Fiche n°2 :
Préparer le corpus issu de Gromoteur

Sommaire

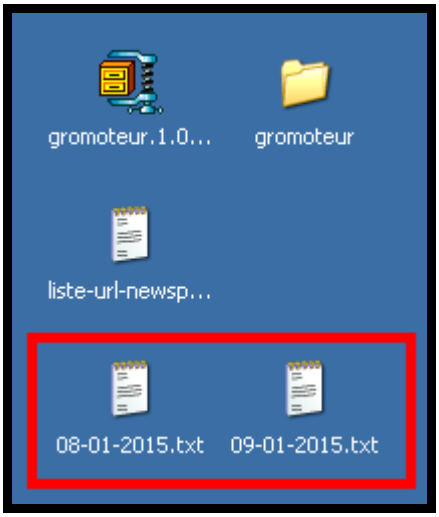
Préambule :

A l'issue de l'étape précédente (Fiche n°1), on dispose ici de 2 fichiers (en général, on aura beaucoup de fichiers…) produits par le module d'export de Gromoteur.

Chacun des fichiers a été constitué à 2 dates différentes : le 8 janvier 2015 et le 9 janvier 2015.

Chacun contient le résultat de l'aspiration des URLs traitées (cf fiche n°1) réalisée à chacune de ces 2 dates.

Les noms de ces fichiers indiquent leur date de création :



Remarque importante :

Cette session de préparation des fichiers issus de Gromoteur est conditionnée par le choix de configuration fait dans la fiche n°1. Dans celle-ci, on a choisi de construire un fichier d'export à chaque phase d'aspiration, on a donc ici autant de fichier d'export que d'aspiration.

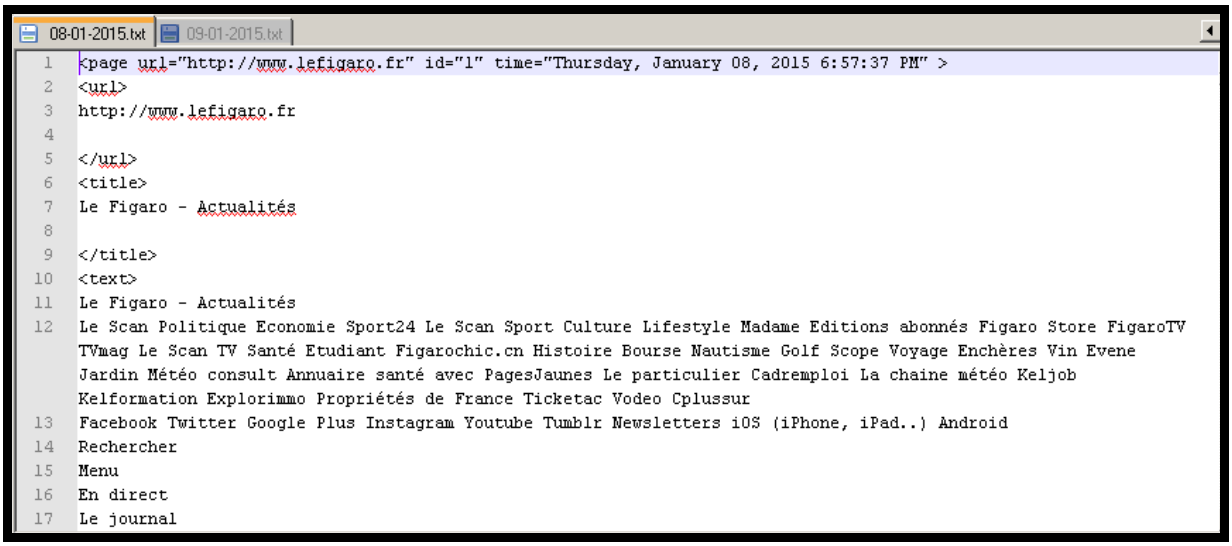
Une autre configuration de Gromoteur aurait pu conduire à une autre stratégie : regroupe directement toutes les aspirations dans un seul fichier d'export.

Etape n°1 : Descriptif du contenu des fichiers issus de Gromoteur

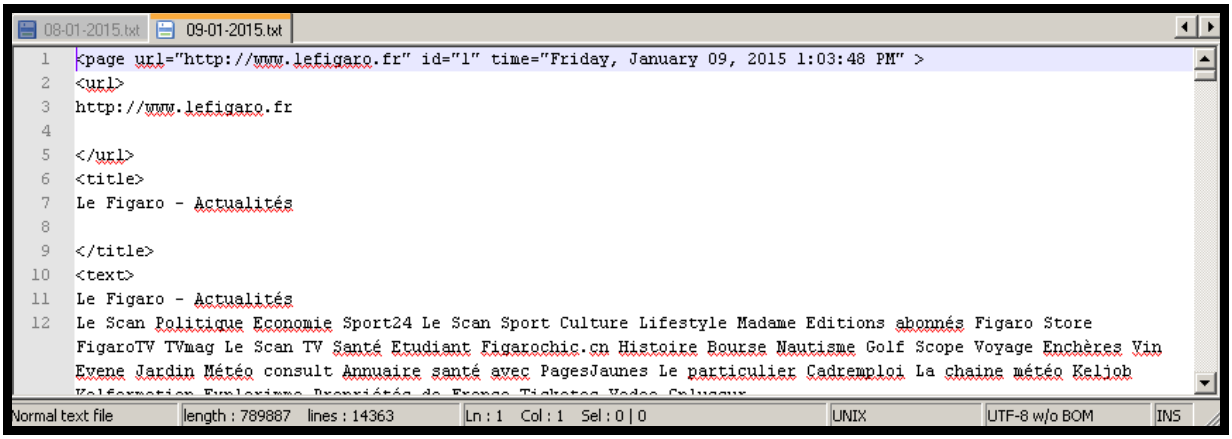
Les 2 fichiers sont au format TXT (et encodés en UTF-8) .

On donne ici à voir l'allure de ces fichiers dans l'éditeur Notepad++ :

1. Début du fichier 08-01-2015.txt :

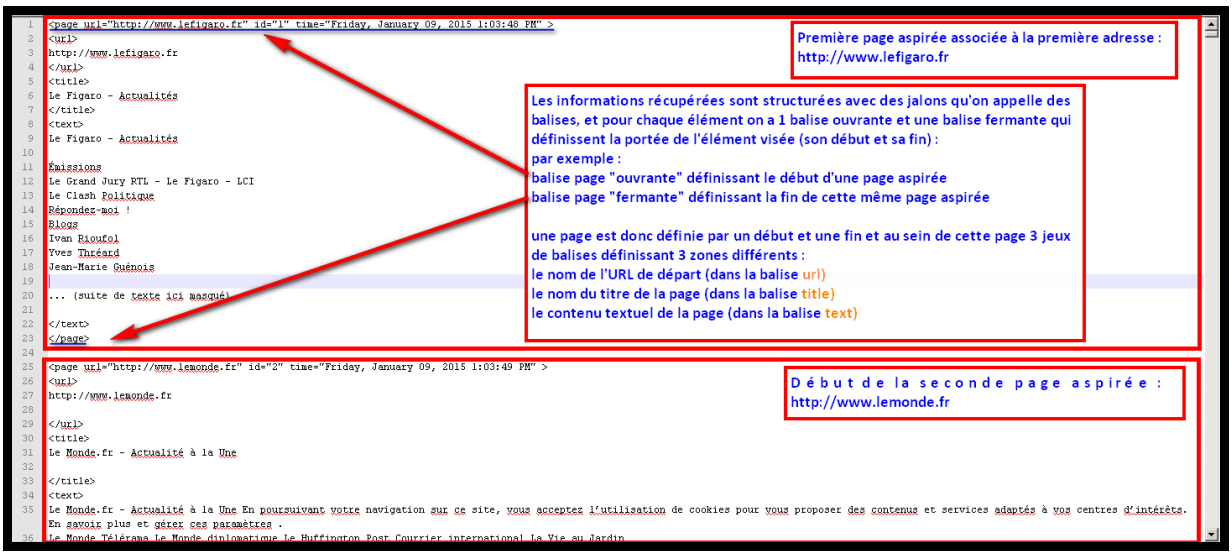


2. Début du fichier 09-01-2015.txt



Les 2 fichiers ont la même structure.

Ils empilent les différentes adresses aspirées suivant le format suivant :



Etape n°2 : Préambule à la préparation du corpus chronologique

Objectif : construire un seul fichier regroupant tous les fichiers aspirés et le préparer pour être chargé dans le logiciel Le Trameur

Deux étapes doivent donc être réalisées :

1. Concaténer tous les fichiers
2. Préparer le fichier final : nettoyage éventuel, ajout de balises pour définir la chronologie

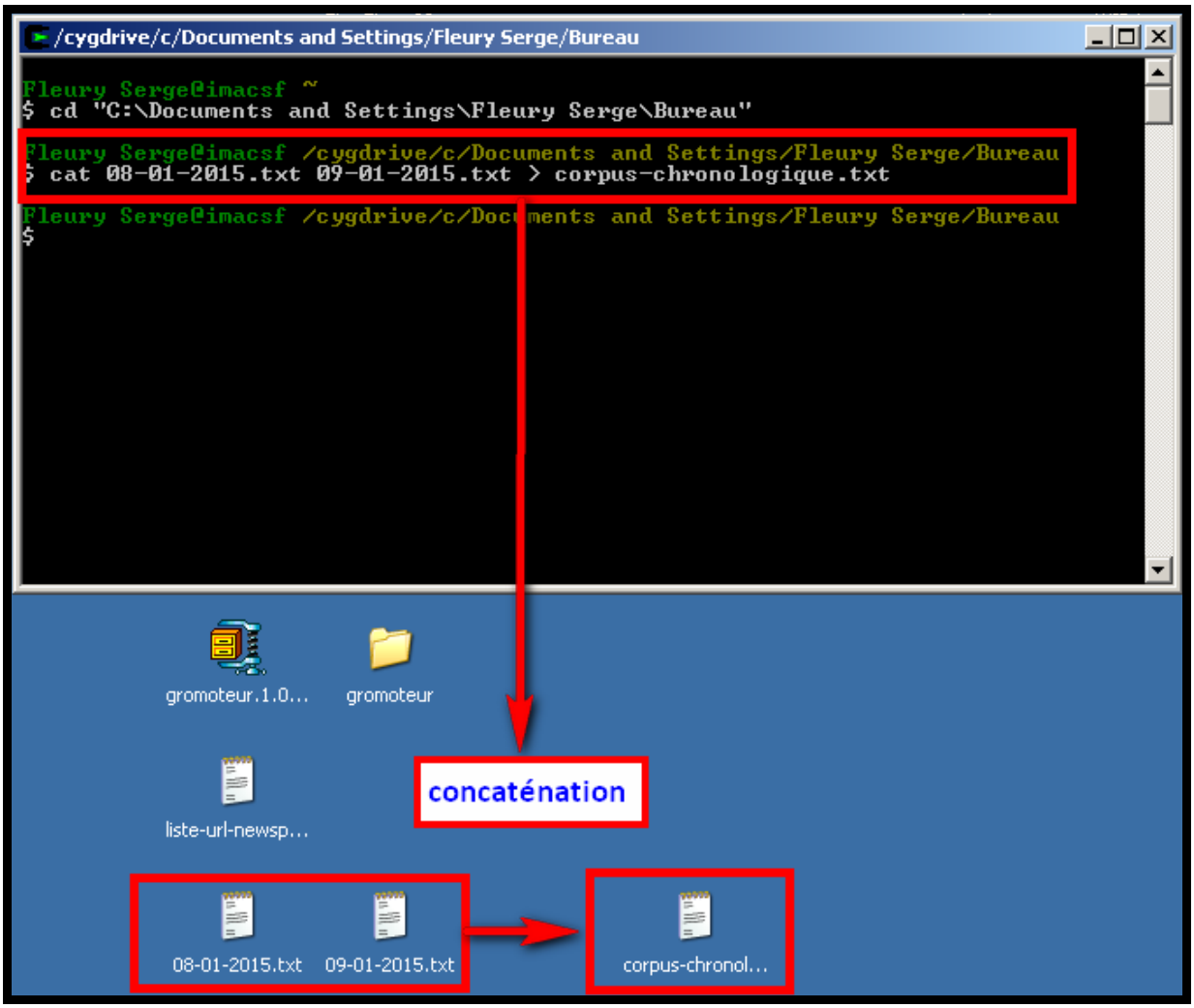
Etape n°3 : Concaténation des fichiers issus de Gromoteur

La concaténation des fichiers peut être réalisée de différentes manières :

1. Construire le fichier final dans un éditeur en y insérant un à un tous les fichiers d'export
2. Utiliser des programmes idoines (la commande cat par exemple dans une fenêtre de commandes)

Ces 2 étapes seront détaillées en cours.

Ci-dessous, la concaténation réalisée dans une fenêtre de commandes :



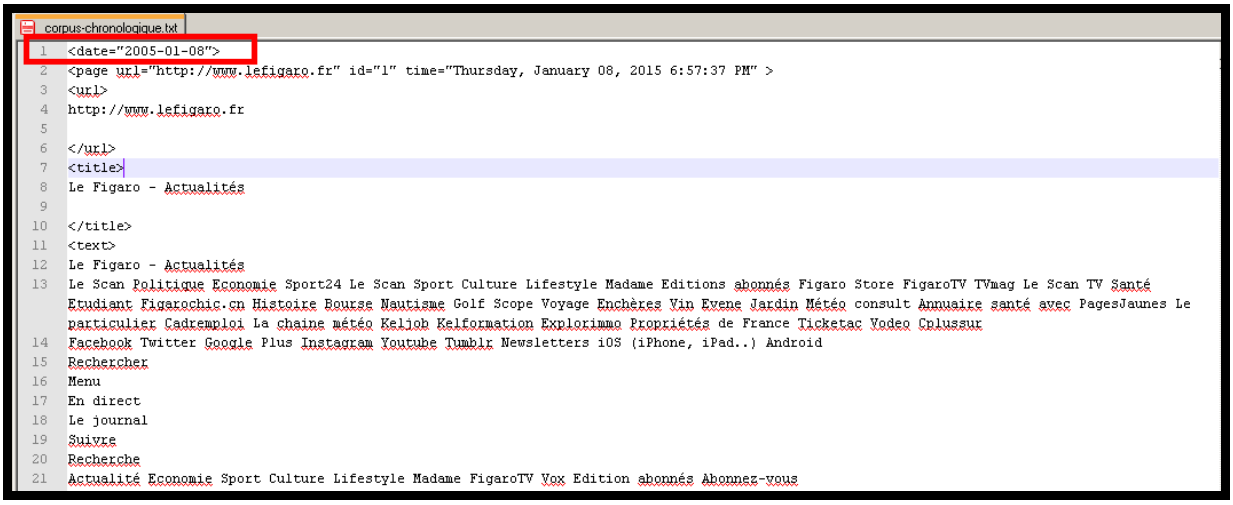
Etape n°4 : Préparation du fichier final

Une fois les fichiers concaténés, on dispose désormais d'un seul fichier nommé :

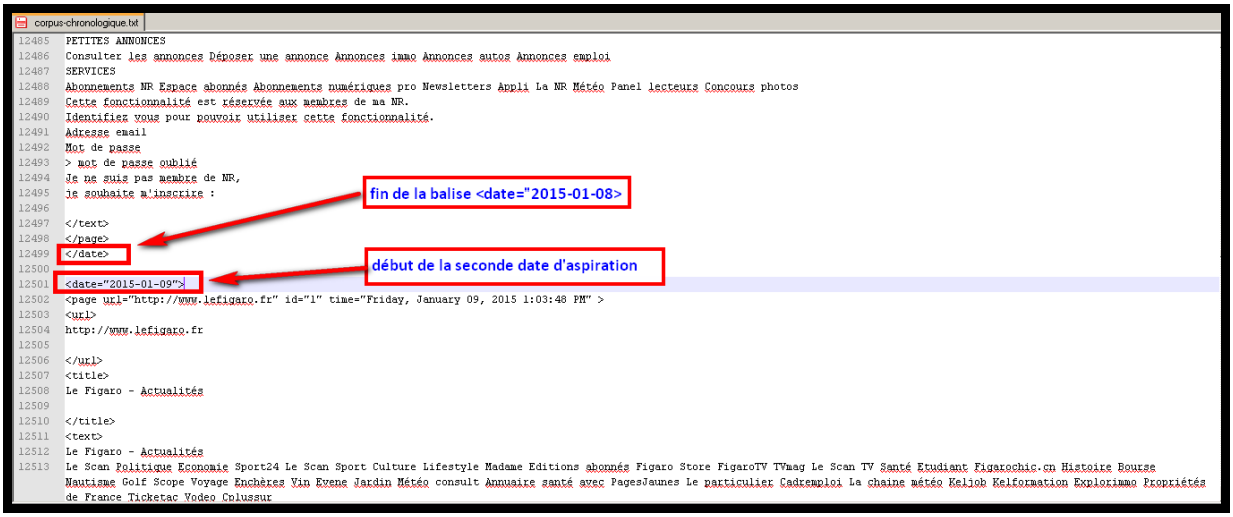
corpus-chronologique.txt

La première étape va conduire à intégrer dans ce fichier final des « jalons textuels » permettant de mettre au jour la chronologie de manière explicite.

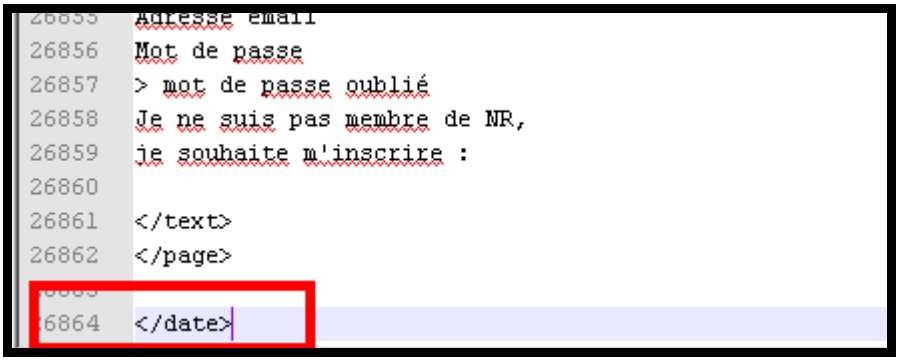
Pour cela, on va introduire une nouvelle balise au début et à la fin de chacun des fichiers concaténés. Ci-dessous, on insère la base (ouvrante) date indiquant la date de la première aspiration :



Un peu plus loin dans le fichier, on insère la balise de fin de cette première date d'aspiration, puis la balise de date (ouvrante) de l'aspiration suivante :



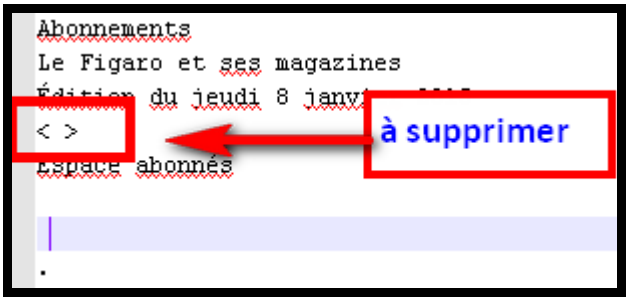
Il reste à insérer à la fin du fichier la balise de fin de la seconde date d'aspiration :



La seconde étape de préparation va consister à s'assurer que le fichier final est « propre » : certains caractères peuvent nécessiter des opérations supplémentaires.

C'est le cas notamment des caractères < et > qui sont utilisés pour écrire des balises, ils ne peuvent pas être utilisés ailleurs dans le fichier.

Si on examine le fichier final, on remarque par exemple ceci :



ci-dessus on remarque la séquence : < >

elle « ressemble » à une balise mais elle ne correspond pas à une balise utilisée dans notre corpus pour le structurer !

Il faut donc supprimer ce type de séquence.

Etape n°5 : Raffinements ultimes

Une dernière étape va conduire à simplifier l'écriture des balises issues de gromoteur.

Essentiellement pour des raisons de visibilité dans le logiciel à venir.

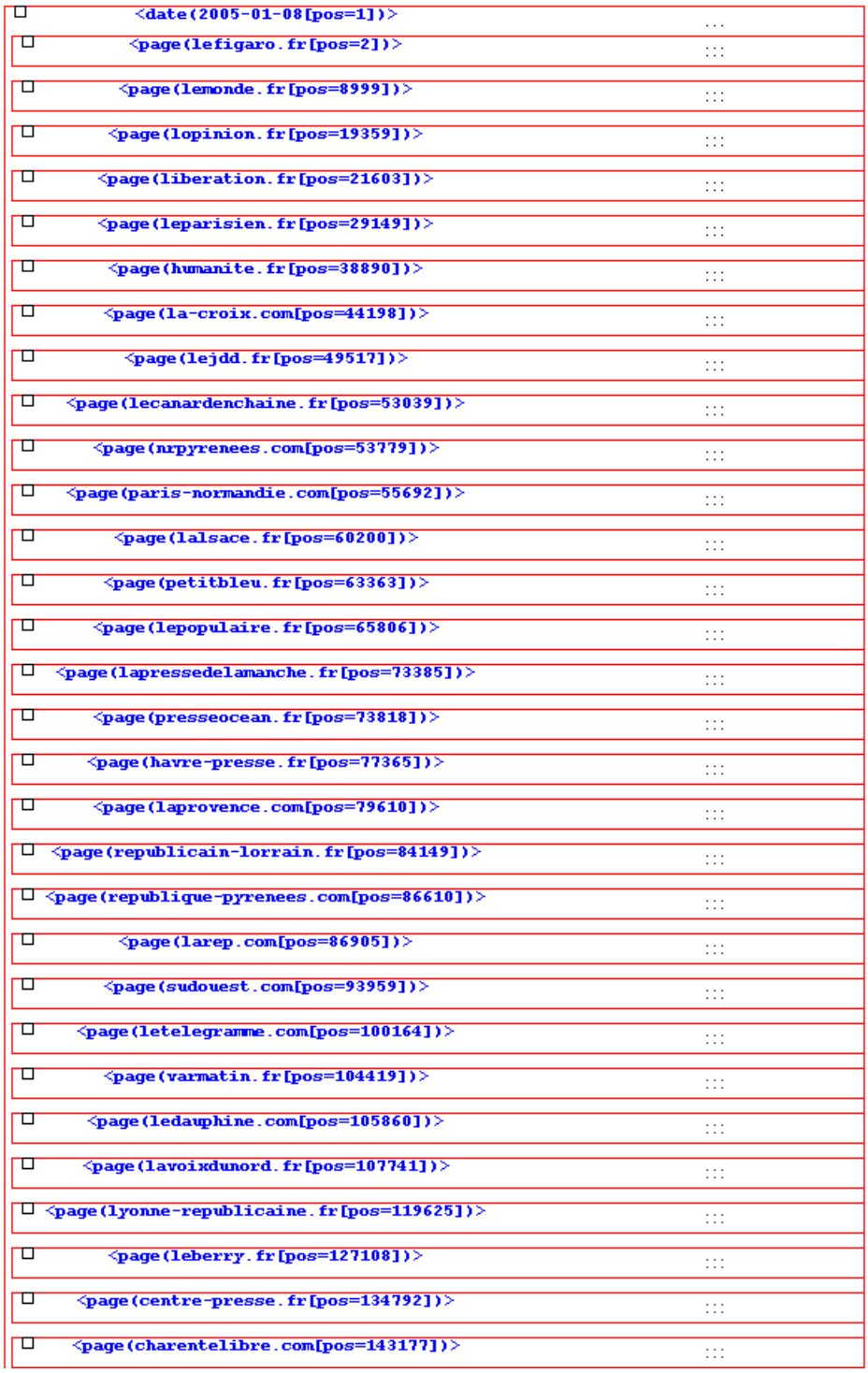
Cette étape va consister à transformer l'écriture de certaines balises de la manière suivante :

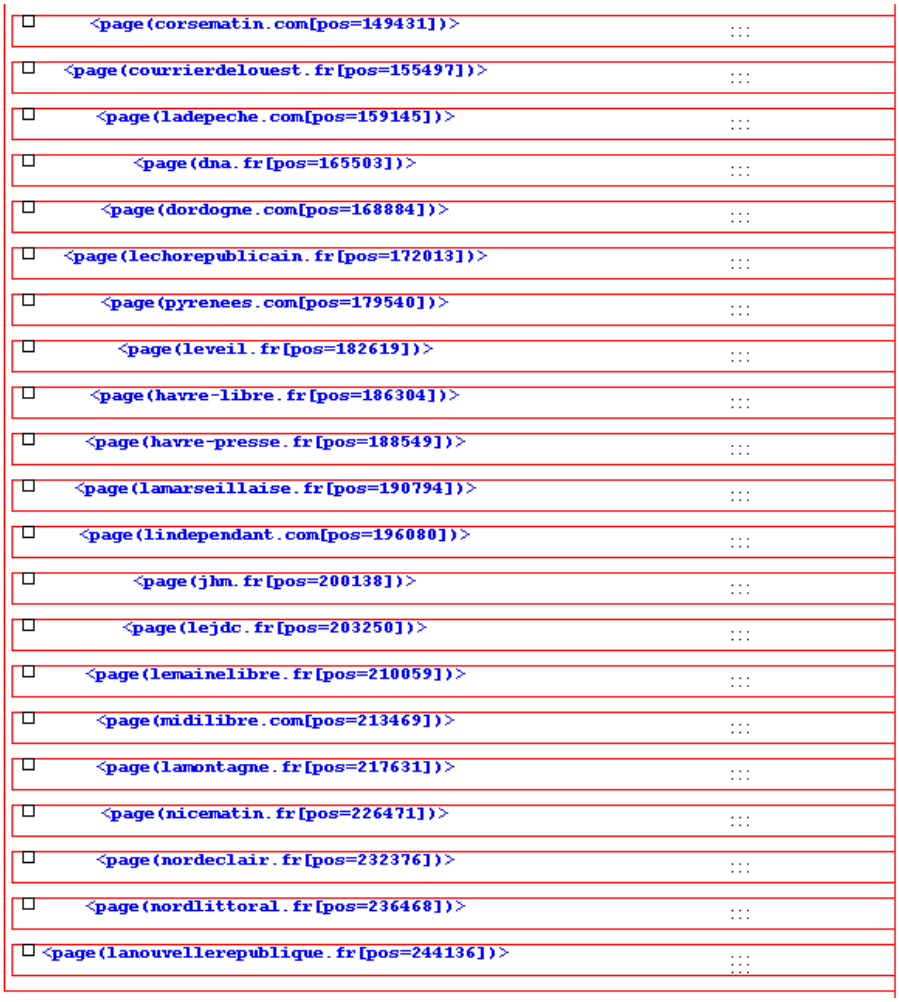
```
<date="2005-01-08">
<page="lefigaro.fr">
```

On ne gardera donc dans la balise page que la référence à l'adresse de la page et simplifiant cette adresse (suppression de http://www.)

Etape n°6 : Corpus final, un emboitement de parties

Au final, le corpus final est prêt à être chargé dans Le Trameur, il est structuré en différentes parties. La représentation graphique suivante (extrait) est produite par Le Trameur : elle donne à voir le système de parties qui sera disponible pour contraster les différentes parties du corpus





On ne voit ici que la structuration sur la première date d'aspiration, contenant toutes les pages aspirées. Une structuration similaire est disponible pour les autres dates d'aspiration.