

Exemple de document (suite 2)

Résultat:

Les pâtes Caielii Smari® soni une source très:æ élevée
_ de fibres alimentaires. Voici une manièm déiicic'sum
__ _'3_/ d'augmenter votre apport en iihrea ei d'améliumr v
&J' alimentation. il suffit de préparer les pâtes Cai9iii S
dela même manière que les pâtes ordinaires et de proiitg;
ensuite d'un goût exceptionneiiement savoureux. Voilà la un
d'obtenir plus de fibres pour toute la iamiiie.

Manipuler des documents PDF: pdftk

pdftk est un outil qui permet de manipuler des fichiers PDF. Il permet de:

- ▶ Fusionner et extraire des section de documents

```
pdftk a1.pdf a2.pdf cat output a1+a2.pdf
```

- ▶ Extraire une section de document

```
pdftk a1+a2.pdf cat 1 output b1.pdf
```

- ▶ Éclater un document en pages séparées

```
pdftk a1+a2.pdf burst
```

pdftk: Appliquer un filigrane ou ajouter un logo

On peut ajouter une image en superposition en utilisant l'option stamp. Par exemple, pour identifier un document comme confidentiel.

Code:

```
pdftk presentation.pdf stamp confidentiel.pdf \  
  output presentation-conf.pdf
```

Versionnement de documents

Il existe plusieurs logiciels de versionnement de code, dont git, qui peuvent aussi être utilisés pour versionner les documents avec une syntaxe en format texte, tels que markdown, \LaTeX et les formats Wiki. Cependant, il peut être difficile de versionner des documents de format OpenDocument ou PDF, car ils incluent de la compression de données ou des insertions binaires qui ne sont pas réversibles avec un outil tel que Pandoc.

Outils de versionnement

Une solution est d'utiliser un extracteur de texte. En voici quelques uns:

- ▶ pdftohtml, un utilitaire simple qui convertis un document PDF en document HTML. Il est ensuite possible d'utiliser Pandoc pour convertir vers un autre format. C'est la meilleure solution pour faire un coup vite.
- ▶ Apache Tika, une arme de guerre pour extraire le contenu textuel et les métadonnées d'une multitude de formats de données. Disponible sous forme de serveur web ou de librairie Java. C'est la meilleure solution pour gérer une masse de documents.