

Fiche n°2 : Préparer le corpus issu de Gromoteur

Sommaire

Préambule :	2
Etape n°1 : Descriptif du contenu des fichiers issus de Gromoteur.....	3
Etape n°2 : Préambule à la préparation du corpus chronologique.....	5
Etape n°3 : Concaténation des fichiers issus de Gromoteur	6
Etape n°4 : Préparation du fichier final	7
Etape n°5 : Raffinements ultimes.....	9
Etape n°6 : Corpus final, un emboitement de parties.....	10

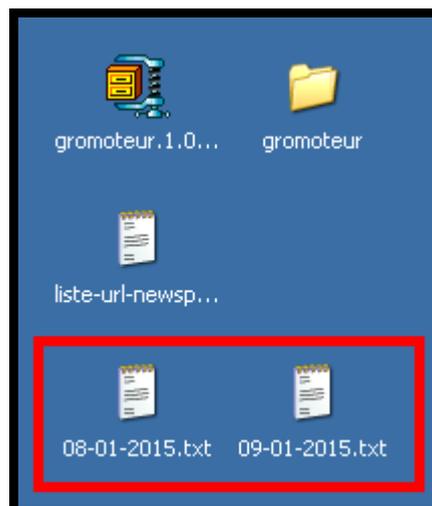
Préambule :

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.

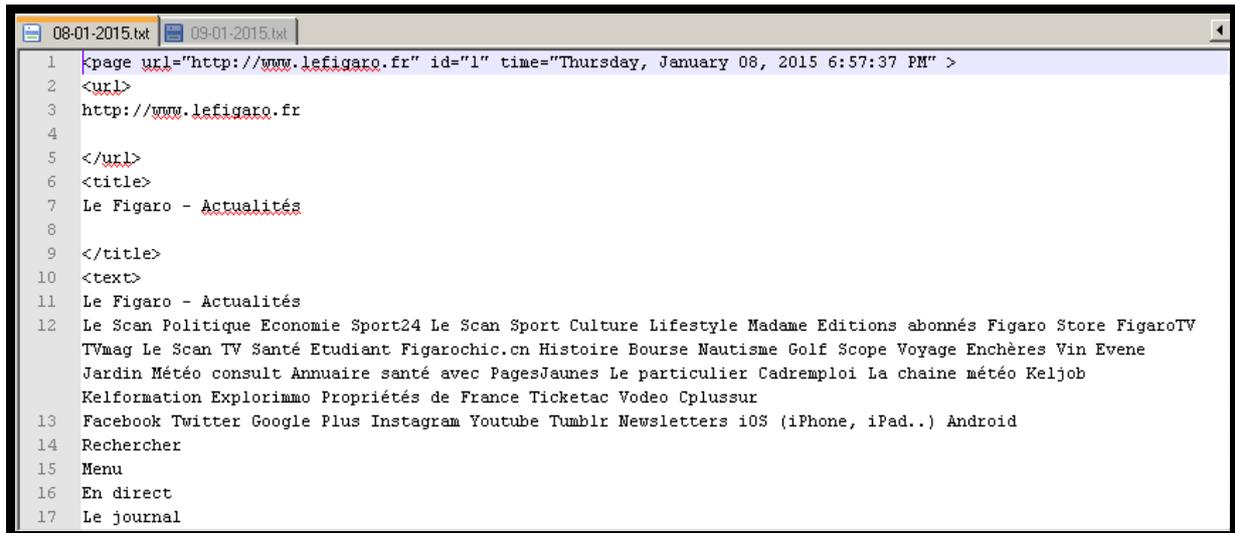
Une autre configuration de Gromoteur aurait pu conduire à une autre stratégie : regroupe directement toutes les aspirations dans un seul fichier d'export.

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

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

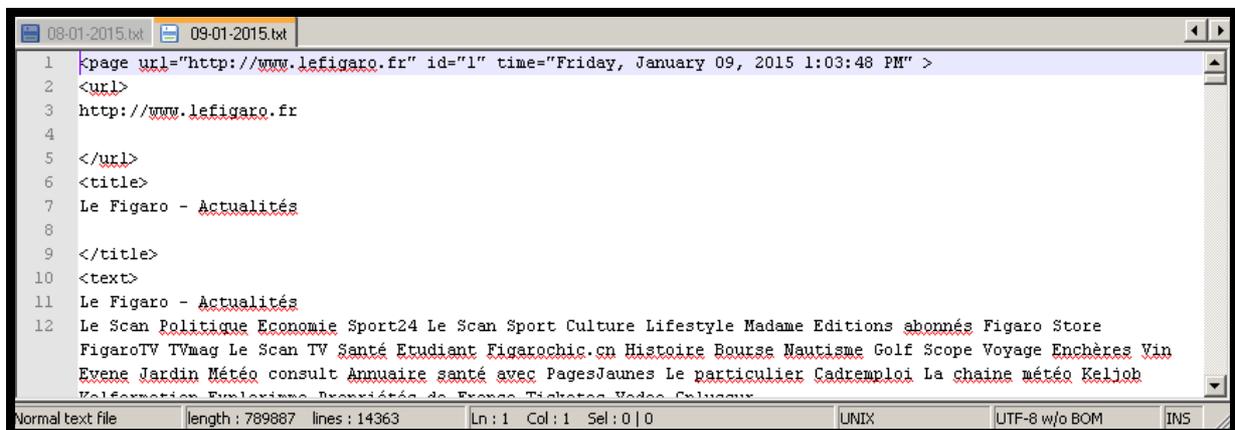
On donne ici à voir l'allure de ces fichiers dans l'éditeur Notepad++ :

1. Début du fichier 08-01-2015.txt :



```
08-01-2015.txt 09-01-2015.txt
1 <page url="http://www.lefigaro.fr" id="1" time="Thursday, January 08, 2015 6:57:37 PM" >
2 <url>
3 http://www.lefigaro.fr
4
5 </url>
6 <title>
7 Le Figaro - Actualités
8
9 </title>
10 <text>
11 Le Figaro - Actualités
12 Le Scan Politique Economie Sport24 Le Scan Sport Culture Lifestyle Madame Editions abonnés Figaro Store FigaroTV
13 TVmag Le Scan TV Santé Etudiant FigaroChic.cn Histoire Bourse Nautisme Golf Scope Voyage Enchères Vin Evene
14 Jardin Météo consult Annuaire santé avec PagesJaunes Le particulier Cadremploi La chaine météo Keljob
15 Kelformation Explorimmo Propriétés de France Ticketac Vodeo Cplussur
16 Facebook Twitter Google Plus Instagram Youtube Tumblr Newsletters iOS (iPhone, iPad..) Android
17 Rechercher
18 Menu
19 En direct
20 Le journal
```

2. Début du fichier 09-01-2015.txt



```
08-01-2015.txt 09-01-2015.txt
1 <page url="http://www.lefigaro.fr" id="1" time="Friday, January 09, 2015 1:03:48 PM" >
2 <url>
3 http://www.lefigaro.fr
4
5 </url>
6 <title>
7 Le Figaro - Actualités
8
9 </title>
10 <text>
11 Le Figaro - Actualités
12 Le Scan Politique Economie Sport24 Le Scan Sport Culture Lifestyle Madame Editions abonnés Figaro Store
13 FigaroTV TVmag Le Scan TV Santé Etudiant FigaroChic.cn Histoire Bourse Nautisme Golf Scope Voyage Enchères Vin
14 Evene Jardin Météo consult Annuaire santé avec PagesJaunes Le particulier Cadremploi La chaine météo Keljob
15 Kelformation Explorimmo Propriétés de France Ticketac Vodeo Cplussur
Normal text file length : 789887 lines : 14363 Ln : 1 Col : 1 Sel : 0 | 0 UNIX UTF-8 w/o BOM INS
```

Les 2 fichiers ont la même structure.

Ils empilent les différentes adresses aspirées suivant le format suivant :

The image shows a screenshot of HTML code for two pages, with red boxes and arrows providing annotations. The first page is from lefigaro.fr and the second is from lemonde.fr.

Annotations:

- Première page aspirée associée à la première adresse :** <http://www.lefigaro.fr>
- Les informations récupérées sont structurées avec des jalons qu'on appelle des balises, et pour chaque élément on a 1 balise ouvrante et une balise fermante qui définissent la portée de l'élément visée (son début et sa fin) :**
 - par exemple :
 - balise page "ouvrante" définissant le début d'une page aspirée
 - balise page "fermante" définissant la fin de cette même page aspirée
- une page est donc définie par un début et une fin et au sein de cette page 3 jeux de balises définissant 3 zones différents :**
 - le nom de l'URL de départ (dans la balise `uri`)
 - le nom du titre de la page (dans la balise `title`)
 - le contenu textuel de la page (dans la balise `text`)
- Début de la seconde page aspirée :** <http://www.lemonde.fr>

HTML Code (Line numbers 1-36):

```
1 <page uri="http://www.lefigaro.fr" id="1" time="Friday, January 09, 2015 1:03:48 PM" >
2 <uri>
3 http://www.lefigaro.fr
4 </uri>
5 <title>
6 Le Figaro - Actualités
7 </title>
8 <text>
9 Le Figaro - Actualités
10
11 Émissions
12 Le Grand Jury RTL - Le Figaro - LCI
13 Le Clash Follage
14 Répondre-xxx !
15 Blogs
16 Ivan Rioufol
17 Yves Thériard
18 Jean-Marie Guénois
19
20 ... (suite de texte ici masqué)
21
22 </text>
23 </page>
24
25 <page uri="http://www.lemonde.fr" id="2" time="Friday, January 09, 2015 1:03:49 PM" >
26 <uri>
27 http://www.lemonde.fr
28
29 </uri>
30 <title>
31 Le Monde.fr - Actualité à la Une
32
33 </title>
34 <text>
35 Le Monde.fr - Actualité à la Une En poursuivant votre navigation sur ce site, vous acceptez l'utilisation de cookies pour vous proposer des contenus et services adaptés à vos centres d'intérêts.
36 En savoir plus et gérer ces paramètres .
37 Le Monde Télérama Le Monde diplomatique Le Huffington Post Courrier international La Vie au Jardin
```

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

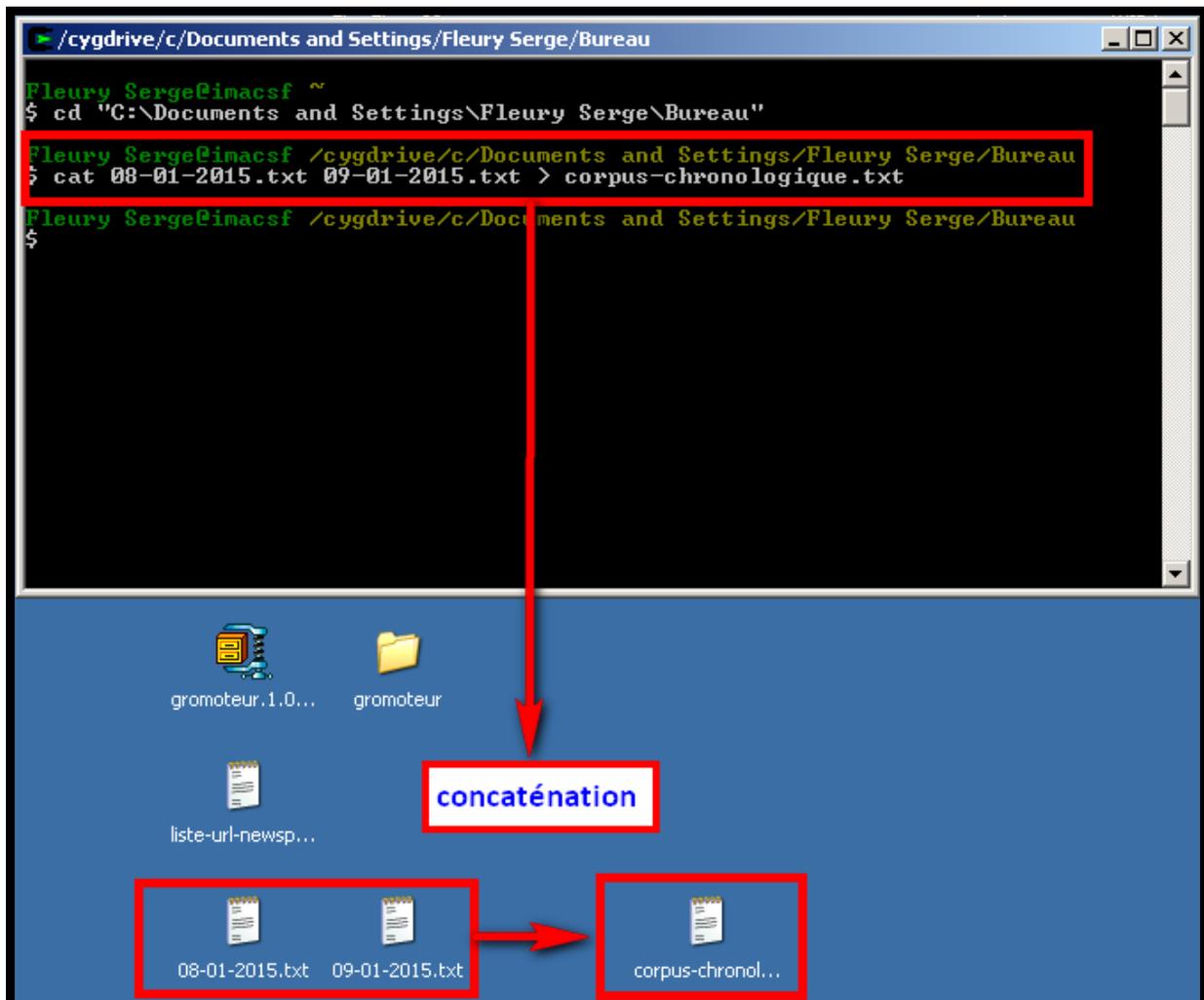
Etape n°3 : Concaténation des fichiers issus de Gromoteur

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 :



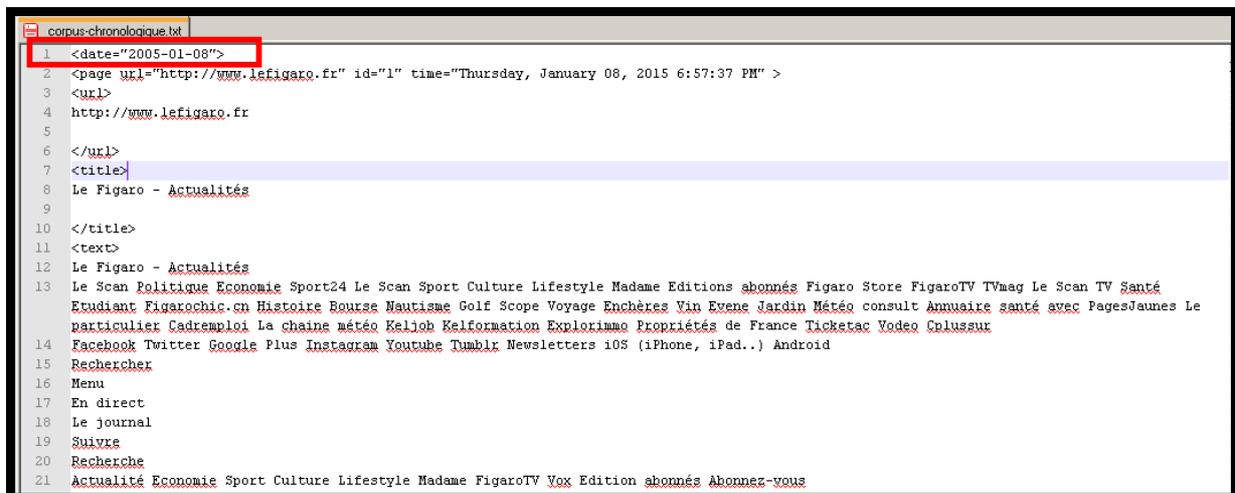
Etape n°4 : Préparation du fichier final

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

corpus-chronologique.txt

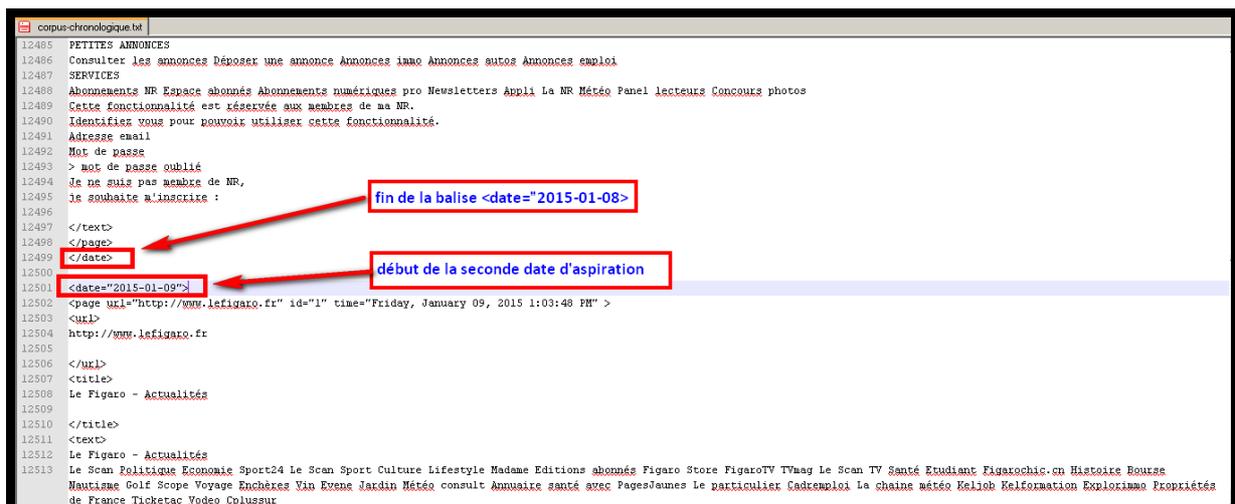
La première étape va conduire à intégrer dans ce fichier final des « jalons textuels » permettant de mettre au jour la chronologie de manière explicite.

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 :



```
corpus-chronologique.txt
1 <date="2015-01-08">
2 <page url="http://www.lefigaro.fr" id="1" time="Thursday, January 08, 2015 6:57:37 PM" >
3 <url>
4 http://www.lefigaro.fr
5
6 </url>
7 <title>
8 Le Figaro - Actualités
9
10 </title>
11 <text>
12 Le Figaro - Actualités
13 Le Scan Politique Economie Sport24 Le Scan Sport Culture Lifestyle Madame Editions abonnés Figaro Store FigaroTV TVmag Le Scan TV Santé
14 Etudiant FigaroChic.cn Histoire Bourse Nautisme Golf Scope Voyage Enchères Vin Evéne Jardin Météo consult Annuaire santé avec PagesJaunes Le
15 particulier Cadremploi La chaîne météo Keljob Kelformation Explorimmo Propriétés de France Ticketac Vodeo Cplussur
16 Facebook Twitter Google Plus Instagram Youtube Tumblr Newsletters iOS (iPhone, iPad..) Android
17 Rechercher
18 Menu
19 En direct
20 Le journal
21 Suivre
22 Recherche
23 Actualité Economie Sport Culture Lifestyle Madame FigaroTV Vox Edition abonnés Abonnez-vous
```

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 :



```
corpus-chronologique.txt
12485 PETITES ANNONCES
12486 Consulter les annonces Déposer une annonce Annonces immo Annonces autos Annonces emploi
12487 SERVICES
12488 Abonnements NR Espace abonnés Abonnements numériques pro Newsletters Appli La NR Météo Panel lecteurs Concours photos
12489 Cette fonctionnalité est réservée aux membres de ma NR.
12490 Identifier vous pour pouvoir utiliser cette fonctionnalité.
12491 Adresse email
12492 Mot de passe
12493 > mot de passe oublié
12494 Je ne suis pas membre de NR,
12495 je souhaite m'inscrire :
12496
12497 </text>
12498 </page>
12499 </date>
12500
12501 <date="2015-01-09">
12502 <page url="http://www.lefigaro.fr" id="1" time="Friday, January 09, 2015 1:03:48 PM" >
12503 <url>
12504 http://www.lefigaro.fr
12505
12506 </url>
12507 <title>
12508 Le Figaro - Actualités
12509
12510 </title>
12511 <text>
12512 Le Figaro - Actualités
12513 Le Scan Politique Economie Sport24 Le Scan Sport Culture Lifestyle Madame Editions abonnés Figaro Store FigaroTV TVmag Le Scan TV Santé Etudiant FigaroChic.cn Histoire Bourses
12514 Nautisme Golf Scope Voyage Enchères Vin Evéne Jardin Météo consult Annuaire santé avec PagesJaunes Le particulier Cadremploi La chaîne météo Keljob Kelformation Explorimmo Propriétés
12515 de France Ticketac Vodeo Cplussur
```

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

```
26853 Adresse email
26856 Mot de passe
26857 > mot de passe oublié
26858 Je ne suis pas membre de MR,
26859 je souhaite m'inscrire :
26860
26861 </text>
26862 </page>
26863
26864 </date>
```

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 :

```
Abonnements
Le Figaro et ses magazines
Édition du jeudi 8 janv. 1999
< >
Espace abonnés
.
```

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

<input type="checkbox"/>	<date(2005-01-08[pos=1])>	...
<input type="checkbox"/>	<page(lefigaro.fr[pos=2])>	...
<input type="checkbox"/>	<page(lemonde.fr[pos=8999])>	...
<input type="checkbox"/>	<page(lopinion.fr[pos=19359])>	...
<input type="checkbox"/>	<page(liberation.fr[pos=21603])>	...
<input type="checkbox"/>	<page(leparisien.fr[pos=29149])>	...
<input type="checkbox"/>	<page(humanite.fr[pos=38890])>	...
<input type="checkbox"/>	<page(la-croix.com[pos=44198])>	...
<input type="checkbox"/>	<page(lejdd.fr[pos=49517])>	...
<input type="checkbox"/>	<page(lecanardenchaine.fr[pos=53039])>	...
<input type="checkbox"/>	<page(nrpyrenees.com[pos=53779])>	...
<input type="checkbox"/>	<page(paris-normandie.com[pos=55692])>	...
<input type="checkbox"/>	<page(lalsace.fr[pos=60200])>	...
<input type="checkbox"/>	<page(petitbleu.fr[pos=63363])>	...
<input type="checkbox"/>	<page(lepopulaire.fr[pos=65806])>	...
<input type="checkbox"/>	<page(lapressedelamanche.fr[pos=73385])>	...
<input type="checkbox"/>	<page(presseocean.fr[pos=73818])>	...
<input type="checkbox"/>	<page(havre-presse.fr[pos=77365])>	...
<input type="checkbox"/>	<page(laprovence.com[pos=79610])>	...
<input type="checkbox"/>	<page(republicain-lorrain.fr[pos=84149])>	...
<input type="checkbox"/>	<page(republique-pyrenees.com[pos=86610])>	...
<input type="checkbox"/>	<page(larep.com[pos=86905])>	...
<input type="checkbox"/>	<page(sudouest.com[pos=93959])>	...
<input type="checkbox"/>	<page(letelegramme.com[pos=100164])>	...
<input type="checkbox"/>	<page(varmatin.fr[pos=104419])>	...
<input type="checkbox"/>	<page(ledauphine.com[pos=105860])>	...
<input type="checkbox"/>	<page(lavoixdunord.fr[pos=107741])>	...
<input type="checkbox"/>	<page(lyonne-republicaine.fr[pos=119625])>	...
<input type="checkbox"/>	<page(leberry.fr[pos=127108])>	...
<input type="checkbox"/>	<page(centre-presse.fr[pos=134792])>	...
<input type="checkbox"/>	<page(charentelibre.com[pos=143177])>	...

<input type="checkbox"/>	<page(corsematin.com[pos=149431])>	...
<input type="checkbox"/>	<page(courrierdelouest.fr[pos=155497])>	...
<input type="checkbox"/>	<page(ladepeche.com[pos=159145])>	...
<input type="checkbox"/>	<page(dna.fr[pos=165503])>	...
<input type="checkbox"/>	<page(dordogne.com[pos=168884])>	...
<input type="checkbox"/>	<page(lechorepublicain.fr[pos=172013])>	...
<input type="checkbox"/>	<page(pyrenees.com[pos=179540])>	...
<input type="checkbox"/>	<page(veille.fr[pos=182619])>	...
<input type="checkbox"/>	<page(havre-libre.fr[pos=186304])>	...
<input type="checkbox"/>	<page(havre-presse.fr[pos=188549])>	...
<input type="checkbox"/>	<page(lamarseillaise.fr[pos=190794])>	...
<input type="checkbox"/>	<page(lindependant.com[pos=196080])>	...
<input type="checkbox"/>	<page(jhm.fr[pos=200138])>	...
<input type="checkbox"/>	<page(lejdc.fr[pos=203250])>	...
<input type="checkbox"/>	<page(lemainelibre.fr[pos=210059])>	...
<input type="checkbox"/>	<page(midilibre.com[pos=213469])>	...
<input type="checkbox"/>	<page(lamontagne.fr[pos=217631])>	...
<input type="checkbox"/>	<page(nicematin.fr[pos=226471])>	...
<input type="checkbox"/>	<page(nordeclair.fr[pos=232376])>	...
<input type="checkbox"/>	<page(nordlittoral.fr[pos=236468])>	...
<input type="checkbox"/>	<page(lanouvellerepublique.fr[pos=244136])>	...

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.