ARIBARD Antoine
BRZUSTOWSKI Matthias
CART-GRANDJEAN Louis
CASANOVA Arthur

Rapport SPRINT 1 :

L'objectif de ce premier sprint était la découverte et l'utilisation des deux convertisseurs de PDF en texte, à savoir les commandes **pdftotext** et **pdf2txt**, le but étant d'évaluer et de comparer la sortie en plein texte des deux convertisseurs.
Ces deux convertisseurs possèdent des options utilisables spécifiant la façon dont l'on veut convertir un PDF.

Pour la commande pdftotext, les options sont :

- **-f *numero* et -l *numero***, pour préciser respectivement la première et dernière page à convertir
- **-r**, pour préciser la résolution
- **-raw**, pour garder le texte dans le même ordre de contenu
- **-htmlmeta**, pour générer un simple fichier HTML
- **-layout**, pour garder (au mieux) la forme originale du texte
- **-enc *encoding-name*** , pour définir l'encodage à utiliser pour le texte (par défaut UTF-8)
- **-eol *unix* | *dos* | *mac***
- **nopgbrk**, pour ne pas insérer des sauts de pages
- **-opw / -upw *password***, pour préciser le mot de passe du propriétaire / de l'utilisateur pour accéder au PDF
- **-q**, pour ne pas afficher de messages d'erreurs
- **-cfg *config-file***
- **-v,** pour afficher des informations sur la version et les copyrights
- **-h** (ou -help) pour obtenir de l'aide par rapport à la commande

Pour la commande pdf2txt, les principales options sont :

- **-p *numero* / --page-numbers *numero***, pour définir une liste de numéro de pages à convertir
- **-m *nombremax*,** pour définir le maximum de pages à convertir
- **-o *fichier_sortie***, pour définir le fichier texte en sortie
- **-c *encodage***, pour définir l'encodage du texte
- **-s *nombre***, pour définir la taille de police du texte

Cependant les fichiers texte sortis après l'utilisation de chaque convertisseur ne sont pas identiques. Les avantages et les inconvénients quant à l'usage de **pdftotext** et **pdf2txt** sont les suivants :

PDFTOTEXT

AVANTAGES	INCONVÉNIENTS
- le texte en sortie est bien conservé dans l'ensemble - les colonnes sont relativement bien conservées	- la mise en page n'est pas respectée (pas de colonnes) - les tableaux / schémas ne sont pas reconnus, les schémas sont illisibles - il y a un caractère spécial à chaque changement de page - il y a des espaces après les majuscules spéciales - les opérations mathématiques ne sont pas respectées

PDF2TXT

AVANTAGES	INCONVÉNIENTS
- le texte en sortie est bien conservé dans l'ensemble	- les tableaux / schémas ne sont pas respectées, les schémas sont illisibles - les emplacements d'images sont remplacées par un espace vide - certains caractères ne sont pas reconnus - les opérations mathématiques ne sont pas respectées

Après plusieurs tests, nous avons décidé d'utliliser le package **pdftotext** pour la réalisation de notre programme de convertisseur PDF.