

Programme d'extraction des fichiers Dbx vers Word

Tout d'abord merci à ceux qui ont participé à la préparation de ce programme par leurs test et leurs conseils avisés, en particulier André et Fabrice qui ont bien voulu essayer les plâtres.

Merci aussi à JLN pour sa participation involontaire par la mise à disposition du source (<http://www.vbfrance.com/code.aspx?ID=5657>) du superbe calendrier perpétuel que j'ai utilisé ici.

Objectif : Extraire d'un fichier Dbx (Outlook Express) des messages et en transférer le texte contenu dans un fichier word.

Déroulement :

Étape 1 : lecture des messages contenus dans le fichier Dbx. Les auteurs des messages sont inscrits par ordre alphabétique dans une liste déroulante. Un avertissement clignotant indique que l'extraction est en cours ; les dates extrêmes des messages sont affichées (plus récent et plus ancien dans la période indiquée) un message s'affiche dans une boîte de dialogue en fin d'extraction.

Des sélections de deux types peuvent être combinées :

- limitation éventuelle aux messages récents avec trois paliers
- limitation aux messages marqués ou suivis.

L'option « Messages supprimés » ajoute à la sélection tout le contenu des messages supprimés récupérables, c'est à dire tant qu'il n'y a pas eu de compactage du dossier par Outlook Express.

Étape 2 : Génération du document Word, soit à partir de tous les messages extraits à l'étape 1, soit avec une sélection supplémentaire sur leur(s) auteur(s) et avec une nouvelle limitation de la période concernée. Un message s'affiche dans une boîte de dialogue pour annoncer la fin de génération.

Le résultat peut être, au choix, trié sur la date, par auteur (et date croissante) ou par conversation.

Si on souhaite réduire la période de temps pour la génération du document word, cliquer sur le bouton, un premier calendrier permet de saisir la nouvelle limite inférieure, puis un autre pour la limite supérieure. Cliquer ensuite sur « Générer »

Le bouton "Interrompre" fonctionne en étape 1 aussi bien qu'en étape 2, mais dans les deux cas il faut recommencer au début de l'étape interrompue.

Résultat :

Pour chaque message :

- Le sujet du message en style « Titre 1 » ou « Titre 2 »
- La date du message (si sélectionnée) en style « DbxDate »
- L'auteur (si sélectionné) avec le style « DbxAuteur »
- Le texte en style Normal. Lorsque seul l'en-tête a été chargé dans Outlook Express, une mention l'indique dans le document Word.

Si les styles spécifiques cités ci-dessus, n'existent pas, ils sont créés dans le document en début de traitement.

L'option « petit nettoyage » prend en compte le fichier "LignesAignorer.ini". Toutes les lignes des messages dont le texte figure en totalité dans une des lignes du fichier indiqué sont ignorées. La comparaison se fait sans tenir compte de la casse ; ainsi Merci et merci sont équivalents. Chacun peut faire évoluer ce fichier comme il l'entend.

Quelques remarques pratiques :

L'option de vérification de l'orthographe en cours de frappe est enlevée pour éviter des messages d'avertissement qui interviennent et perturbent la génération.

Si le document Word reçoit plus de 40 messages, le fichier obtenu est enregistré dans le répertoire temp sous le nom ~~TempLireDbx0.doc, ~~TempLireDbx1.doc etc.

L'affichage Word est partiellement neutralisé pendant la génération du document, pour accélérer (?) le processus qui demeure cependant assez lent.

Il vaut mieux faire une copie du fichier dbx à traiter. Si on traite un fichier alors que OE est connecté, le contenu du fichier pouvant évoluer pendant le traitement, on peut avoir un résultat un peu fantaisiste ; des messages d'erreurs d'incohérence peuvent alors apparaître.

Si on coche la case « **mémoriser la configuration** » un petit fichier LireDbxEnWord.ini est inscrit dans le répertoire du programme avec le nom complet du fichier lu et les choix des différents paramètres, hormis :

- « tous auteurs » positionné systématiquement
- « messages supprimés » dont l'usage est considéré comme exceptionnel.

L'enregistrement ne se fait que si une extraction a été effectuée, lors de l'arrêt normal du programme en utilisant le bouton ad hoc. Lors de l'utilisation suivante, le dernier fichier traité sera proposé par défaut.

Ce fichier LireDbxEnWord.ini, peut être supprimé sans problème.

Le programme a été écrit de manière à ne pas trop monopoliser la capacité de calcul de l'ordinateur malgré sa complexité et les volumes traités, il peut donc durer assez longtemps.

Cas particulier du tri par conversation : la situation idéale est celle où la question d'origine et les différentes réponses sont « suivies ». Dans ce cas la question est en style Titre 1, et les réponses – par ordre d'ancienneté – en Titre 2. Comme on ne se trouve pas toujours dans cette situation, si la question n'est pas « suivie », c'est le message le plus ancien qui aura le niveau 1.

Cas particulier de la récupération des messages supprimés : Outlook Express n'efface physiquement les messages supprimés que lors de l'opération de compactage du dossier. Tant que cette opération n'a pas été faite on peut récupérer une grande partie des messages. Le texte brut de ces messages est repris en fin du document Word généré, sans aucune mise en forme. La perte d'information joue en particulier sur quatre caractères par paquet de 500 caractères environ, ils sont remplacés par des espaces. Attention : les pièces jointes (images ou autres) et le HTML sont récupérés au moins temporairement, ce qui peut donner des volumes monstrueux et des temps de traitement impressionnants. Normalement ils ne sont pas restitués sur le fichier word. Une limitation avec message d'avertissement est affiché par le programme quand le programme a atteint sa limite.

Limitations :

- La taille des messages extraits est limitée arbitrairement à 400 lignes.
- Récupération des messages supprimés limitée à 1.000.000 caractères environ.
- Il vaut mieux ne pas avoir d'autre fichier Word ouvert en même temps que la génération du document, le traitement peut les modifier, ça ne se produit pas toujours (de moins en moins d'ailleurs) et je n'ai pas encore complètement compris pourquoi.

Il reste de petits défauts dans le texte, en particulier les coupures faites par les différents lecteurs de messages, mais on ne peut qu'y intervenir à la main. Des texte « parasites » persistent aussi tels que « NextPart_ », il faut bien garder un peu de travail pour la maintenance.

On peut faire une petite macro qui remplace toutes les fins de paragraphe, suivis ou non de >, par une espace et garder la dernière fin de paragraphe de la sélection.

Voici un exemple de macro qui, de plus, remplace les doubles sauts de paragraphe par un seul, l'objectif étant de conserver les retours à la ligne volontaires.

Option Base 1

```
Sub TexteDBx()
Dim Textes(50, 2) As String
Dim i As Integer
i = 1
Textes(i, 1) = "^p^p"
Textes(i, 2) = "$£$"
```

```

i = i + 1
Textes(i, 1) = "^p> > > > "
Textes(i, 2) = " "
i = i + 1
Textes(i, 1) = "^p> > > "
Textes(i, 2) = " "
i = i + 1
Textes(i, 1) = "^p> > "
Textes(i, 2) = " "
i = i + 1
Textes(i, 1) = "^p> "
Textes(i, 2) = " "
i = i + 1
Textes(i, 1) = "^p"
Textes(i, 2) = " "
i = i + 1
Textes(i, 1) = "$£$"
Textes(i, 2) = "^p"

For i = 1 To UBound(Textes)
    If Len(Textes(i, 1)) = 0 Then Exit For
    Selection.Find.ClearFormatting
    Selection.Find.Replacement.ClearFormatting
    With Selection.Find
        .Text = Textes(i, 1)
        .Replacement.Text = Textes(i, 2)
        .Forward = True
        .Wrap = wdFindStop
        .Format = False
        .MatchCase = False
        .MatchWholeWord = False
        .MatchWildcards = False
        .MatchSoundsLike = False
        .MatchAllWordForms = False
    End With
    Selection.Find.Execute Replace:=wdReplaceAll
Next i
Selection.EndOf 'Unit:=
Selection.TypeParagraph
End Sub

```