

Formation *Lodel/PoPuPS*

5. Nettoyer les fichiers .docx avec *Pandoc*

Bernard Pochet – ULiège Library

2021 (cc-by)

Les documents *Word* qui sont passés de mains en mains (auteurs, reviewers, éditeur...) contiennent souvent des **styles indésirables**, pas toujours faciles à détecter et qui provoquent des **erreurs de chargement** dans *Lodel* ou des **erreurs d'affichage** si *Lodel* a néanmoins réussi l'importation.

Il est possible de nettoyer les fichiers :

- légèrement, en imposant les styles “Style paragraphe par défaut” ou “Corps du texte” sur l'ensemble du document mais certains styles résistent
- plus drastiquement, en utilisant le style “Effacer le formatage” mais alors on perd aussi les styles de texte (gras, italique, exposant, indice...)

Le passage par un langage de balisage léger (Markdown) avec l'utilisation d'un logiciel de transformation comme *Pandoc* va vous permettre de nettoyer le fichier en profondeur.

Markdown

Markdown est un langage de balisage (très) léger originellement mis au point pour créer des pages web. Il peut maintenant produire des documents complexes (grâce à *Pandoc* et *LaTeX*).

Markdown (cette présentation est produite à partir d'un fichier Markdown) contient une vingtaine de balises qui sont faciles à mémoriser (ce que vous ne devrez pas faire) et qui permettent de créer tous types de documents.

Il n'est pas nécessaire de maîtriser Markdown pour réaliser cette transformation.

Nettoyage

L'objectif de cette courte présentation est de vous guider pas à pas pour enlever tous les styles non nécessaires dans un document .docx

Cette transformation se fait en deux étapes :












- avec *Pandoc*, on va transformer un fichier .docx en fichier Markdown
- toujours avec *Pandoc*, on va réaliser la transformation inverse (Markdown en .docx)

Il faut passer par le mode “Terminal”¹ et introduire trois commandes simples (la première est destinée à vous placer dans le dossier où se trouve l'article à transformer) :

```
XPS-13-7390:~$ cd Documents/Articles
XPS-13-7390:~/Documents/Articles$ pandoc 722.docx -o 722.md
XPS-13-7390:~/Documents/Articles$ pandoc 722.md -o 722.md.docx
```

¹Sous Linux ou OSX, utilisez “Terminal”. Pour Windows, il faut entrer “cmd” dans la zone de recherche.

Vos fichiers

personnel Documents Articles ▶		Q	☰
Nom	Taille	Type	
 722.docx	5,5 Mo	Document	
 722.lodel.odt	63,1 ko	Document	
 722.md	53,4 ko	Texte	
 722.md.docx	32,1 ko	Document	
 figure1.png	288,8 ko	Image	
 figure2.png	763,1 ko	Image	
 figure3.png	844,5 ko	Image	
 figure4.png	2,3 Mo	Image	
 figure5.png	1,2 Mo	Image	
 formule1.png	5,1 ko	Image	
 formule2.png	5,9 ko	Image	

722.docx² est le fichier de départ, **722.md** est le fichier Markdown intermédiaire et **722.md.docx** est le fichier nettoyé qui sera importé dans **722.lodel.odt** pour être balisé et importé dans *PoPuPS*.

²Il est important de nommer correctement vos fichiers pour éviter les risques de confusion.

Pour pouvoir réaliser cette transformation, il faut installer le logiciel *pandoc* (à partir du site <https://pandoc.org/>).

Pandoc a universal document converter

[About](#)

[Installing](#)

[Getting started](#)

[Demos ▾](#)

Installing pandoc

The simplest way to get the latest pandoc release is to use the installer.

[Download the latest installer](#)

For alternative ways to install pandoc, see below under the heading for your operating system.

[Windows](#)

[macOS](#)

[Linux](#)

[Chrome OS](#)