

Guide de l'utilisateur du script « XMLConverter.cd »

1 Préambule

Ce script, écrit en CDuce¹, est destiné à modifier un document XML généré par Adobe Acrobat Pro afin de le mettre à la norme DtBook² du Daisy Consortium³.

Il a été réalisé dans le cadre d'un projet étudiant suite à la demande de l'association FAF-LR⁴.

Bien que fonctionnel, ce script ne permet pas de gérer tous les cas pouvant être rencontrés.

2 Avant d'utiliser le script

Il est nécessaire d'installer CDuce pour pouvoir exécuter ce script.

De plus, ce script nécessite un nettoyage du document devant être transformé. Les nettoyages à effectuer sont les suivants.

2.1 Retrait des informations générées par Adobe Acrobat Pro

La présence des informations générées par Adobe Acrobat Pro cause des problèmes empêchant l'exécution du script. Il faut donc retirer les balises présentes avant la balise <TaggedPDF-Doc> (il faut également supprimer leur contenu). Il faut également, à l'intérieur de cette balise <TaggedPDF-Doc>, supprimer les balises situées avant les balises <?xpacket>, y compris ces dernières.

2.2 Retrait des balises <Sect>

Il est nécessaire de retirer ces balises. Cependant, leur contenu doit rester à l'intérieur du document afin d'éviter les pertes.

<Sect> et </Sect> sont donc les deux seuls éléments à devoir être supprimés.

2.3 Nettoyage des balises <Figure>

Ces balises peuvent contenir des balises <ImageData>, des balises <Caption> et/ou du texte. Cependant, seules les balises <ImageData> sont gérées. Il faut donc retirer les balises <Caption> ainsi que le texte.

Afin d'éviter la perte de données, il reste préférable de garder une trace de ces contenus, afin de les rajouter dans le document (dans des balises <caption> cependant).

2.4 Vérification du bon enchaînement des titres

Les balises de titre <H1>, <H2> ... <H6> potentiellement présentes dans le document original doivent, pour pouvoir être transformées, suivre un certain ordre. Ainsi, lorsque l'on a une balise d'un niveau élevé la balise de titre suivante devra être d'un niveau supérieur ou égal au niveau inférieur.

¹ [Site web du langage](#)

² DTD créé par le Daisy Consortium pour l'édition de livres.

³ Association à but non lucratif ayant pour but de créer « la meilleure façon de lire et publier ».

⁴ Fédération des aveugles et amblyopes de France – Languedoc Roussillon

Par exemple, une balise <H2> pourra être suivie par une balise <H3>, une balise <H2> ou une balise <H1>.

Si des balises ne se suivent pas l'ordre, il y a deux façons de corriger cela :

1. Remonter les balises (remplacer les balises de mauvais niveau par des balises de plus haut niveau)
2. Rajouter les balises manquantes (tout en pouvant les supprimer une fois la transformation effectuée).

3 Utilisation du script

Il existe deux façons équivalentes d'utiliser le script.

Il faut entrer l'une des deux commandes écrites ci-dessous dans le terminal⁵.

```
cduce C:\Users\Arthur\Desktop\CDuce-Script\XMLConverter.cd --arg  
«J:\PI\SOURCES\Hist 1e p50_79 part1 ch2 xml\doc.xml»
```

```
cduce --script C:\Users\Arthur\Desktop\CDuce-Script\XMLConverter.cd  
«J:\PI\SOURCES\Hist 1e p50_79 part1 ch2 xml\doc.xml»
```

Il est possible de faire un glisser déposer du script et du document à modifier pour aller plus vite. Si vous faites un glisser déposer, les guillemets seront automatiquement places.

Cependant, ainsi écrites, ces commandes afficheront dans le terminal le résultat de la transformation. Pour que ce résultat soit écrit à l'intérieur d'un fichier, il suffit de rajouter « > *chemin vers le fichier* » à la suite de la commande. Par exemple, pour la première commande, si l'on souhaite mettre le résultat dans le même dossier que doc.xml, il faut entrer la commande :

```
cduce C:\Users\Arthur\Desktop\CDuce-Script\XMLConverter.cd --arg  
«J:\PI\SOURCES\Hist 1e p50_79 part1 ch2 xml\doc.xml» > «J:\PI\SOURCES\Hist 1e p50_79 part1  
ch2 xml\resultat.xml »
```

4 Après l'utilisation du script

Une fois le script exécuté, il reste à rajouter les balises enlevées dans le nettoyage des balises <Figure> (si souhaité), qui sont devenues des balises <imggroup>. Les textes retirés et les contenus des balises <Caption> doivent être placés dans des balises <caption>.

Il faut également supprimer les balises [à l'étape 2.4](#), si la deuxième solution a été choisie.

En cas de soucis, vous pouvez contacter les créateurs du script aux adresses suivantes :

[Charles BORDIER](#)

[Arthur JAMELOT](#)

⁵ Ou invite de commande, selon les systemes.