Architecture du projet

" Un corpus de veille : le journal *Le Monde* "

Partie 1 : « *Le Monde en Surface* »

Partie 2 : « *Le Monde Profond* »

Document de travail - Décembre 2005

Fleury Serge, CLA²T/SYLED Serge.fleury@univ-paris3.fr

1 Sommaire

2 Table des figures

3 Préambule

Ce document présente d'une part (*Partie 1*) l'architecture construite pour traiter les fils RSS[1] mis à disposition sur le site Web du journal Le Monde (d'autres fils sont aussi traités dans cette architecture, en particulier celui du site de l'AFP) et d'autre part (*Partie 2*) l'architecture mise en œuvre pour construire et analyser un corpus chronologique de la version électronique du journal *Le Monde*.

Partie 1 « *Le Monde en Surface* »

L'architecture construite est composée de 2 modules.

Le premier (« **Fil(s) de presse** ») correspond au module permettant de traiter un fil de presse donné (au format RSS) et de construire des traitements sur le contenu de ce fil (au départ, un « nuage de mots »).

Le second (« **Archivage des Fils de Presse** ») correspond au module permettant d'archiver les fils de manière continue et automatique afin de constituer la mémoire de ces fils.

Ce projet a commencé en Octobre 2005 *i.e.* on dispose à ce jour d'un corpus de fils RSS archivés toutes les heures et d'une série d'outils de traitement de ces fils (en développement).

Partie 2 « *Le Monde Profond* »

Chaque version quotidienne du journal *Le Monde* est régulièrement récupérée sur le site web du journal[2] : dans sa version HTML et dans sa version PDF. La version HTML[3] du journal est traitée pour produire différents états :

un état quotidien des contenus textuels du journal sous la forme d'une version normalisée au format XML et une version compatible avec le logiciel ***Lexico3***

des états statistiques quotidiens

Les états quotidiens des contenus textuels sont ensuite nettoyés et concaténés pour produire des corpus chronologiques couvrant l'ensemble des dates de récupération.

Le démarrage de ce processus a commencé le 12 avril 2003 *i.e.* on dispose à ce jour d'un corpus regroupant l'ensemble des versions électroniques de chaque journée depuis cette date.

---

[1] RSS est un format de diffusion (syndication) de contenus. Le principe est simple : les sites/blogs mettent en place des flux RSS avec un format de données automatiquement structuré (en RDF ou en XML) et les utilisateurs peuvent les lire dans des outils dédiés (aggrégateurs, utilitaires mail, navigateurs). *Cf* cours URFIST *infra* http://www.ccr.jussieu.fr/urfist/rss/

[2] http://www.lemonde.fr/

[3] La version HTML traitée ici est celle dite "simplifiée (sans image de la une et sans menu déroulant)"

4 Le(s) site(s) du projet

Le site « officiel » du projet (dit « site CLMC ») est en accès restreint[4] à l'adresse suivante :

Hypertoile : http://sfmac.no-ip.com/corpusLeMonde/[5]

Une partie du projet « *Le Monde en Surface* » est visible en ligne et sans restriction à cette adresse : http://tal.univ-paris3.fr/filpresse/

De même, une partie des résultats construits pour la partie du projet « *Le Monde Profond* » est aussi disponible sans restriction aux adresses suivantes :

Weblog (pluri)TAL:

http://tal.univ-paris3.fr/blogtal/

On trouvera en particulier sur ce blog des rapports de navigation textométrique ou des résultats produits dans le cadre de ce projet.

Rapports *Lexico3* :

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/rapportsL3.htm

Rapports *Lexico3* et spécificité chronologique :

http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm

[4] accès restreint sur demande à serge.fleury@univ-paris3.fr)

[5] *cf infra* « parcours du site »

5 Partie 1 : « Le Monde en Surface »

Ce projet est composé de 2 modules.

Le premier (« **Fil(s) de presse** ») correspond au module permettant de traiter un fil de presse donné (au format RSS) et de construire des traitements sur le contenu de ce fil (au départ, un nuage de mots).

Le second (« **Archivage des Fils de Presse** ») correspond au module permettant d'archiver les fils de manière continue et automatique afin de constituer la mémoire de ces fils.

5.1 La plateforme d'archivage des Fils de Presse

Un processus expérimental a été mis en place pour archiver les fils de presse. L'idée est la suivante :

- on a à disposition le corpus Le Monde depuis Avril 2003[6] (*cf* Partie 2 : « ***le Monde PROFOND*** »)

- on peut aussi avoir accès au fils RSS publiés quotidiennement (« ***le Monde EN SURFACE*** »)

En archivant régulièrement les fils on a donc à portée de main le *PROFOND* et la *SURFACE*. Le processus mis en place aspire régulièrement les fils visés et créé des pages de navigation pour donner à voir les données archivées et les nuages de mots créés sur chacun des fils (*cf infra* le projet « Fil(s) de Presse » : programme construisant un nuage de mots à partir des contenus textuels présents dans un fil donné). Les données sont visibles ici :

http://sfmac.no-ip.com/fils-presse-arch/index.xml (accès restreint)

L'archivage mis en place concerne les fils du journal Le Monde (*cf infra*) et celui de l'AFP[7]. La figure suivante donne une représentation de l'organisation de cet archivage :

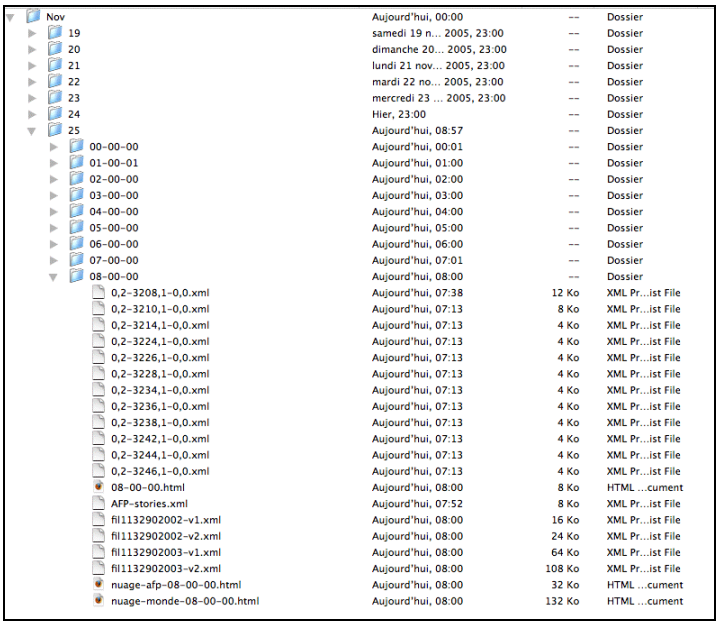


Figure 1 : Archivage des fils, arborescence

[6] http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/veille.htm

[7] http://www.afp.fr/francais/rss/stories.xml

Le processus d'archivage est déclenché toutes les heures et produit à chaque lancement un archivage des fils, des pages de navigation et les données nécessaires pour construire les nuages de mots.

5.2 Le projet Fil(s) de Presse

Le programme construit prend en entrée des fils RSS disponibles sur des sites de presse (Le Monde[8], Le Figaro[9], Libération[10]…) et produit des résultats donnant à voir :

- des nuages de mots
- une présentation des fils scrutés au format HTML et des comptages lexicométriques à partir des contenus textuels **des descriptions des articles** (disponibles dans les fils) mis à la disposition par les journaux.

Les figures suivantes présentent les différents types de nuages construits :

Figure 2 : Nuage de mots sans lien

Dans cette première figure, le nuage de mots donne à voir l'ensemble des mots présents dans les descriptions des articles des fils d'un journal en ligne à un moment donne (ici Le Figaro).

[8] http://www.lemonde.fr/web/rss/0,48-0,1-0,0.html

[9] http://www.lefigaro.fr/xml/

[10] http://www.liberation.fr/page.php?Article=149907

Figure 3 : Nuage de mots avec liens

Dans la seconde, on peut voir un nuage similaire dans lequel chaque mot donne accès *via* un clic aux contextes dans lesquels ce mot apparaît (colonne de droite) : le contexte est constitué par le titre de l'article, sa description et son URL.

Figure 4 : Nuages de mots avec "carte des sections" (1 section = 1 carré = 1 article)

Dans la troisième, on y voit toujours le même nuage de mots sur la gauche, dans lequel chaque mot donne accès *via* un clic à une « *représentation cartographique*[11] » du contenu du fil scruté dans

[11] Ce développement s'inscrit dans les travaux faits autour de Lexico3 pour construire des représentations des textes donnant à voir les unités textuelles manipulées à travers des objets graphiques :

http://lexico3.no-ip.org/, http://tal.univ-paris3.fr/CE-query/

laquelle le contenu textuel de la description d'un article est représenté par un carré, les articles contenant le mot cliqué sont associées à des carrés rouges ■ et les autres à des carrés blancs □ . Chaque carré est donc associé à un article en ligne (si on clique sur le carré on accède à l'article en ligne).

Dans les trois figures, la taille de la police de caractères utilisée pour afficher le mot dans le nuage est déterminée par la fréquence du mot dans l'ensemble des articles scrutés pour un journal donné.

5.3 Architecture du projet « nuage de mots »

Dans le projet initial [Herrington, 2005], l'architecture « en amont » de l'application a l'allure suivante :

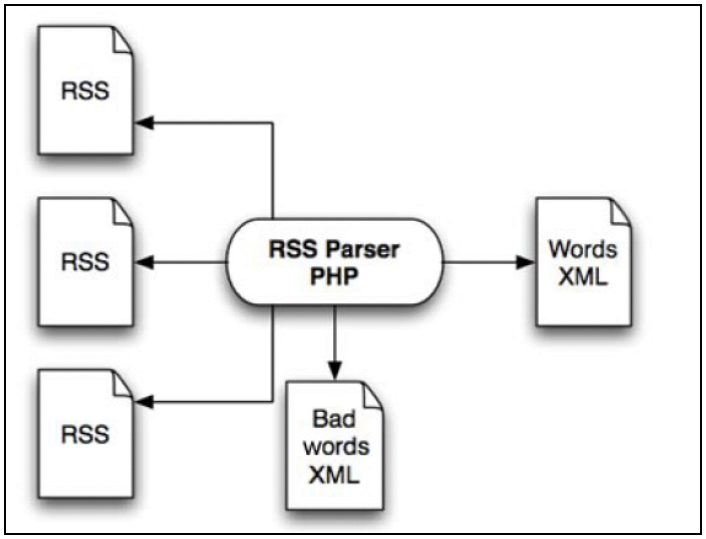


Figure 5 : Architecte initiale « en amont »

L'application lit des flux RSS et déclenche un *parser* RSS (écrit en PHP) qui a pour tâche de sélectionner les zones de texte à explorer puis de lancer une opération de segmentation de ces contenus textuels en ne retenant que les mots non présents dans une liste prédéterminée (mots vides). L'architecture maintenue pour le projet présenté ici est la suivante :

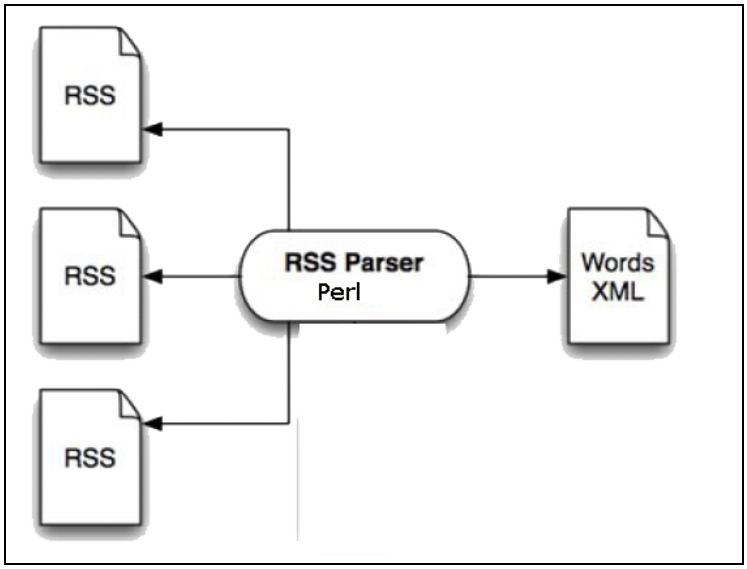


Figure 6 : Architecture modifiée « en amont »

Le principe général est conservé, tout le code est réécrit en Perl, *parser* compris. Tous les mots présents dans les contenus textuels scrutés sont conservés. Les mots retenus et comptés sont sauvegardés au format XML. Le fichier produit a l'allure suivante :

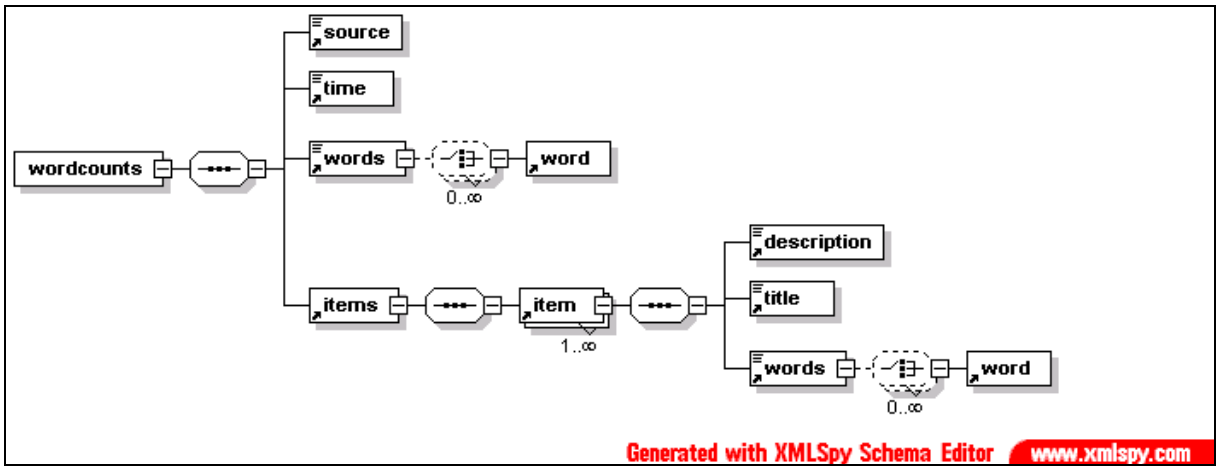


Figure 7 : Schéma du lexique construit

Dans ce schéma, l élément `words` contient la liste de tous les mots (et leur fréquence) pour un fil de presse donné, l'élément `item` contenant la liste de tous les mots pour un article donné contenu dans ce fil.

On présente ci-dessous un extrait du lexique construit :

```
<?xml version="1.0" encoding="iso-8859-1"?>
<?xml-stylesheet type="text/xsl" href="parserss.xsl"?>
<wordcounts>
<source>LIBERATION</source>
<time>Wed Oct 26 08:06:28 2005</time>
<words>
<word text="de" count="16" />
<word text="le" count="5" />
<word text="la" count="5" />
<word text="d" count="5" />
…
</words>
<items>
<item url="http://www.liberation.fr/page.php?Article=332624" title="">
<description><![CDATA[Mort du dessinateur aux personnages diaphanes et angéliques, rendu
célèbre par un générique d'Antenne 2.]]></description>
<title><![CDATA[Feu Folon]]></title>
<words>
<word text="Mort" text2="Mort" />
<word text="du" text2="du" />
<word text="dessinateur" text2="dessinateur" />
<word text="aux" text2="aux" />
<word text="personnages" text2="personnages" />
<word text="diaphanes" text2="diaphanes" />
<word text="et" text2="et" />
<word text="angéliques" text2="angeliques" />
<word text="rendu" text2="rendu" />
<word text="célèbre" text2="celebre" />
<word text="par" text2="par" />
<word text="un" text2="un" />
<word text="générique" text2="generique" />
<word text="d" text2="d" />
<word text="Antenne" text2="Antenne" />
<word text="2" text2="2" />
</words>
</item>
…
</items>
</wordcounts>
```

Une modification mineure a été réalisée dans la grammaire du fichier lexique produit par rapport à l'application initiale. La présence de caractères accentués dans les mots posant des problèmes pour la seconde partie de l'application (celle utilisant le script établissant le lien entre le mot et ses contextes), un attribut a été ajouté dans les éléments décrivant les mots, celui-ci contenant après transcodage, la forme graphique normalisée du mot sans caractères accentués (`générique` est réécrit `generique`).

Dans un deuxième temps, l'application construit le nuage des mots en utilisant le lexique produit et en appliquant sur la sortie XML contenant ce lexique une feuille de style XSL (utilisant un script Javascript).

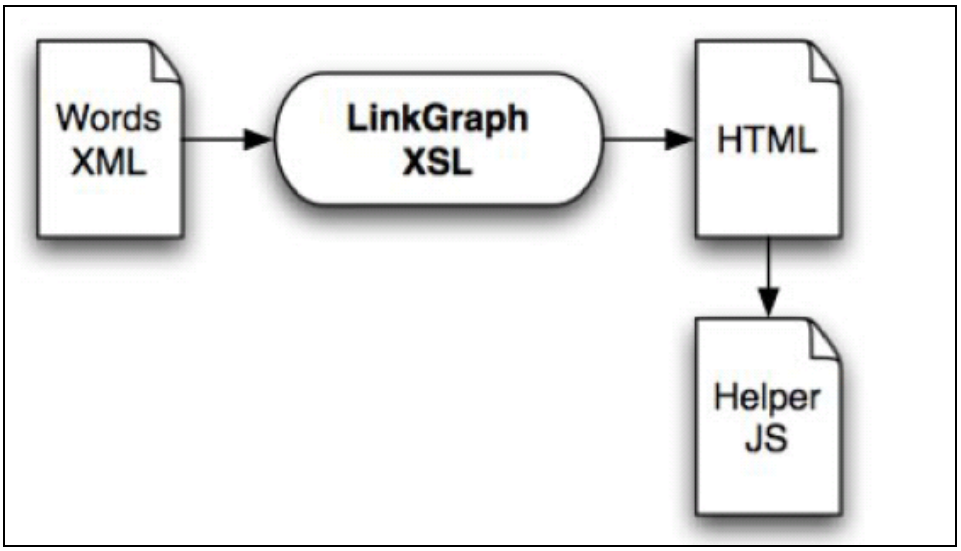


Figure 8 : Architecture "en aval"

Cette architecture « en aval » maintient intégralement le principe présenté dans [Herrington, 2005]. Plusieurs modifications ont cependant été apportées :

La feuille de style initiale a d'abord été réécrite pour ne produire qu'un nuage de mot sans lien *i.e.* sans liens vers les contextes originaux contenant les mots scrutés.

Elle a aussi été modifiée pour affiner les sorties produites (contextes ou carte des sections)

5.4 Boîte à outils

5.4.1 Chronofil

Chronofil est un programme CGI permettant de construire en ligne un graphique de ventilation d'une forme donnée dans l'ensemble des Fils de presse archivés. L'interface web de ce programme est disponible à cette adresse :

http://sfmac.no-ip.com/cgi-bin/chronofil.cgi

On y accède aussi à partir de la page d'accueil du site du projet (rubrique *Chronofil* dans le menu de gauche). La figure suivante présente cette page :

Figure 9 : Interface Chronofil

Une fois entrée une forme graphique, le programme déclenché scrute les fils de presse archivés et construit in fine le graphique de ventilation correspondant à la forme. Les figures qui suivent montrent la ventilation de la forme LAICITE :

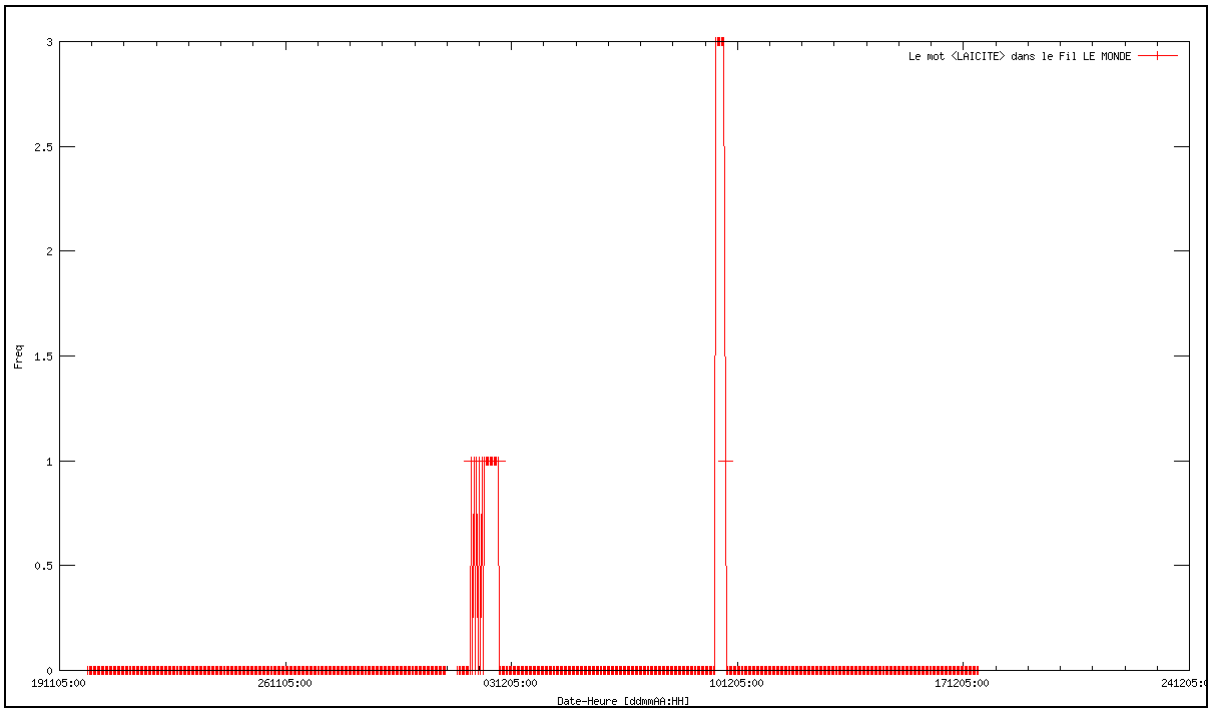


Figure 10 : Chronofil -Ventilation de la forme LAICITE (Fils Le Monde)

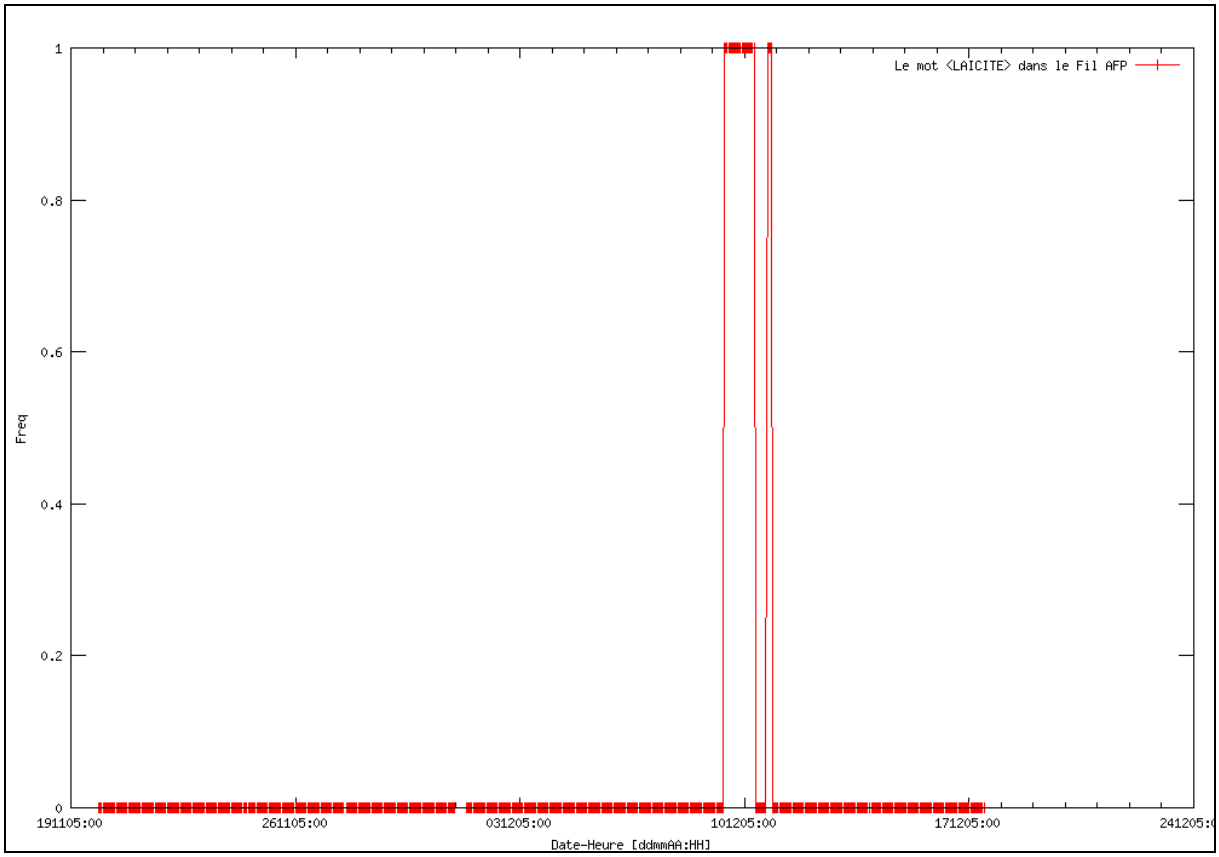


Figure 11 : Chronofil -Ventilation de la forme LAICITE (Fil AFP)

5.4.2 Chronofiltre

Chronofiltre est un programme CGI permettant de sélectionner en ligne les contenus textuels des fils archivés contenant un mot donné. L'interface web de ce programme est disponible à cette adresse :

http://sfmac.no-ip.com/cgi-bin/chronofiltre.cgi

On y accède aussi à partir de la page d'accueil du site du projet (rubrique *Chronofiltre* dans le menu de gauche). La figure suivante présente cette page :

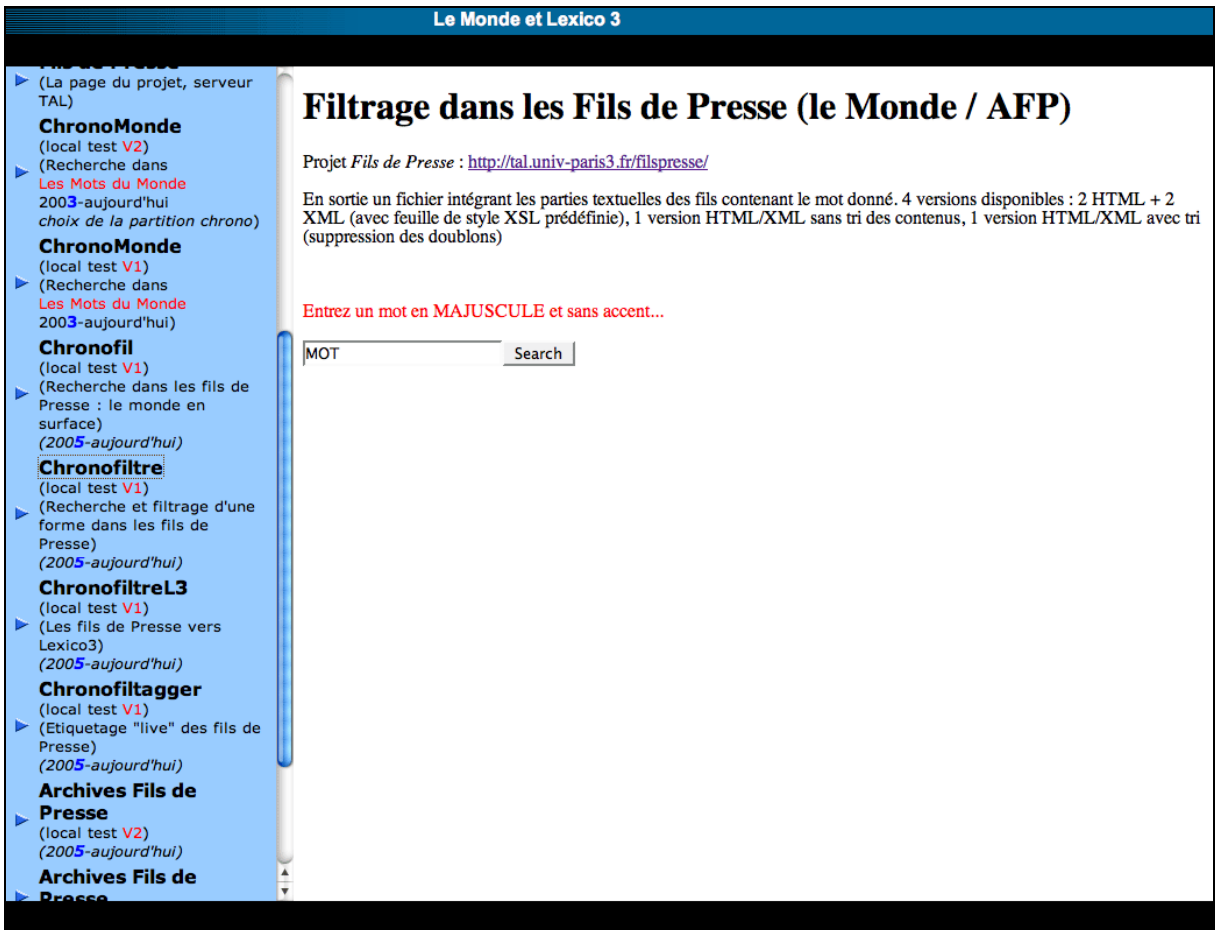
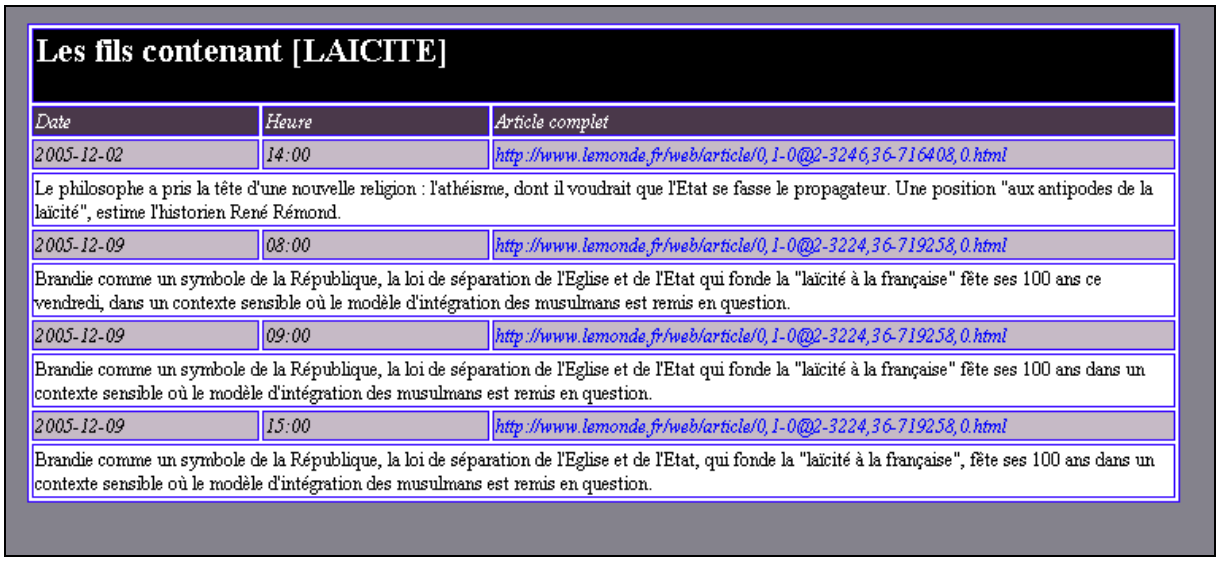


Figure 12 : Interface Chronofiltre

Une fois entrée une forme graphique, le programme déclenché scrute les fils de presse archivés et construit *in fine* une page regroupant les contenus des fils contenant la forme cherchée. Deux états sont disponibles : un premier état concatène l'ensemble des contenus textuels sans filtrage des doublons, un deuxième filtre les éventuels doublons. De plus ces 2 états sont disponibles sous 2 formats : HTML et XML (avec feuille de style XSLT).

La figure suivante présente le résultat (filtré) sur la forme LAICITE :



Les fils contenant [LAICITE]

Date	*Heure*	*Article complet*
2005-12-02	*14:00*	*http://www.lemonde.fr/web/article/0,1-0@2-3246,36-716408,0.html*
Le philosophe a pris la tête d'une nouvelle religion : l'athéisme, dont il voudrait que l'Etat se fasse le propagateur. Une position "aux antipodes de la laïcité", estime l'historien René Rémond.		
2005-12-09	*08:00*	*http://www.lemonde.fr/web/article/0,1-0@2-3224,36-719258,0.html*
Brandie comme un symbole de la République, la loi de séparation de l'Eglise et de l'Etat qui fonde la "laïcité à la française" fête ses 100 ans ce vendredi, dans un contexte sensible où le modèle d'intégration des musulmans est remis en question.		
2005-12-09	*09:00*	*http://www.lemonde.fr/web/article/0,1-0@2-3224,36-719258,0.html*
Brandie comme un symbole de la République, la loi de séparation de l'Eglise et de l'Etat qui fonde la "laïcité à la française" fête ses 100 ans dans un contexte sensible où le modèle d'intégration des musulmans est remis en question.		
2005-12-09	*15:00*	*http://www.lemonde.fr/web/article/0,1-0@2-3224,36-719258,0.html*
Brandie comme un symbole de la République, la loi de séparation de l'Eglise et de l'Etat, qui fonde la "laïcité à la française", fête ses 100 ans dans un contexte sensible où le modèle d'intégration des musulmans est remis en question.		

Figure 13 : Sortie Chronofiltre

5.4.3 ChronofiltreL3

ChronofiltreL3 est un programme CGI permettant de construire en ligne un corpus au format ***Lexico3*** rassemblant l'ensemble des Fils de presse archivés. L'interface web de ce programme est disponible à cette adresse :

http://sfmac.no-ip.com/cgi-bin/chronofiltreL3.cgi

On y accède aussi à partir de la page d'accueil du site du projet (rubrique *ChronofiltreL3* dans le menu de gauche). La figure suivante présente cette page :

Figure 14 : Interface ChronofiltreL3

La figure suivante présente un extrait du résultat produit :

```
 <descDateFil="2005111921">
  LE MANS (AFP) - La possibilité d'un rassemblement de tous les socialistes au congrès du Mans é
tait suspendue samedi à  un compromis entre François Hollande et le courant du Nouveau Parti so
cialiste (NPS).L'événement

 <descDateFil="2005111922">
 Après deux années de négociations en vue d'un vaste accord de "ciel ouvert", la Commission euro
péenne et les Etats-Unis ont franchi une première étape, vendredi, que les 25 pays de l'Union d
oivent valider.

 <descDateFil="2005111922">
 BAGDAD (AFP) - Au moins trente personnes ont été tuées et 40 blessées samedi dans un attentat s
uicide à  la voiture piégée contre une tente de condoléances, à  Abou Saïda, localité de la pro
vince de Diyala, au nord-est de Bagdad, selon le chef de la police de la province.

 <descDateFil="2005111923">
 En Espagne, le gouvernement ouvre le débat sur les moyens d'inciter à la poursuite d'activité j
usqu'à 70 ans. En Allemagne et en Grande-Bretagne, il faudra bientôt travailler jusqu'à 67 ans

 <descDateFil="2005112000">
 Ce devait être le premier accord interprofessionnel de l'"ère Parisot".
```

Figure 15 : Sortie ChronofiltreL3

5.4.4 Chronofiltagger

Chronofiltagger est un programme CGI permettant de construire en ligne un corpus étiqueté avec *treetagger* rassemblant l'ensemble des Fils de presse archivés. L'interface web de ce programme est disponible à cette adresse :

http://sfmac.no-ip.com/cgi-bin/chronofiltagger.cgi

On y accède aussi à partir de la page d'accueil du site du projet (rubrique *Chronofiltagger* dans le menu de gauche). La figure suivante présente cette page :

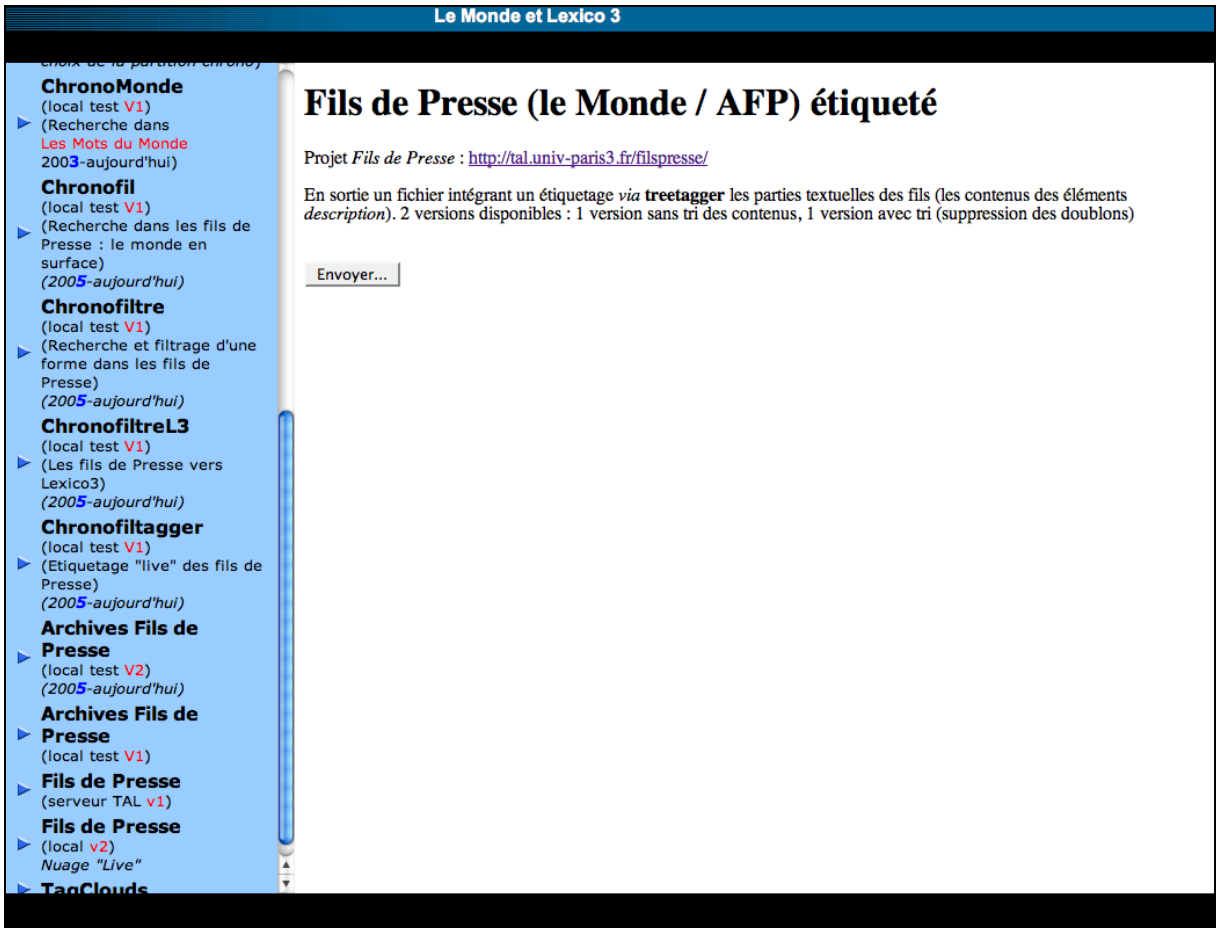
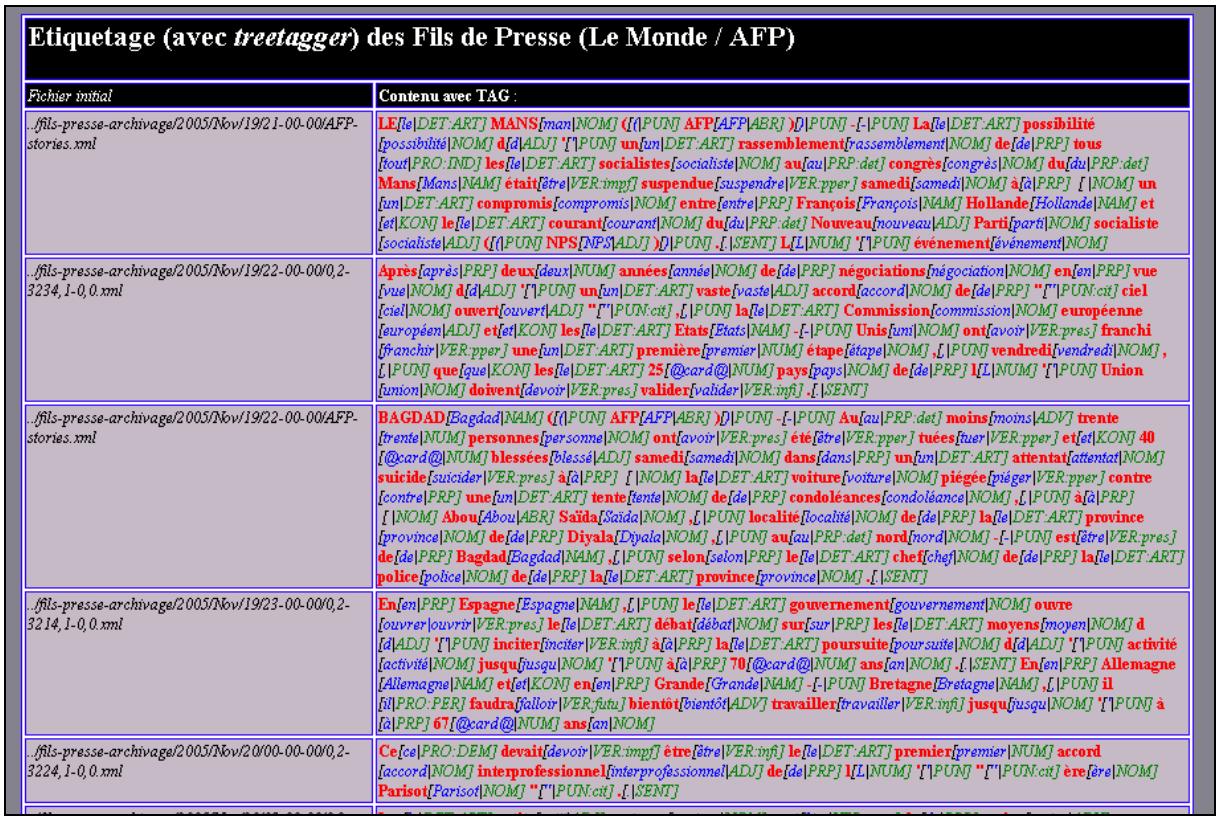


Figure 16 : Interface Chronofiltagger

La figure suivante présente un extrait du résultat produit :



Etiquetage (avec *treetagger*) des Fils de Presse (Le Monde / AFP)

Fichier initial	**Contenu avec TAG** :
../fils-presse-archivage/2005/Nov/19/21-00-00/AFP-stories.xml	**LE**[le|DET:ART] **MANS**[man|NOM] ([(|PUN] **AFP**[AFP|ABR])D|PUN] -[-|PUN] **La**[le|DET:ART] **possibilité**[possibilité|NOM] **d**[d|ADJ] '['|PUN] **un**[un|DET:ART] **rassemblement**[rassemblement|NOM] **de**[de|PRP] **tous**[tout|PRO:IND] **les**[le|DET:ART] **socialistes**[socialiste|NOM] **au**[au|PRP:det] **congrès**[congrès|NOM] **du**[du|PRP:det] **Mans**[Mans|NAM] **était**[être|VER:impf] **suspendue**[suspendre|VER:pper] **samedi**[samedi|NOM] **à**[à|PRP] [|NOM] **un**[un|DET:ART] **compromis**[compromis|NOM] **entre**[entre|PRP] **François**[François|NAM] **Hollande**[Hollande|NAM] **et**[et|KON] **le**[le|DET:ART] **courant**[courant|NOM] **du**[du|PRP:det] **Nouveau**[nouveau|ADJ] **Parti**[parti|NOM] **socialiste**[socialiste|ADJ] ([(|PUN] **NPS**[NPS|ADJ])D|PUN] .[.|SENT] **L**[L|NUM] '['|PUN] **événement**[événement|NOM]
../fils-presse-archivage/2005/Nov/19/22-00-00/0,2-3234,1-0,0.xml	**Après**[après|PRP] **deux**[deux|NUM] **années**[année|NOM] **de**[de|PRP] **négociations**[négociation|NOM] **en**[en|PRP] **vue**[vue|NOM] **d**[d|ADJ] '['|PUN] **un**[un|DET:ART] **vaste**[vaste|ADJ] **accord**[accord|NOM] **de**[de|PRP] "["|PUN:cit] **ciel**[ciel|NOM] **ouvert**[ouvert|ADJ] "["|PUN:cit] ,[,|PUN] **la**[le|DET:ART] **Commission**[commission|NOM] **européenne**[européen|ADJ] **et**[et|KON] **les**[le|DET:ART] **Etats**[Etats|NAM] -[-|PUN] **Unis**[uni|NOM] **ont**[avoir|VER:pres] **franchi**[franchir|VER:pper] **une**[un|DET:ART] **première**[premier|NUM] **étape**[étape|NOM] ,[,|PUN] **vendredi**[vendredi|NOM] ,[,|PUN] **que**[que|KON] **les**[le|DET:ART] **25**[@card@|NUM] **pays**[pays|NOM] **de**[de|PRP] **l**[L|NUM] '['|PUN] **Union**[union|NOM] **doivent**[devoir|VER:pres] **valider**[valider|VER:infi] .[.|SENT]
../fils-presse-archivage/2005/Nov/19/22-00-00/AFP-stories.xml	**BAGDAD**[Bagdad|NAM] ([(|PUN] **AFP**[AFP|ABR])D|PUN] -[-|PUN] **Au**[au|PRP:det] **moins**[moins|ADV] **trente**[trente|NUM] **personnes**[personne|NOM] **ont**[avoir|VER:pres] **été**[être|VER:pper] **tuées**[tuer|VER:pper] **et**[et|KON] **40**[@card@|NUM] **blessées**[blessé|ADJ] **samedi**[samedi|NOM] **dans**[dans|PRP] **un**[un|DET:ART] **attentat**[attentat|NOM] **suicide**[suicider|VER:pres] **à**[à|PRP] [|NOM] **la**[le|DET:ART] **voiture**[voiture|NOM] **piégée**[piéger|VER:pper] **contre**[contre|PRP] **une**[un|DET:ART] **tente**[tente|NOM] **de**[de|PRP] **condoléances**[condoléance|NOM] ,[,|PUN] **à**[à|PRP] [|NOM] **Abou**[Abou|ABR] **Saïda**[Saïda|NOM] ,[,|PUN] **localité**[localité|NOM] **de**[de|PRP] **la**[le|DET:ART] **province**[province|NOM] **de**[de|PRP] **Diyala**[Diyala|NOM] ,[,|PUN] **au**[au|PRP:det] **nord**[nord|NOM] -[-|PUN] **est**[être|VER:pres] **de**[de|PRP] **Bagdad**[Bagdad|NAM] ,[,|PUN] **selon**[selon|PRP] **le**[le|DET:ART] **chef**[chef|NOM] **de**[de|PRP] **la**[le|DET:ART] **police**[police|NOM] **de**[de|PRP] **la**[le|DET:ART] **province**[province|NOM] .[.|SENT]
../fils-presse-archivage/2005/Nov/19/23-00-00/0,2-3214,1-0,0.xml	**En**[en|PRP] **Espagne**[Espagne|NAM] ,[,|PUN] **le**[le|DET:ART] **gouvernement**[gouvernement|NOM] **ouvre**[ouvrer|ouvrir|VER:pres] **le**[le|DET:ART] **débat**[débat|NOM] **sur**[sur|PRP] **les**[le|DET:ART] **moyens**[moyen|NOM] **d**[d|ADJ] '['|PUN] **inciter**[inciter|VER:infi] **à**[à|PRP] **la**[le|DET:ART] **poursuite**[poursuite|NOM] **d**[d|ADJ] '['|PUN] **activité**[activité|NOM] **jusqu**[jusqu|NOM] '['|PUN] **à**[à|PRP] **70**[@card@|NUM] **ans**[an|NOM] .[.|SENT] **En**[en|PRP] **Allemagne**[Allemagne|NAM] **et**[et|KON] **en**[en|PRP] **Grande**[Grande|NAM] -[-|PUN] **Bretagne**[Bretagne|NAM] ,[,|PUN] **il**[il|PRO:PER] **faudra**[falloir|VER:futu] **bientôt**[bientôt|ADV] **travailler**[travailler|VER:infi] **jusqu**[jusqu|NOM] '['|PUN] **à**[à|PRP] **67**[@card@|NUM] **ans**[an|NOM]
../fils-presse-archivage/2005/Nov/20/00-00-00/0,2-3224,1-0,0.xml	**Ce**[ce|PRO:DEM] **devait**[devoir|VER:impf] **être**[être|VER:infi] **le**[le|DET:ART] **premier**[premier|NUM] **accord**[accord|NOM] **interprofessionnel**[interprofessionnel|ADJ] **de**[de|PRP] **l**[L|NUM] '['|PUN] "["|PUN:cit] **ère**[ère|NOM] **Parisot**[Parisot|NOM] "["|PUN:cit] .[.|SENT]

Figure 17 : Sortie Chronofiltagger

6 Partie 2 : « Le Monde Profond »

6.1 Chaînes de traitement

6.1.1 Récolte des données

Le journal *Le Monde* met à la disposition de ses abonnés à la version électronique du journal différentes versions électroniques du journal : une version au format PDF et 2 versions au format HTML (une avec image et l'autre sans). La version PDF et la version HTML sans image sont récupérées sur le site du journal quotidiennement et stockées sur disque. Seule la version HTML simplifiée est traitée par la chaîne de traitement décrite *infra*.

6.1.2 Préparation et manipulation des données

Après récupération, chaque version quotidienne est soumis à différentes manipulations.

6.1.2.1 Traitements automatiques

Des scripts Perl ont été construits pour réaliser en plusieurs étapes différentes opérations :

Une première étape consiste à "normaliser" la version originale en utilisant le programme webxref utilisé dans les projets TYPWEB et SENSNET (*cf* Phase 1).

Une seconde étape construit une version XML et ***Lexico3*** de la version électronique quotidienne du journal (*cf* Phase 1).

Une troisième étape construit l'index quotidien des mots du journal (*cf* Phase 1).

Une quatrième étape permet d'obtenir une version complète du corpus au format XML (par concaténation et structuration) (*cf* Phase 1).

Enfin une dernière étape permet de construire à la volée les différents corpus thématiques (quotidiens et complets) (*cf* Phase 2).

6.1.2.2 Traitements complémentaires

Chaque version quotidienne du corpus au format ***Lexico3*** est insérée dans les corpus de chronologique complet de référence (corpus complet, corpus France, corpus INTERNATIONAL). Certaines balises sont ajoutées pour marquer le changement de semaine, de mois ou d'année.

Un programme se charge ensuite d'insérer la balise RUBRIQUE dans le corpus complet au format ***Lexico3***.

Cette étape de préparation des données n'est pas réalisée quotidiennement, mais en général une fois par quinzaine.

6.1.3 Synthèse des traitements

Phase n°0	
Récupération des données sur le site du Monde.fr	
Phase n°1	
`load-webxref.pl`	Lancement de la normalisation WEBXREF
`load-makeCorpusL3-date.pl`	Génération de corpus quotidien XML et pour L3 Index des mots
Phase n°2	
`load-makeAllThemaCorpus.pl`	Génération des corpus thématiques (toutes rubriques confondues)
`load-makeAllThemaCorpus-v2.pl`	Génération des corpus thématiques (toutes rubriques confondues) (v2)
`load-makeThemaCorpus.pl`	Génération des corpus thématiques "France" et "International"
`load-makeThemaCorpus-international-v2.pl`	Génération du corpus thématique "International" (v2)
`load-makeThemaCorpus-france-v2.pl`	Génération du corpus thématique "France" (v2)
`load-makeThemaCorpus-societe.pl`	Génération du corpus thématique "Société"
`load-makeThemaCorpus-societe-v2.pl`	Génération du corpus thématique "Société" (v2)
`load-makeThemaCorpus-select.pl`	Génération de corpus thématique à la volée
`load-makeThemaCorpus-select-v2.pl`	Génération de corpus thématique à la volée (v2)
Phase n°3	
"Rebalisage" manuel ou automatique des corpus produits	
Phase n°4	
Traitements lexicométriques	

6.1.4 Traitements lexicométriques

A l'issue de l'étape précédente, on dispose quotidiennement d'un état du corpus prêt pour être analysé par ***Lexico3***. (*cf* en annexe un exemple de rapport construit avec ***Lexico3***)

6.2 Boîte à outils

6.2.1 ChronoMonde

Chronomonde est un programme CGI permettant de construire en ligne un graphique de ventilation d'une forme donnée dans le corpus complet. L'interface web de ce programme est disponible à cette adresse :

http://sfmac.no-ip.com/cgi-bin/chronomonde.cgi

On y accède aussi à partir de la page d'accueil du site du projet (rubrique *Chronomonde* dans le menu de gauche). La figure suivante présente cette page :

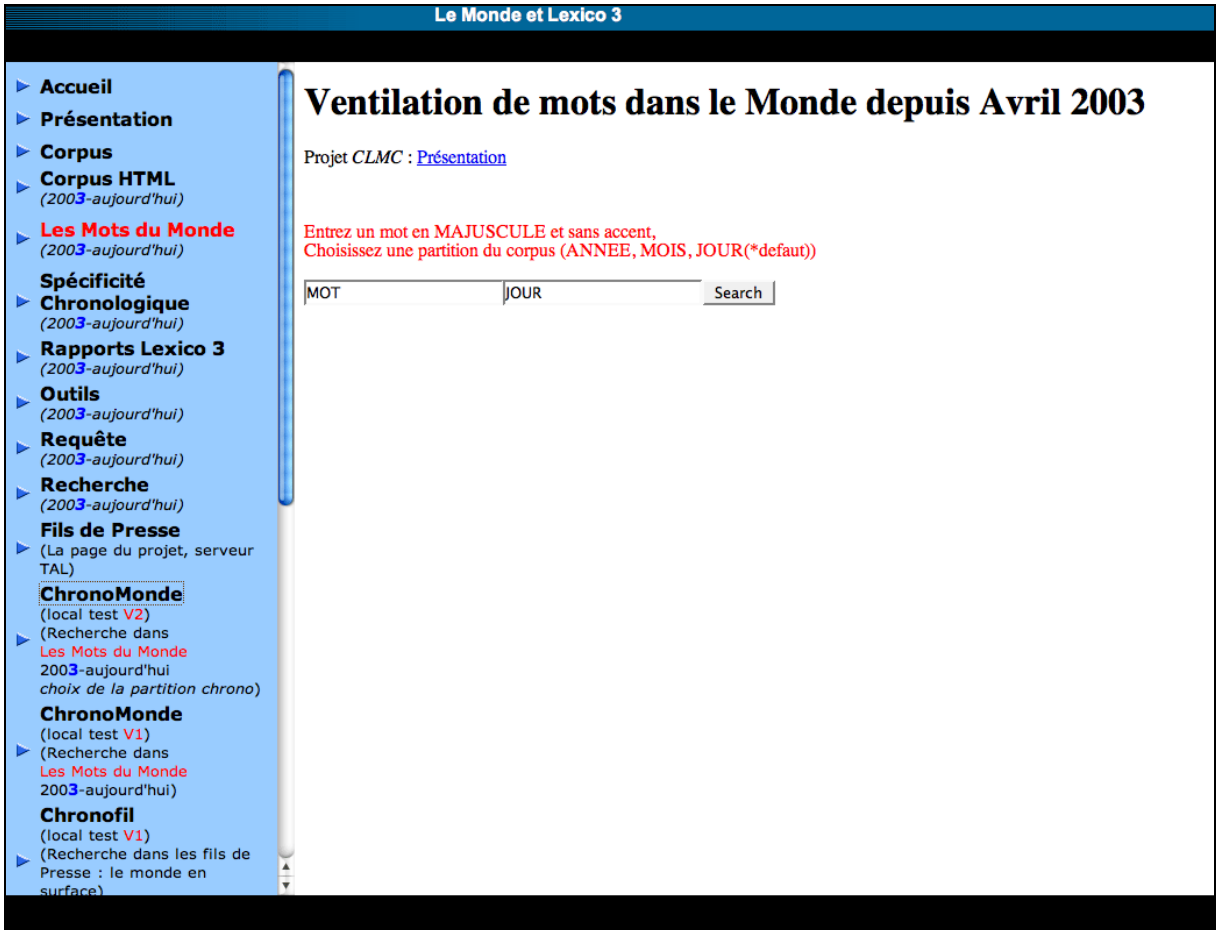


Figure 18 : Interface Chronomonde

Une fois entrée une forme graphique et après avoir choisi une partition pour la ventilation (JOUR, MOIS, ANNEE), le programme déclenché scrute le corpus et construit *in fine* le graphique de ventilation correspondant à la forme. La figure qui suit montre la ventilation de la forme LAICITE sur une partie du corpus (Avril 2003 – Novembre 2004), partition MOIS :

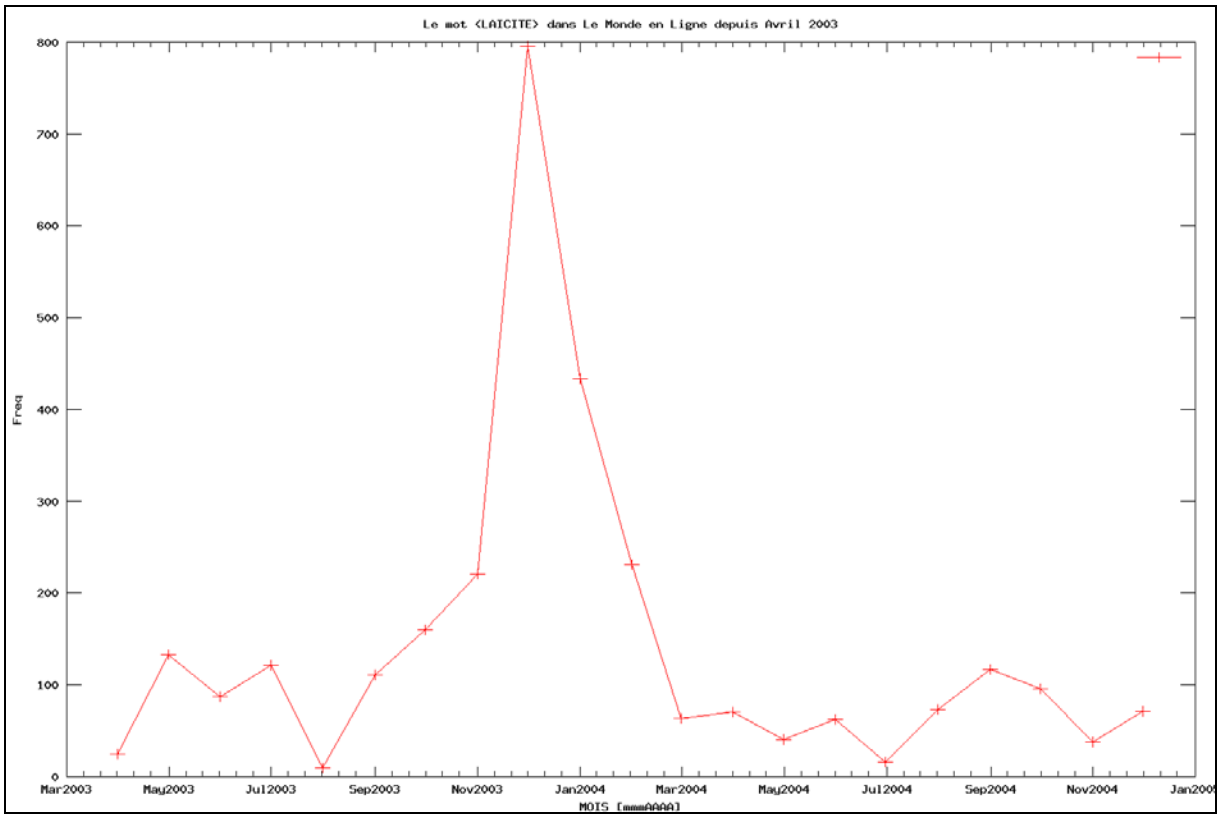


Figure 19 : sortie Chronomonde

6.2.2 Moteur de recherche

Un moteur de recherche permet de lancer des requêtes dans la version HTML du corpus archivé. Ce moteur prend appui sur le moteur d'indexation swish-e[12] qui est paramétré pour indexer l'ensemble des pages HTML du corpus original.

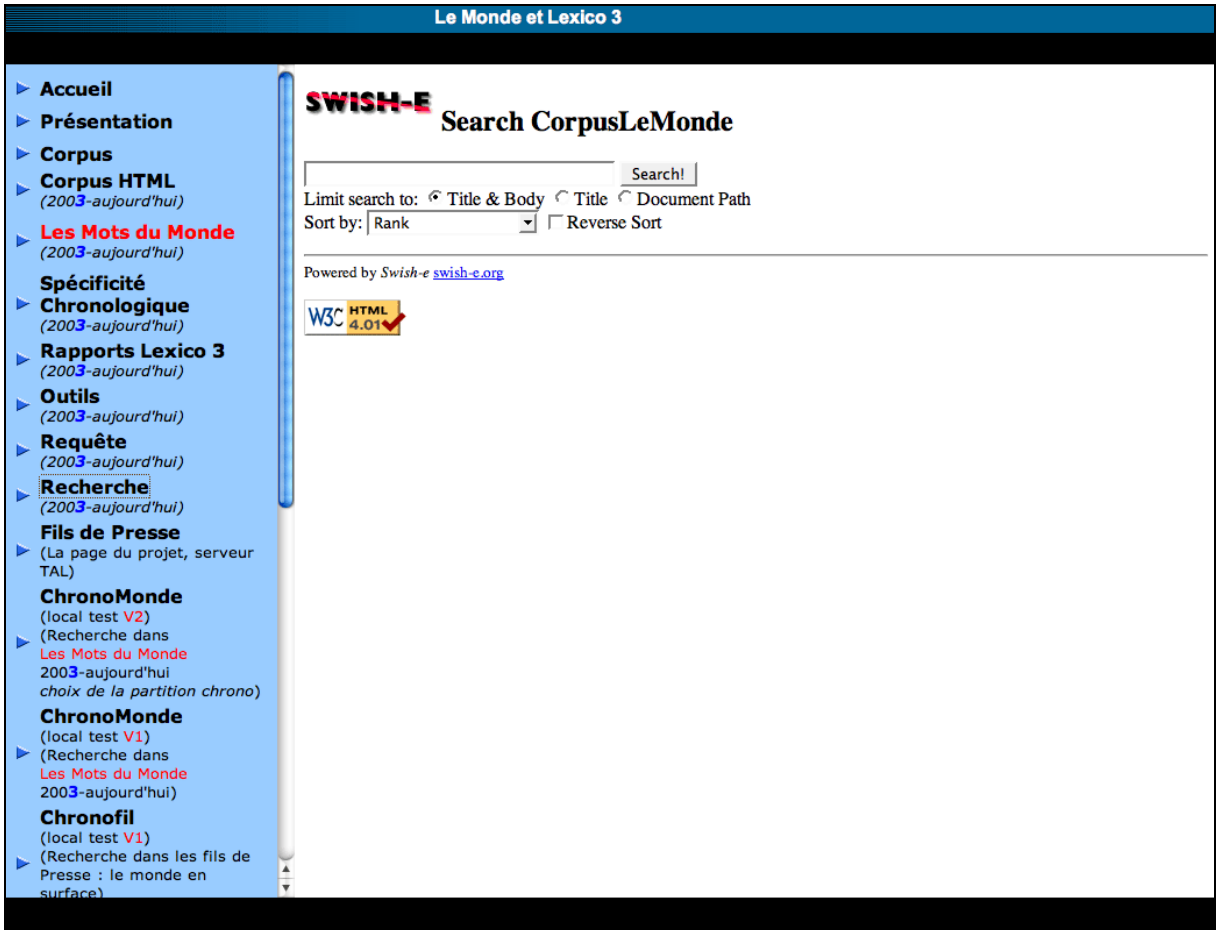


Figure 20 : Interface Recherche

[12] http://swish-e.org

6.3 Perspectives

6.3.1 Travaux en cours

Mise en place d'outils :

- pour structurer les données textuelles traitées
- les récupérer efficacement avant traitement
- et enfin pour les analyser

7 Parcours du site : présentation des données et outils disponibles

7.1 La page d'accueil du site

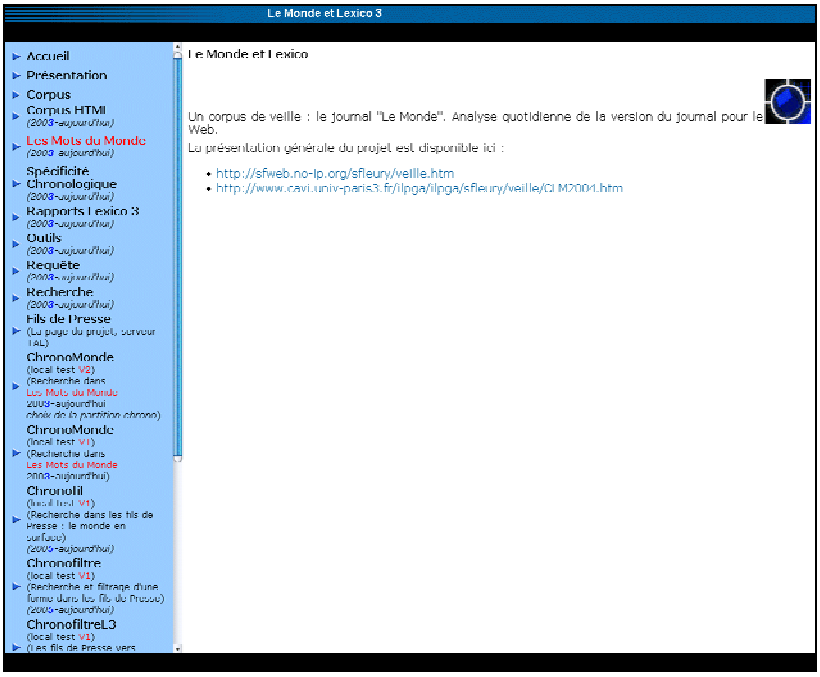


Figure 21 : page d'accueil du site

Le menu de gauche de cette page permet d'accéder aux rubriques présentées ci-dessous.

7.2 Présentation

Présentation générale du projet

7.3 Corpus

Présentation des différentes versions des corpus disponibles

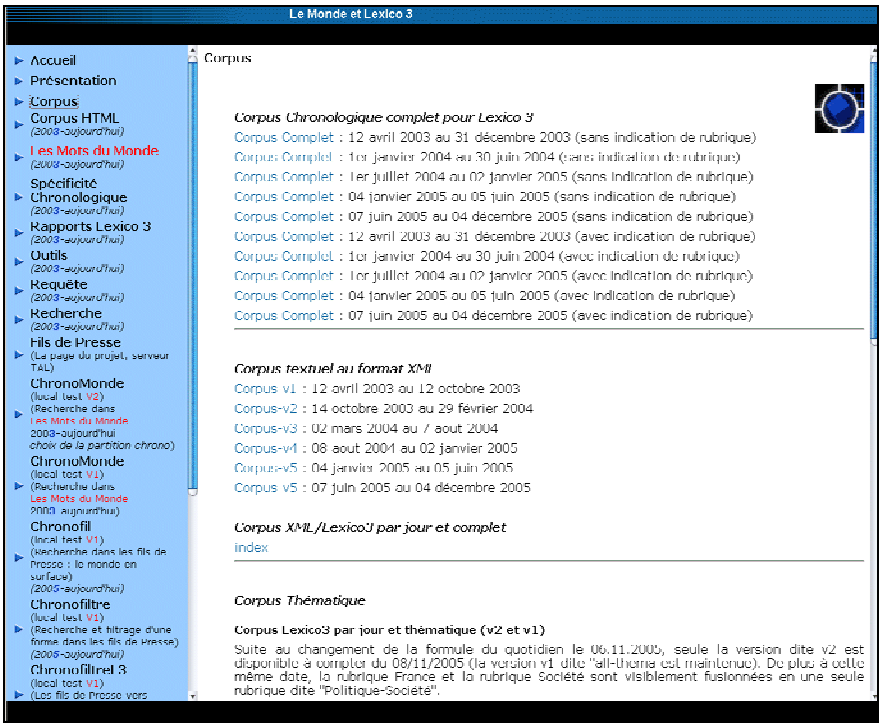


Figure 22 : page corpus

La chaîne de traitement présentée *infra* construit différentes versions des données associées aux versions électroniques du journal.

REMARQUE : Suite au changement de la formule du quotidien le 06.11.2005, seule la version dite v2 présentée *infra* est disponible à compter du 08/11/2005 (la version v1 dite « all-thema » est maintenue). De plus à cette même date, la rubrique France et la rubrique Société sont visiblement fusionnées en une seule rubrique dite "Politique-Société".

7.3.1 Corpus Chronologique pour *Lexico3*

Un corpus complet préparé pour être analysable par ***Lexico3*** est construit par la chaîne de traitement : il rassemble l'intégralité des états électroniques du journal depuis le démarrage de ce projet. Ce corpus est partitionné de la manière suivante : ANNEE, MOIS, SEMAINE, DATE, RUBRIQUE, PAGE (article). En raison de la taille volumineuse de ce corpus, celui-ci est disponible sous la forme de plusieurs fichiers (un fichier couvre une période de 6 mois environ) dont la concaténation construit une version couvrant l'ensemble de la période disponible.

7.3.2 Corpus textuel au format XML

Un corpus complet au format XML et regroupant les contenus textuels est également construit à l'issue de la chaîne de traitement. Ce corpus est utilisé dans le moteur de requête présenté *infra*. Le schéma ci-dessous décrit la structure de ce corpus.

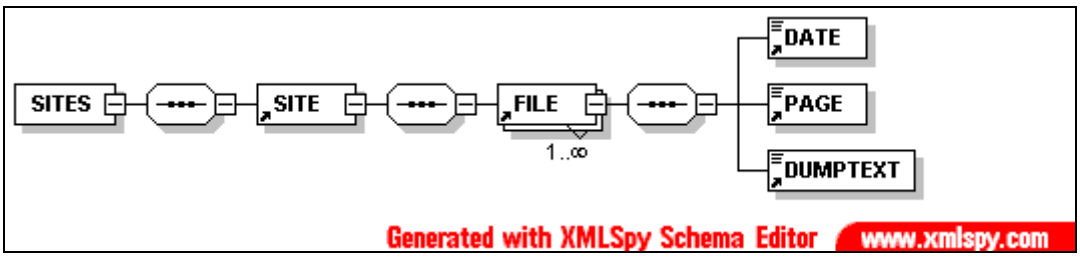


Figure 23 : schéma corpus

7.3.3 Corpus XML/*Lexico3* par jour et thématique (France, International, Société…)

La chaîne de traitement construit des "corpus quotidiens" au format XML et au format ***Lexico3***.

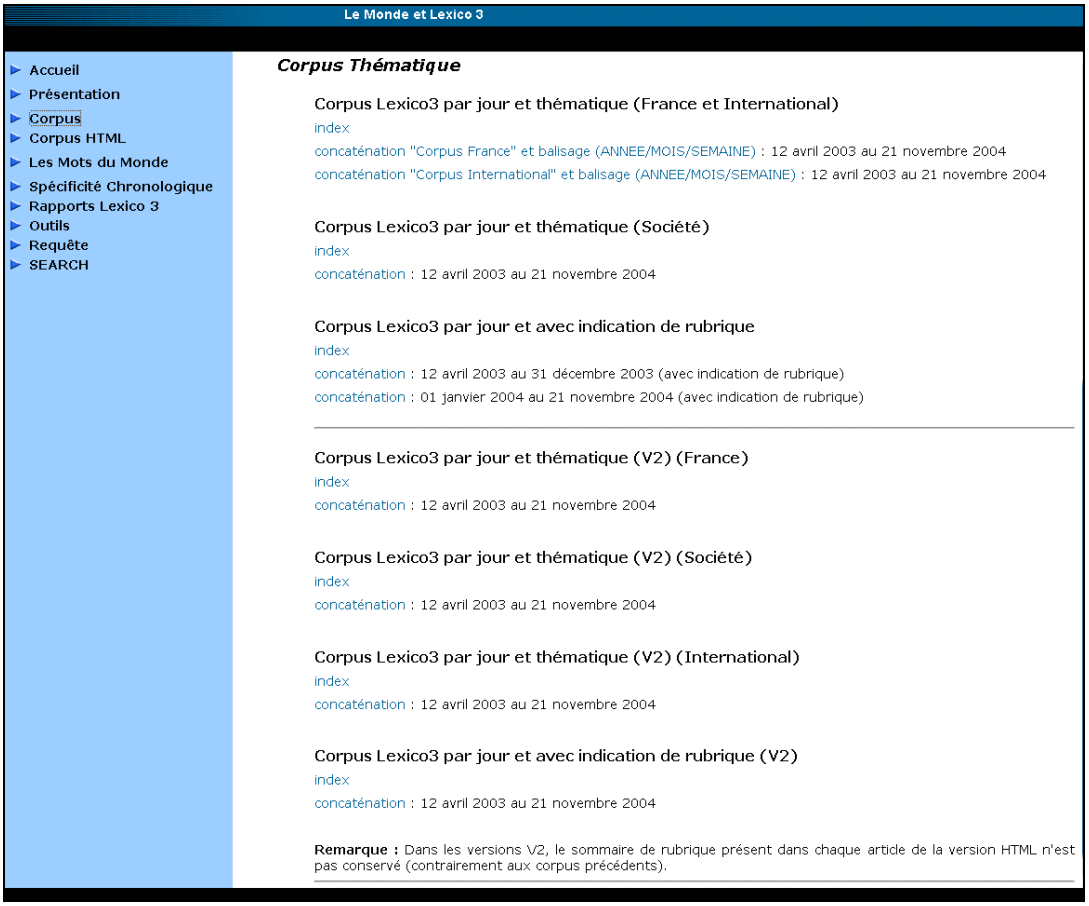


Figure 24 : corpus format Lexico

7.3.4 Corpus thématique complet pour *Lexico3*

Des corpus supplémentaires sont construits sur le même principe que le corpus chronologique complet mais en ne conservant que des rubriques sélectionnées : FRANCE (regroupant les rubriques « France » et « France-Société » du journal), INTERNATIONAL (la rubrique « international » du journal), SOCIETE (la rubrique « société » du journal). Il est possible ne pas se restreindre à ces rubriques particulières et de construire un corpus thématique donné : un générateur de corpus thématique (*i.e.* lié à une rubrique donné) est disponible dans la chaîne de traitement mise en œuvre (*cf infra*).

7.3.5 Corpus thématique quotidien pour *Lexico3*

La chaîne de traitement construit des "corpus quotidiens" au format ***Lexico3*** qui contrairement aux précédents intègre une indication de rubrique. Les articles de navigation (menu et sommaire de la version HTML) ne sont pas conservés dans cette version de données.

7.3.6 Corpus thématique complet pour *Lexico3* avec indication de rubrique

Ce corpus résulte de la concaténation des précédentes données.

7.3.7 Corpus thématique quotidien pour *Lexico3* (version dite V2)

Cette version du corpus se distingue de la précédente de la manière suivante :

Dans la version précédente, le processus de filtrage mis en place des parties textuelles se fait en utilisant la commande Unix `"lynx -dump"` sur les pages HTML originales, or ces pages contiennent systématiquement un sommaire de chaque rubrique pour une journée donnée. On avait donc, sur chaque version "nettoyée" une "surcharge" textuelle correspondant au rappel de ces rubriques.

Dans cette nouvelle version, le processus de filtrage textuel se fait toujours avec la même commande Unix mais un filtrage préalable des zones textuelles est opéré : on commence par isoler les contenus textuels correspondant au nom de la *Rubrique*, du *Titre* de l'article et de son *Contenu* "propre". Les pages HTML originales sont assez bien structurées et permettent d'isoler ces zones et a fortiori de négliger le sommaire qu'elles contiennent aussi.

7.4 Version HTML

La version originale (construite par les éditeurs de la version électronique du journal) des états quotidiens est disponible dans sa version HTML (simplifiée). Cette version permet de reconstruire « à la volée » des corpus thématiques en utilisant les outils construits à cet effet (*cf infra*).

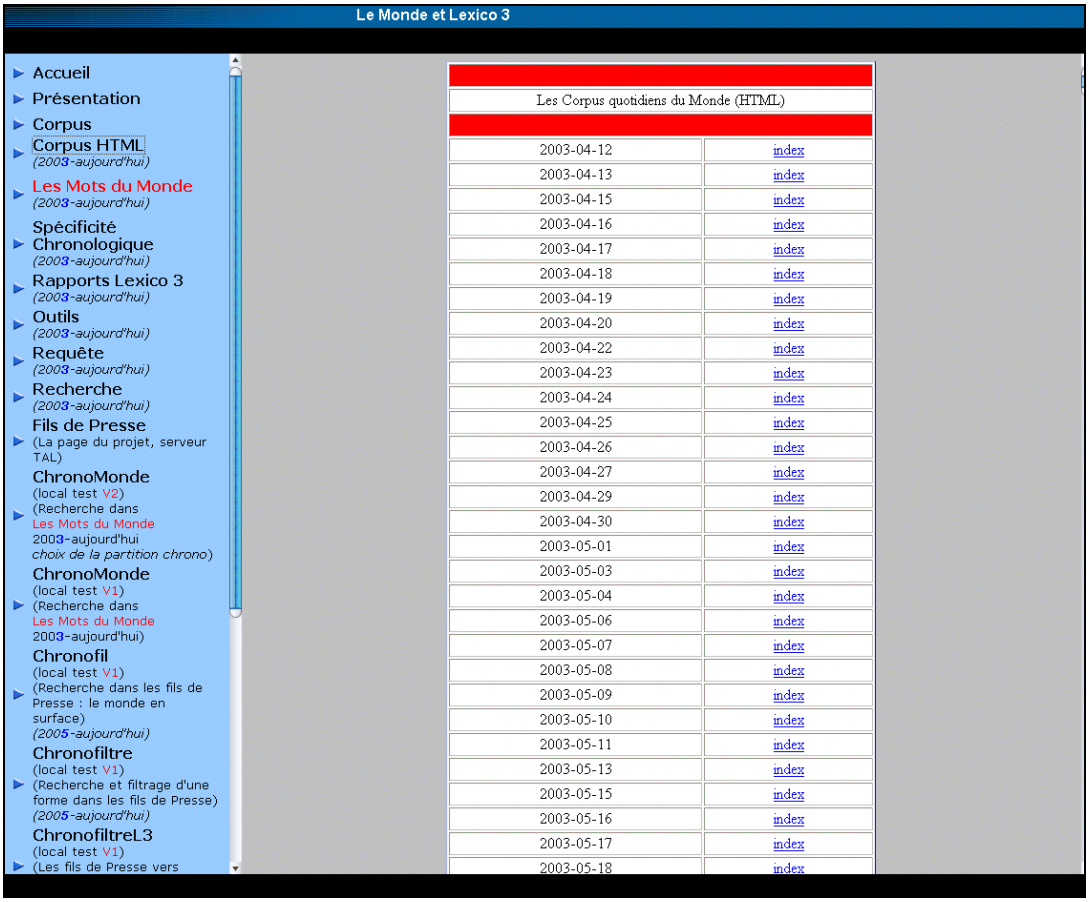


Figure 25 : page "pages HTML"

7.5 Les Mots du Monde

Dictionnaire des mots du journal et leur fréquence pour chaque jour (format HTML et XML)

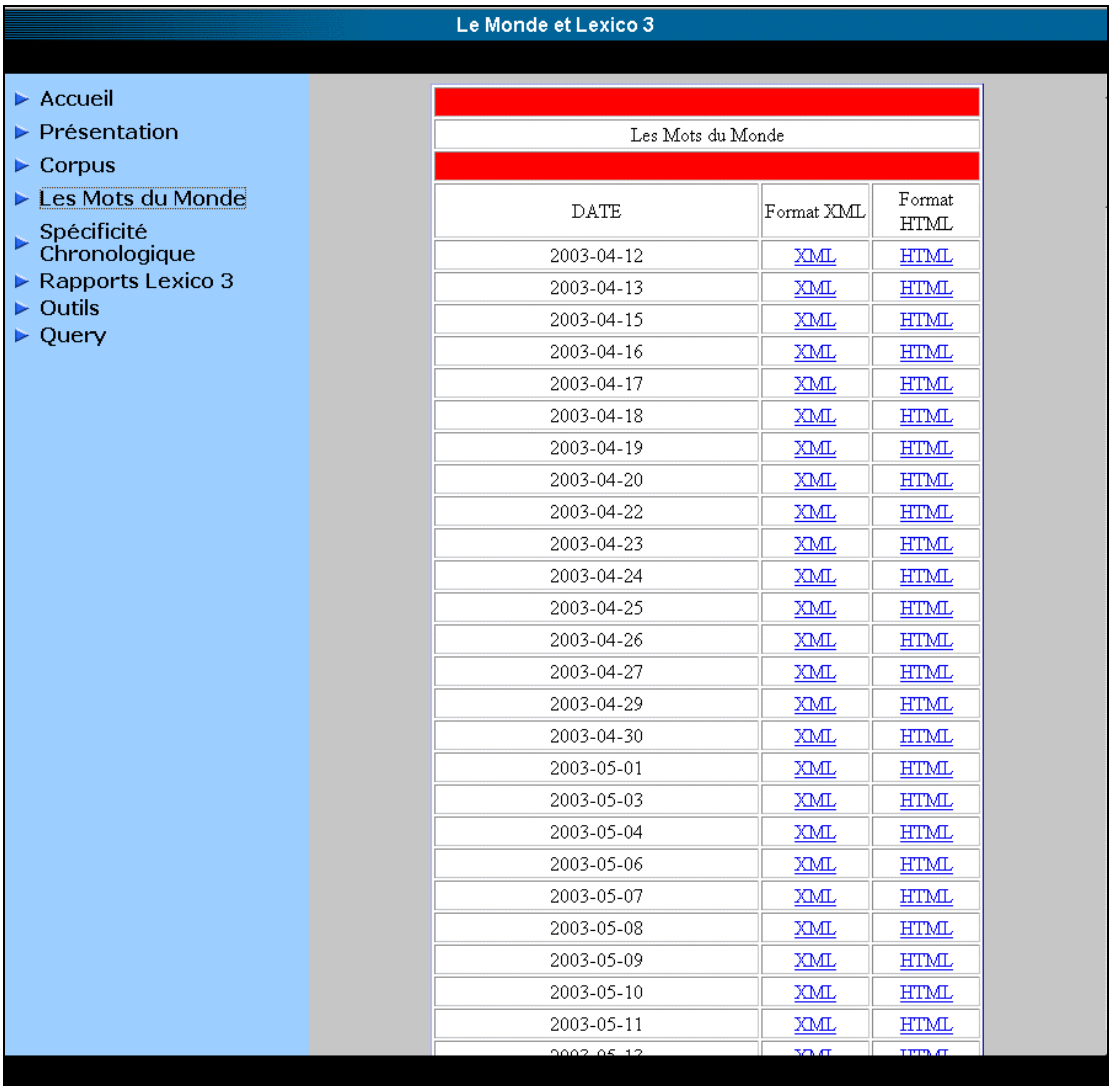


Figure 26 : page "les mots du monde"

La chaîne de traitements construit pour chaque journée traitée un index des mots utilisés et leur fréquence.

7.6 Spécificités chronologiques

L'ensemble des calculs des spécificités chronologiques construites avec **Lexico3** pour certaines dates.

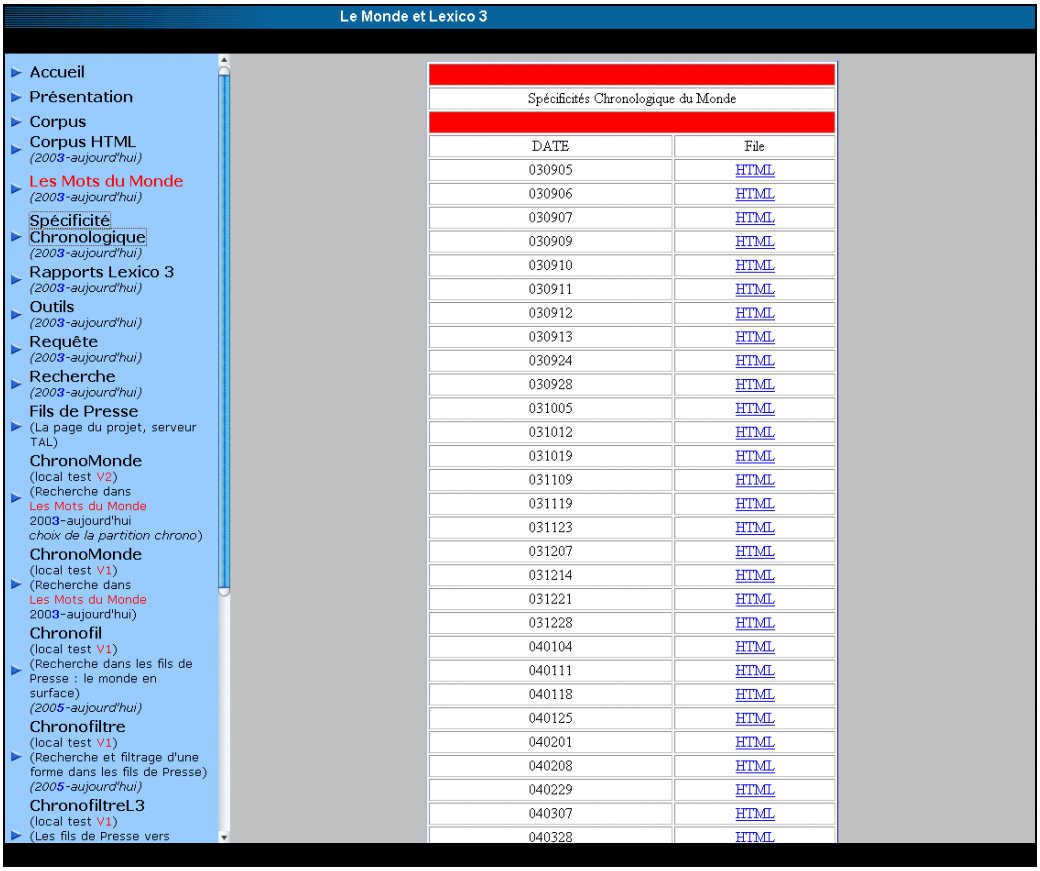


Figure 27 : page "spécificités chronologiques"

Des calculs de spécificités (par jour) sont construits par ***Lexico3*** pour certaines journées, pour le moment cette opération est réalisée manuellement.

7.7 Rapports *Lexico3*

Ensemble des rapports d'analyse construits avec **Lexico3**

Des rapports d'analyse réalisés avec ***Lexico3*** sont construits à partir des différentes versions disponibles du corpus. A ce jour une vingtaine de rapport ont été construits (*cf infra*).

7.8 Outils

Outils associés aux projets[13]

Cette page donnent des liens vers les outils utilisés ou construits pour ce projet.

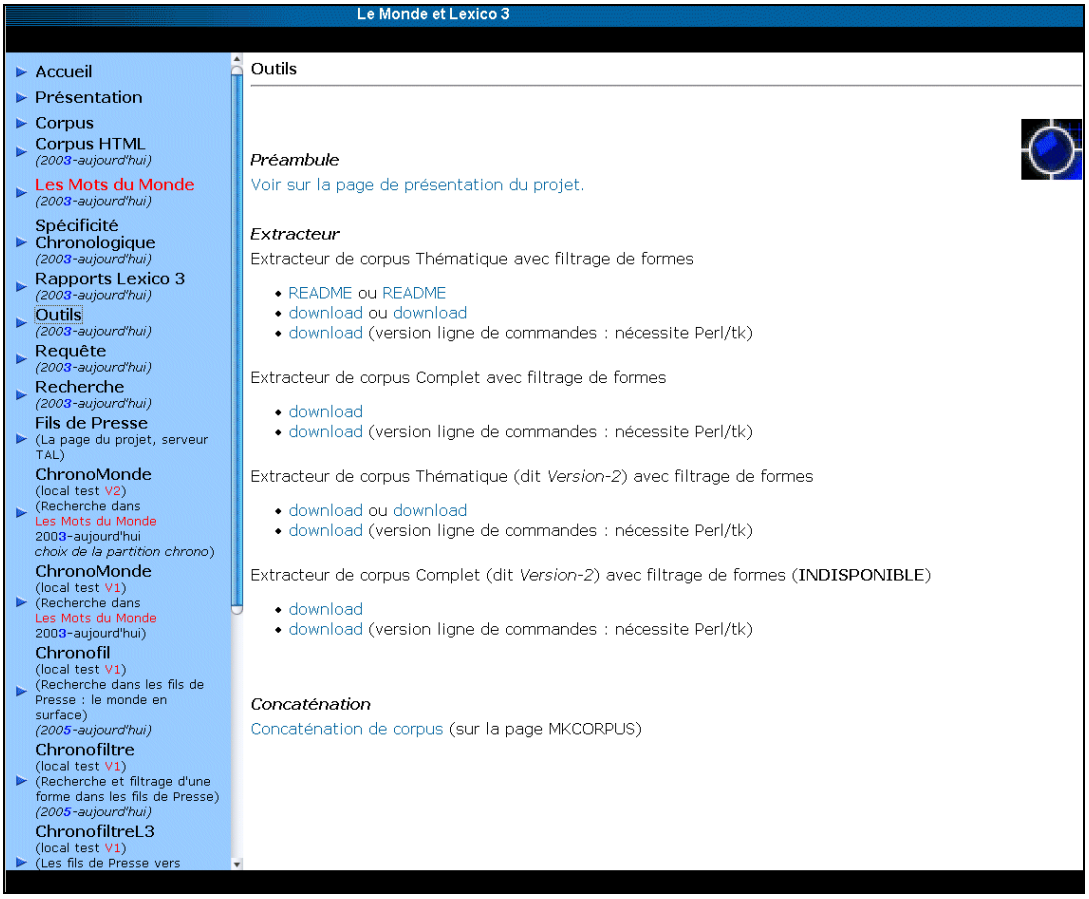


Figure 28 : page Outils

7.8.1 Générateur de corpus Thématique

Un script Perl peut construire à la demande un état complet regroupant l'ensemble des articles d'une même rubrique du journal sur l'ensemble de la période disponible. On dispose à ce jour de tels états pour les rubriques : France, International et Société.

Ce script utilise l'application Lynx.

"*Lynx is a fully-featured* World Wide Web *(*WWW*) browser for users on Unix, VMS, and other platforms running cursor-addressable, character-cell terminals or emulators. That includes vt100 terminals, other character-cell displays, and vt100 emulators such as Kermit or Procomm running on PCs or Macs. Lynx was a product of the Distributed Computing Group within Academic Computing Services of The University of Kansas. Lynx was originally developed by Lou Montulli, Michael Grobe, and Charles Rezac. Garrett Blythe created DosLynx and later joined the Lynx effort as well. Following the departures of Lou and Garrett for positions at Netscape in the summer of 1994,* Craig Lavender *provided support services for Lynx, and* Ravikumar Kolli *for DosLynx. Currently Lynx is being maintained and supported by members of the Internet community coordinated via the*

[13] Ces outils sont disponibles sur la page du projet http://sfmac.no-ip.com

lynx-dev mailing list. *Lynx is copyrighted by the University of Kansas and is distributed without restrictions on usage or redistribution under the GNU General Public License.*"

Mode d'emploi

```
perl MKCLM-TH.pl -at repOut repIn date thema thema2
```

Le script doit être lancé avec 4 paramètres et une option complémentaire :

L'option `-at` permet de re-diriger les sorties vers le répertoire `repOut`

Les paramètres à utiliser sont :

`repIn`	le répertoire contenant un état quotidien du journal au format HTML
`date`	la date traitée
`thema`	la rubrique à conserver
`thema2`	réécriture du nom de la rubrique après nettoyage typographique

Ce script utilise l'application Lynx (*cf supra*).

7.8.2 Générateur de corpus Thématique avec filtrage de formes

Un script Perl permet de construire un état complet regroupant l'ensemble des articles d'une même rubrique et contenant un motif textuel donné (une expression régulière[14] permettant de sélectionner un groupe de mots par exemple).

Ce script utilise l'application Lynx (*cf supra*).

Mode d'emploi

```
perl MKCLM-TH-querystring.pl -at  repOut repIn date thema thema2 queryString
```

Le script doit être lancé avec 5 paramètres et une option complémentaire :

L'option `-at` permet de re-diriger les sorties vers le répertoire `repOut`

Les paramètres à utiliser sont :

`repIn`	le répertoire contenant un état quotidien du journal au format HTML
`date`	la date traitée
`Thema`	la rubrique à conserver
`Thema2`	réécriture du nom de la rubrique après nettoyage typographique
`queryString`	le motif à rechercher dans les articles à conserver

7.8.3 Extracteur de corpus Thématique avec filtrage de formes

Un script Perl/Tk permet d'extraire une sous-partie d'un corpus thématique (construit par les précédents scripts) en ne conservant que les articles contenant un motif donné (exprimé là encore par une expression régulière permettant de sélectionner une famille de mots).

[14] Voir en annexe pour une présentation rapide des expressions régulières

Exemple d'utilisation :

A partir du corpus Thématique France, extraction des articles contenant les formes graphiques : sécurité, sécuritaire, etc.

Ci-dessous l'interface de cet extracteur :

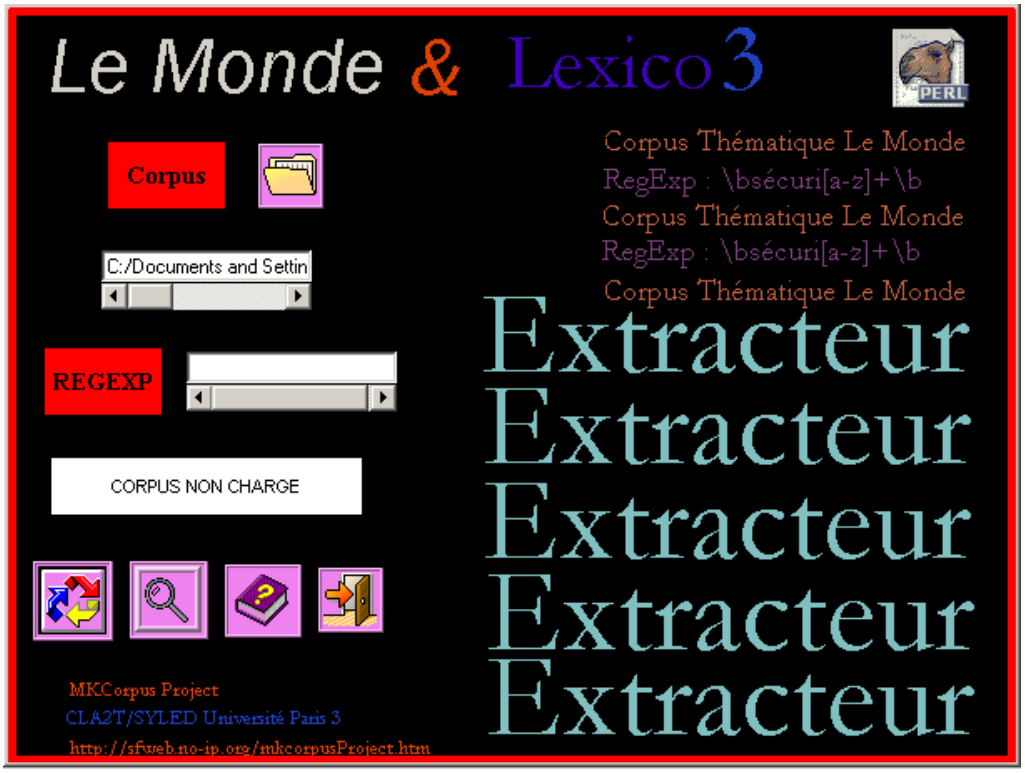


Figure 29 : MKCorpusLeMonde-extracteur

MKCorpusLeMonde-Extracteur :

Ce programme prend en entrée un corpus thématique du journal *Le Monde* et un motif (une expression régulière) et il extrait tous les articles contenant le motif : le format en sortie est compatible avec ***Lexico3***.

Mode d'emploi :

Ce programme existe sous 2 formes :

(1) Un script Perl/Tk qui doit être lancé dans une fenêtre de commandes en respectant la syntaxe suivante :

```
perl MKCorpusLeMonde-Extracteur -at  repOut corpus queryString
```

Le script doit être lancé avec 2 paramètres et une option complémentaire :

L'option `-at` permet de re-diriger les sorties vers le répertoire `repOut`

Les paramètres à utiliser sont :

`Corpus`	Le nom du corpus thématique utilisé
`queryString`	le motif à rechercher dans les articles à conserver

(2) Une application Perl/Tk (un exécutable) dans un environnement Windows. Pour le lancer, double-clic sur le programme.

INPUT :

Le corpus thématique (*i.e.* correspondant à une rubrique donnée) est issu de la chaîne de traitement présentée dans ce document. Pour charger ce corpus, utiliser le bouton `"browse"` puis sélectionner le corpus thématique souhaité. Le corpus en entrée contient un balisage propre à son utilisation avec ***Lexico3***. Ce balisage tient compte essentiellement de l'aspect chronologique du corpus. On y trouve donc en général des balises du type ANNEE, MOIS, SEMAINE, DATE, PAGE (cette dernière balise correspondant en fait aux articles). Les valeurs des balises précédentes correspondent aux périodes temporelles associées aux parties du corpus visées. Ce balisage est maintenu par l'extracteur : le corpus en sortie aura une structure similaire.

FILTRE :

Les filtres permettent de sélectionner des articles (dans le corpus thématique) contenant une chaîne de caractère correspondant au filtre écrit. Un filtre est écrit sous la forme d'une expression régulière.

Exemple : \bsécurit[a-z]+\b

Ce motif permet de rechercher tous les mots qui commencent (\b) par la chaîne de caractères 'sécurit' (peu importe la casse) suivie d'une répétition de un à « une infinité » de n'importe quel caractère compris entre a et z avant la fin du mot (\b). Les mots couverts par ce motif sont par exemple : sécurité, sécurités, sécuritaire etc.

OUTPUT :

Le corpus obtenu à l'issue de l'extraction contient tous les articles du corpus initial contenant le motif décrit par le filtre. Ce corpus en sortie est compatible avec le format de fichier pour ***Lexico3***. De plus, il est possible de relancer le processus d'extraction sur ce fichier pour par exemple affiner le processus de filtrage des articles. Les corpus issus du processus d'extraction sont horodatés, il est conseillé cependant de leur donner un nom plus précis.

7.8.4 Extracteur de corpus complet avec filtrage de formes

Un script Perl/Tk permet d'extraire une sous-partie du corpus complet en ne conservant que les articles contenant un motif donné (exprimé là encore par une expression régulière permettant de sélectionner une famille de mots).

Même principe que le précédent.

Nom du programme : MKCorpusLeMonde-Extracteur-complet.

7.8.5 Générateur de corpus Thématique (dit version v2)

Même principe que ci-dessus.

7.8.6 Générateur de corpus Thématique (dit version v2) avec filtrage de formes

En cours de développement

7.8.7 Extracteur de corpus Thématique (dit version v2) avec filtrage de formes

Même principe que ci-dessus.

7.8.8 Extracteur de corpus complet (dit version v2) avec filtrage de formes

En cours de développement

7.9 Query

Un moteur de requête qui permet pour le moment de lancer des requête sur un des états du corpus

Cette page donne accès à un moteur de requête sur la version XML du corpus. Le détail des outils mis en place dans ce moteur est détaillé *infra*. Ce moteur permet de lancer 2 types de requêtes :

- par date : on obtient en sortie, le contenu textuel du journal relatif à la date choisie. Dans la figure qui suit, la requête sur la date 2004-01-09 permet de recueillir l'intégralité des contenus textuels de cette journée.

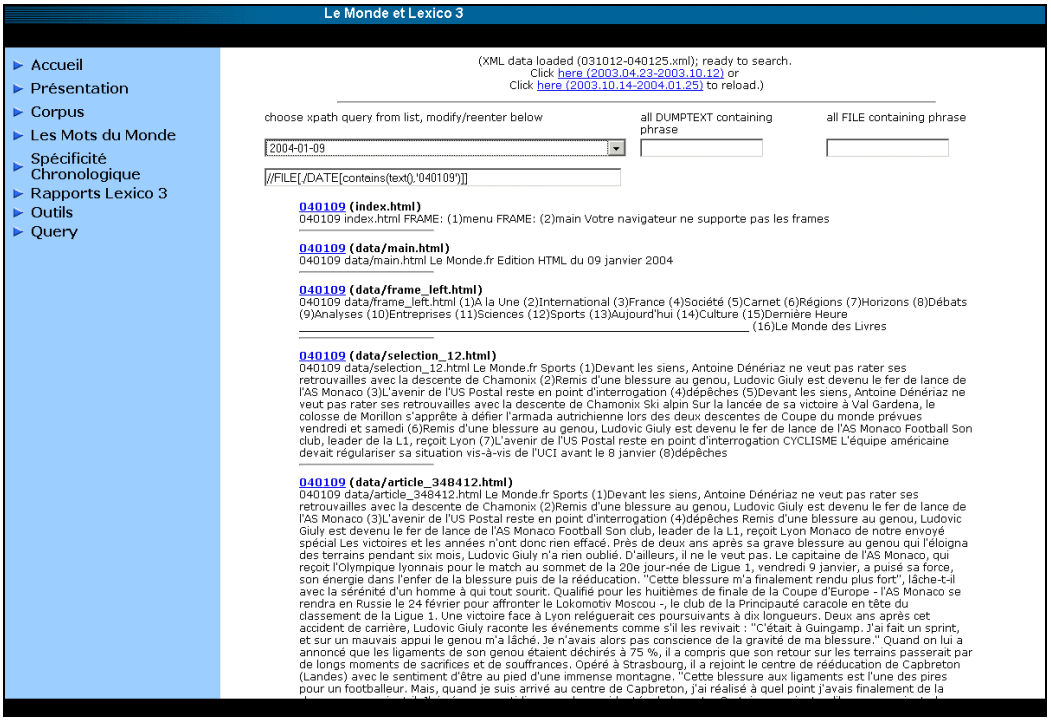


Figure 30 : page Query

- on peut aussi lancer une requête pour extraire les contenus textuels contenant une séquence textuelle choisie par l'utilisateur. Dans la figure qui suit, la requête lancée vise à recueillir tous les articles qui contiennent le mot "voile".

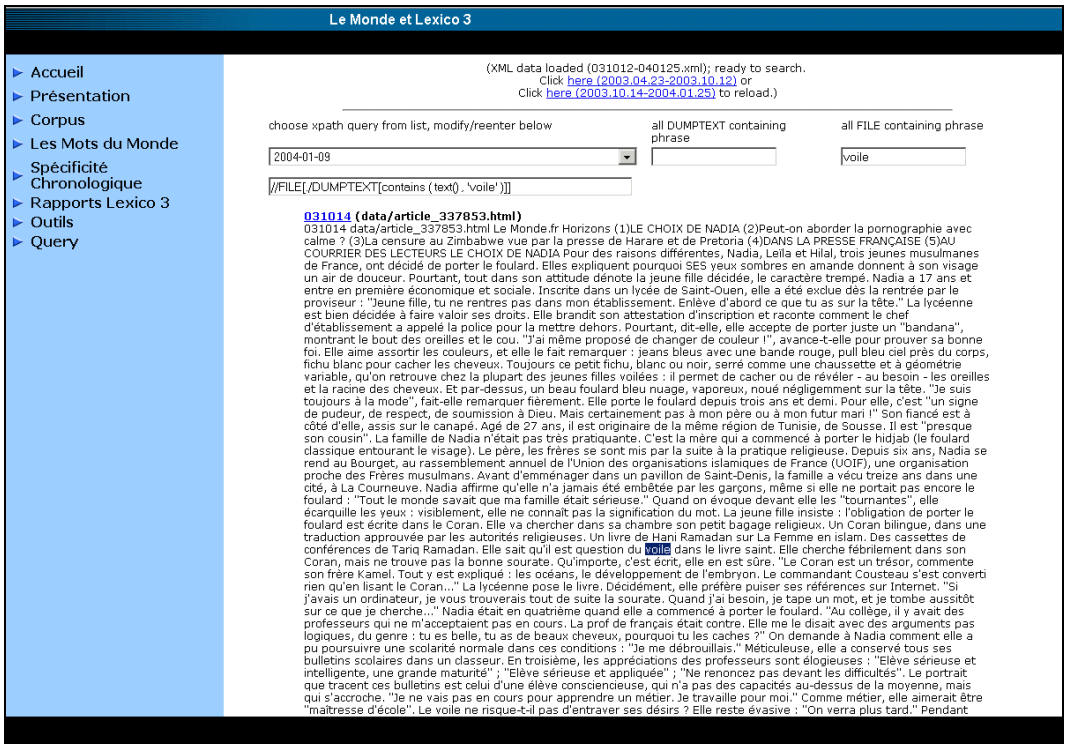


Figure 31 : page Query (2)

Les données utilisées à ce jour dans ce moteur sont constituées de 2 fichiers au format XML, le premier couvre la période suivante : du 12 avril 2003 au 10 décembre 2003 et du 14 décembre 2003 au 25 janvier 2004. Ces deux fichiers font respectivement 90 Mo et 55 Mo.

Des tests ont été opérés pour mesurer les temps de traitement nécessaires pour travailler sur ces données volumineuses dans le contexte web mis en place actuellement :

dans une utilisation sur un serveur local, les temps de réponse sont tout à fait satisfaisants

dans une utilisation externe, il faut compter une heure pour charger les données (1 des 2 fichiers), l'exécution des requêtes est ensuite assez rapide.

7.10 Recherche

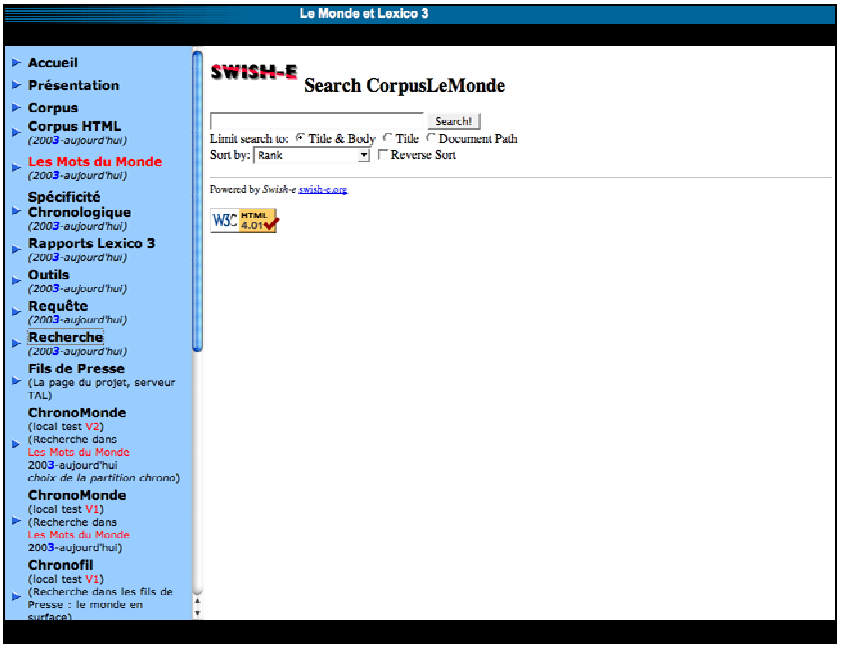


Figure 32 : page Recherche

Cet item donne accès à une page de formulaire permettant de lancer des requêtes (moteur de recherche) sur le corpus original "HTML simplifié et sans image". L'image qui suit montre le résultat d'une requête sur la recherche du mot "laïcité".

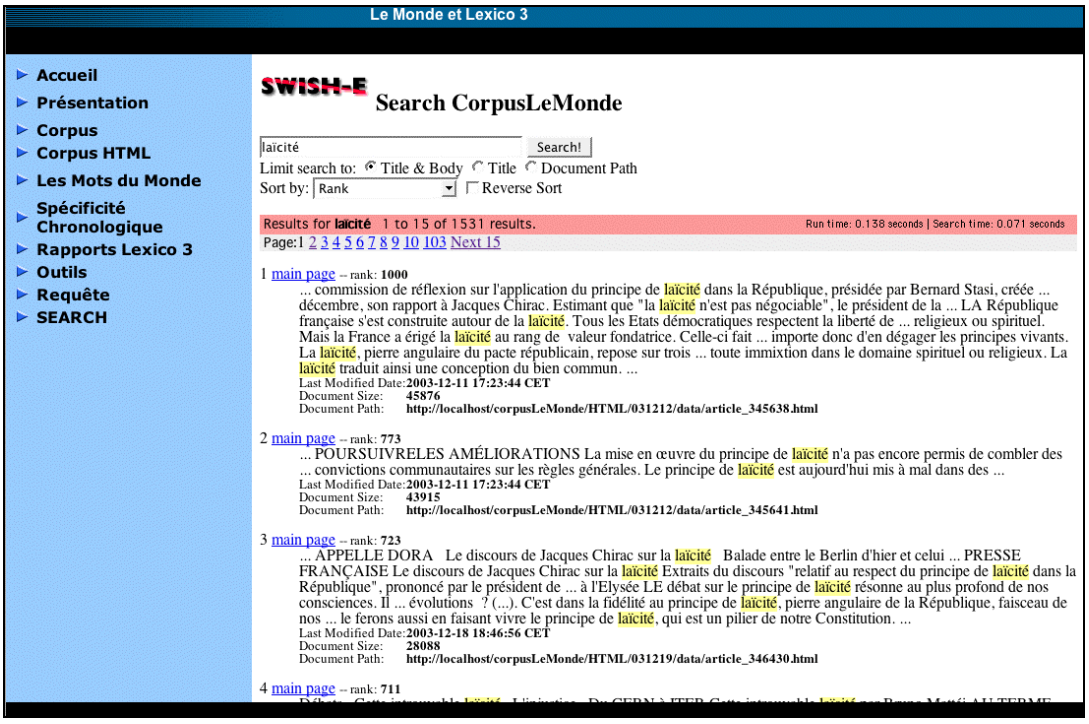


Figure 33 : page Recherche (2)

7.11 Le moteur de requête

Le moteur de requête en cours de développement reprend une application développée par John Udell[15] permettant de sélectionner et de récupérer les rubriques de son weblog classées thématiquement via l'utilisation d'une interface web utilisant des requêtes XPath[16].

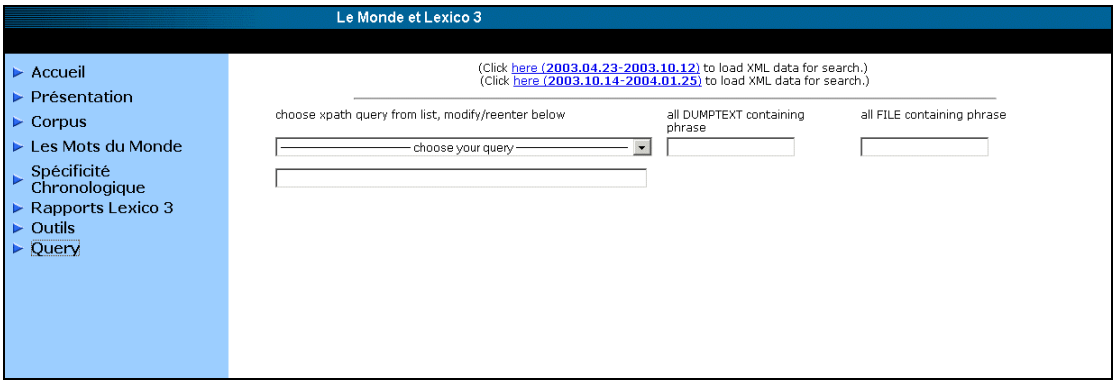


Figure 34 : le moteur de requête (1/3)

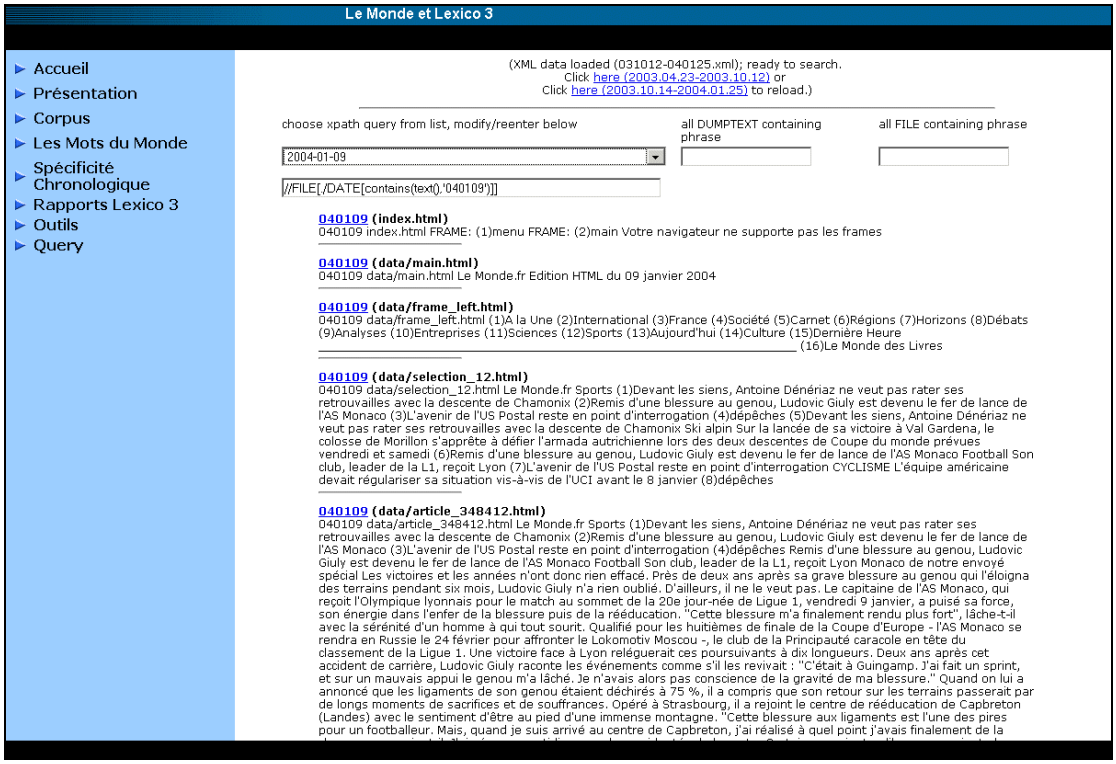


Figure 35 : le moteur de requête (2/3)

[15] http://udell.roninhouse.com

[16] http://udell.infoworld.com:8000/?//blocquote[@cite='infoWorld']

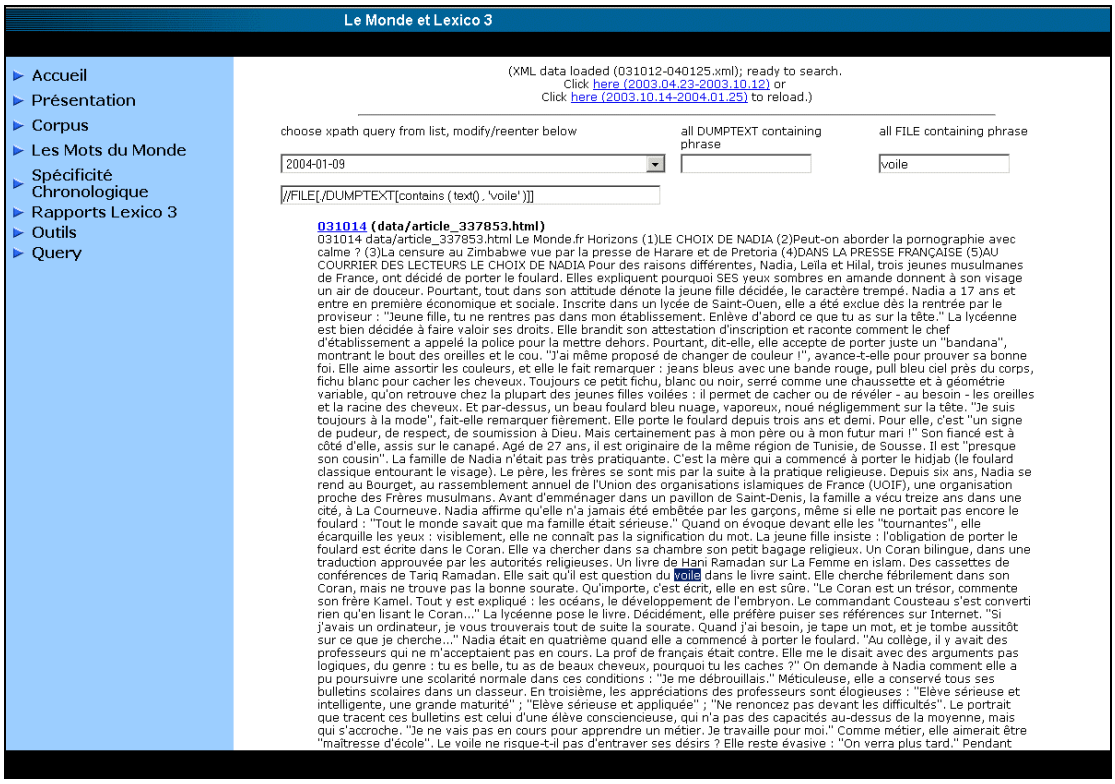


Figure 36 : le moteur de requête (3/3)

7.12 Boîte à outils CGI

7.12.1 Chronofil

Cf supra

Figure 37 : Chronofil

7.12.2 Chronofiltre

Cf supra

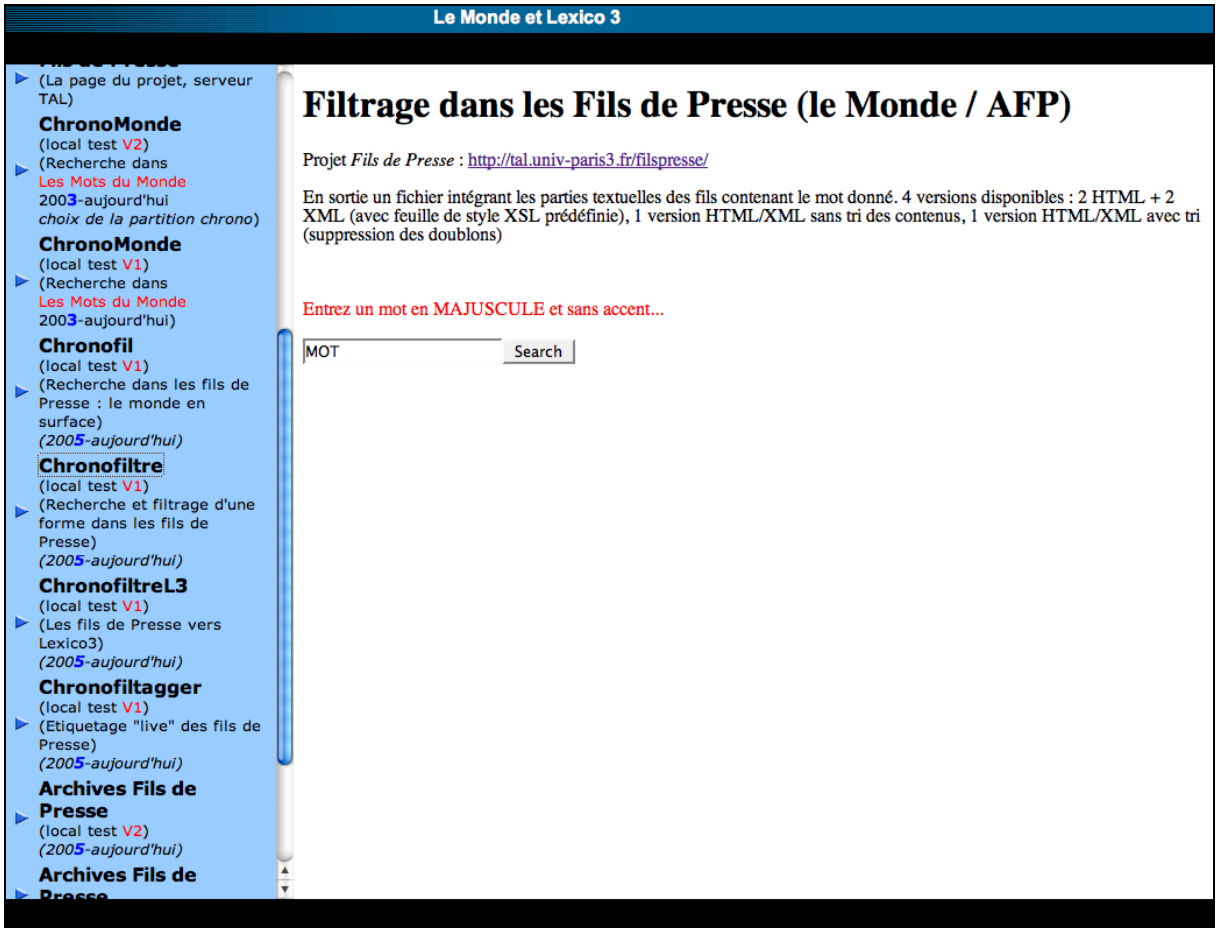


Figure 38 : Chronofiltre

7.12.3 Chronomonde

Cf supra

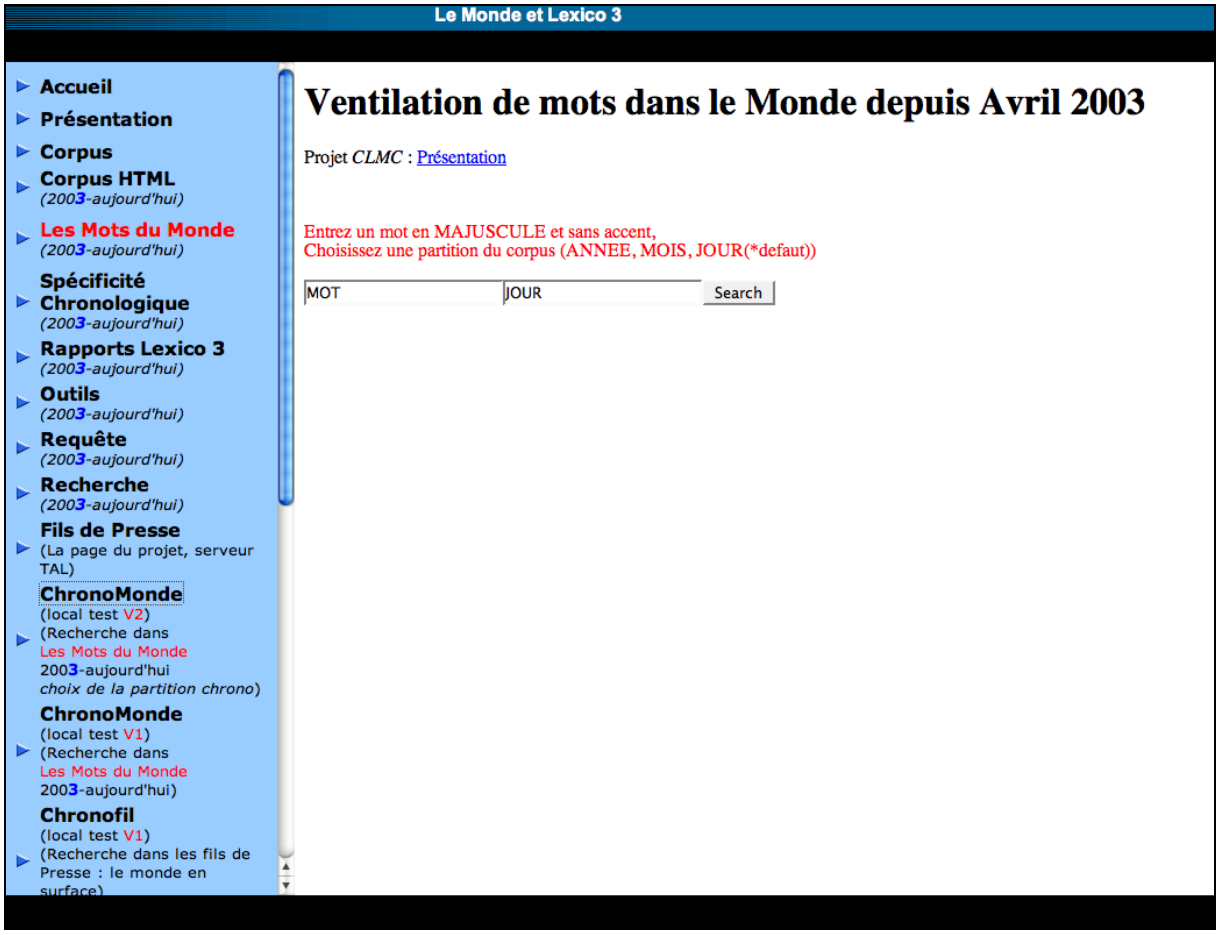


Figure 39 : Chronomonde

7.12.4 ChronofiltreL3

Cf supra

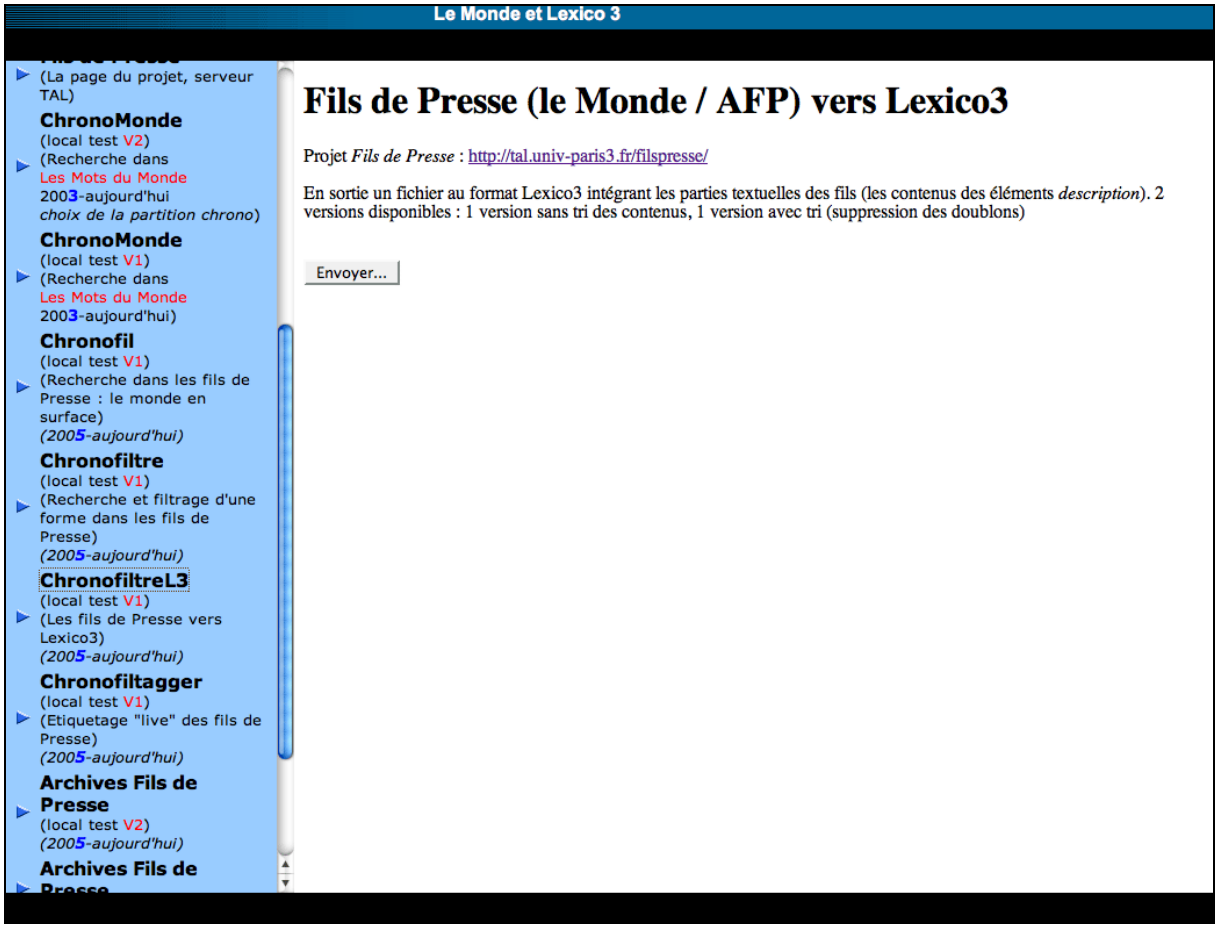


Figure 40 : ChronofiltreL3

7.12.5 Chronofiltagger

Cf supra

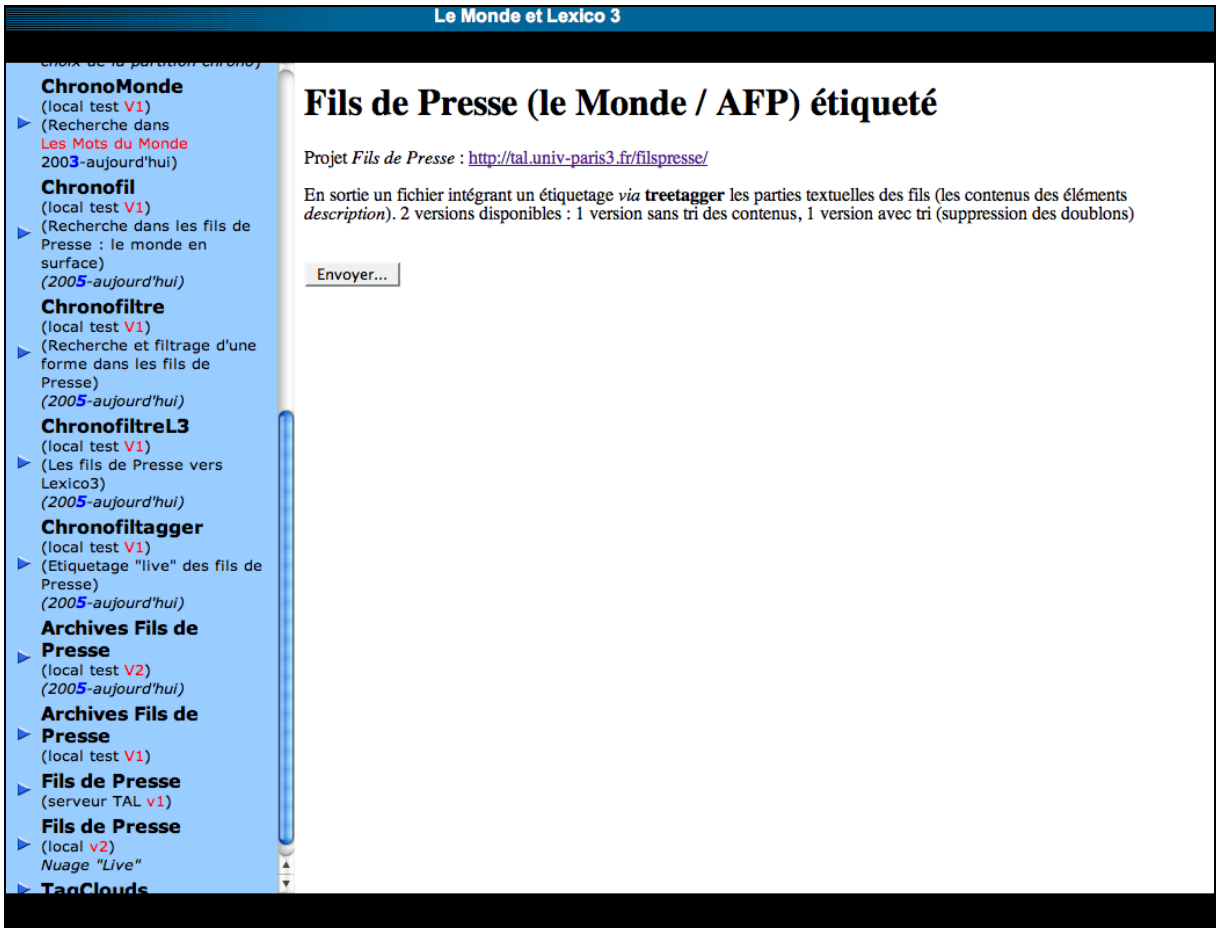


Figure 41 : Chronofiltagger

8 Références bibliographiques

2001, Valérie Beaudouin, Serge Fleury, Benoît Habert, Gabriel Illouz, Christian Licoppe, Marie Pasquier, "TyPWeb : décrire la Toile pour mieux comprendre les parcours", CIUST'01, Colloque International sur les Usages et les Services des Télécommunications, e-Usages, Paris, 12-14 juin (Version PDF)

2001, Cédric Lamalle, William Martinez, Serge Fleury, André Salem, Andrea Kuncova, Aude Maisondieu, "Dix premiers pas avec Lexico3", Manuel d'utilisation abrégé (Version PDF), (Version HTML) (sur le site de Lexico)

2002, Valérie Beaudouin, Serge Fleury, Marie Pasquier, Benoît Habert, Christian Licoppe, "TyPWeb : décrire la Toile pour mieux comprendre les parcours. Sites Personnels et sites marchands", in *RESEAUX*, Volume 20, n°116/2002, "Parcours Sur Internet", pages 19-52, FT&RD/Hermès

9 Annexes

9.1 Annexe Partie 1

9.1.1 Exemple de rapports construits avec *Lexico3*

On présente ici un rapport[17] construit sur le corpus complet avec indication de rubriques sur la période du 12 avril 2003 au 25 janvier 2004.

Cette première page donne une présentation générale de caractéristiques du corpus et un point d'entrée vers les résultats construits :

Principales caratéristiques lexicométriques	
Nombre des occurrences	20814181
Nombre des formes	236878
Fréquence maximale	1054349
Nombre des hapax	95427

Principales caractéristiques de la partition : DATE

Spécifs - Part : DATE Parties sélectionnées : "040125", (1)

Graphique de ventilation pour la partition : MOIS (1)

Graphique de ventilation pour la partition : MOIS (2)

Graphique de ventilation pour la partition : MOIS (3)

Graphique de ventilation pour la partition : MOIS (4)

Graphique de ventilation pour la partition : MOIS (5)

[17] http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/rapportsL3/lemonde/23/index.htm

9.1.1.1 Principales caractéristiques de la partition : DATE

Partie	Nb occurrences	Nb formes	Nb hapax	Fréq. Max	Forme
"030412"	112381	15126	8431	5810	de
"030413"	89082	13106	7320	4397	de
"030415"	99958	13044	7056	5078	de
"030416"	81067	12451	6982	4072	de
"030417"	72207	11741	6719	3498	de
"030418"	107672	15277	8499	5678	de
"030419"	99999	14715	8319	4614	de
"030420"	72675	11453	6375	3762	de
"030422"	54165	10057	6006	2669	de
"030423"	97590	13814	7679	4842	de
"030424"	74333	11761	6686	3814	de
"030425"	92763	14517	8247	4897	de
"030426"	93498	13941	7935	4708	de
"030427"	66391	11151	6411	3468	de
"030429"	100283	13112	7051	5100	de
"030430"	69765	11045	6198	3563	de
"030501"	98056	14705	8307	5004	de
"030503"	92861	13770	7733	4552	de
"030504"	80813	12136	6734	4125	de
"030506"	100027	13820	7759	5119	de
"030507"	81589	12653	7015	4283	de
"030508"	75886	12345	7054	3916	de
"030509"	88439	14165	8216	4527	de
"030510"	109317	14889	8280	5723	de
"030511"	68223	10795	6054	3485	de
"030513"	103051	13630	7437	5411	de
"030515"	101227	14182	8130	5182	de
"030516"	101761	14485	8217	5194	de
"030517"	99793	14218	7774	5065	de
"030518"	78494	12044	6733	4047	de
"030520"	96383	12977	7032	4772	de
"030521"	80838	12374	6924	4003	de
"030522"	81813	12312	6954	4098	de
"030523"	113852	15895	8935	5748	de
"030524"	107205	14783	8406	5270	de
"030525"	77704	11716	6650	3977	de
"030527"	104251	13356	7320	5162	de
"030528"	88341	12957	7299	4616	de
"030529"	76832	12359	7094	3858	de
"030530"	80979	13307	7654	4046	de
"030531"	109412	14862	8352	5275	de
"030603"	116495	14083	7613	5799	de
"030605"	80288	11430	6311	4140	de
"030606"	119069	16655	9402	5836	de
"030607"	105537	14628	8230	5221	de
"030608"	67059	10991	6322	3321	de
"030610"	54493	9684	5628	2741	de
"030611"	110572	14129	7646	5505	de
"030612"	78243	12733	7363	3916	de
"030613"	101523	15187	8584	4989	de
"030614"	101602	14337	8007	5158	de
"030615"	76090	11819	6672	3826	de
"030617"	98870	13461	7330	5026	de
"030618"	82917	12303	6882	4263	de
"030619"	84585	12517	6922	4348	de
"030620"	105921	15254	8488	5377	de
"030621"	108032	15285	8617	5542	de
"030622"	85385	12618	7061	4236	de
"030624"	106686	14365	7912	5554	de
"030625"	79273	12328	6932	4224	de
"030626"	75699	11918	6773	3965	de
"030627"	102711	15086	8589	4869	de
"030628"	110907	14859	8176	5382	de

"030629"	72509	11456	6435	3801	de
"030701"	108993	14438	7868	5790	de
"030702"	86314	12977	7283	4493	de
"030703"	76492	11884	6758	3812	de
"030704"	116897	16261	9159	5936	de
"030705"	109022	15175	8508	5668	de
"030706"	78467	11735	6567	4031	de
"030708"	101853	13748	7539	5003	de
"030709"	83935	12668	7049	4068	de
"030710"	78022	12204	6696	3795	de
"030711"	108573	15424	8605	5696	de
"030712"	104031	14816	8332	5269	de
"030713"	56727	10011	5759	2785	de
"030715"	57836	10087	5819	2886	de
"030716"	60139	10086	5728	2959	de
"030717"	55583	10211	6050	2720	de
"030718"	87642	13973	8094	4492	de
"030719"	72843	12209	7041	3436	de
"030720"	105483	15536	8693	5190	de
"030723"	49936	9409	5584	2515	de
"030724"	44405	8604	5091	2384	de
"030725"	90489	15164	8836	4566	de
"030726"	80010	12458	7197	4129	de
"030727"	54995	10293	6063	2581	de
"030729"	50450	9511	5610	2482	de
"030730"	55655	10362	6185	2952	de
"030731"	51209	10063	6046	2587	de
"030801"	53006	9770	5715	2721	de
"030802"	71026	12127	7053	3619	de
"030803"	48812	9864	5985	2378	de
"030805"	49721	9234	5385	2542	de
"030806"	52703	9712	5654	2710	de
"030807"	48581	9695	5690	2566	de
"030808"	50972	9849	5864	2597	de
"030809"	77070	13008	7653	3820	de
"030810"	46052	9588	5839	2309	de
"030812"	51118	9690	5750	2479	de
"030813"	44908	9105	5504	2261	de
"030814"	42737	8800	5314	2156	de
"030815"	47763	9039	5264	2524	de
"030816"	70527	12222	7318	3537	de
"030817"	49631	9431	5553	2401	de
"030819"	50177	9744	5871	2565	de
"030820"	45334	8767	5123	2480	de
"030821"	58495	9889	5680	2903	de
"030822"	85101	13249	7849	4045	de
"030823"	79258	12953	7671	3658	de
"030824"	53209	10138	6020	2536	de
"030826"	57359	10038	5701	2836	de
"030827"	57430	9831	5645	2907	de
"030828"	58130	9724	5372	2858	de
"030829"	89198	14100	8237	4283	de
"030830"	85353	12512	7146	4195	de
"030831"	52441	9857	5896	2601	de
"030902"	97194	12709	6842	4970	de
"030903"	74363	11459	6352	4064	de
"030904"	70755	11339	6439	3742	de
"030905"	106138	14515	8013	5599	de
"030906"	102418	13743	7568	5250	de
"030907"	74382	11059	6127	3826	de
"030909"	101656	12918	6911	5102	de
"030910"	77427	11875	6671	4056	de
"030911"	74630	12031	6873	4054	de
"030912"	162277	19607	10648	8018	de
"030913"	107828	14859	8273	5165	de
"030914"	90484	12868	7007	4723	de
"030916"	101760	13187	7133	5110	de
"030917"	74875	11525	6502	3980	de
"030918"	77892	11894	6672	4071	de
"030919"	100693	14702	8326	5137	de
"030920"	100292	14049	7819	5273	de
"030921"	102639	14586	8211	4948	de
"030923"	104638	13823	7602	5442	de
"030924"	82413	12257	6833	4315	de

"030925"	78846	11721	6570	4155	de
"030926"	106739	15093	8408	5231	de
"030927"	120516	16544	9258	5598	de
"030928"	73380	11356	6387	3906	de
"030930"	110095	15568	8739	5635	de
"031001"	85954	13032	7277	4473	de
"031002"	72540	11816	6779	3792	de
"031003"	104784	14584	7990	5457	de
"031004"	118509	16419	9112	5750	de
"031005"	74972	11851	6638	3757	de
"031007"	99443	13407	7323	5120	de
"031008"	81500	12698	7219	4295	de
"031009"	111183	14759	8382	4937	de
"031010"	118191	16132	9074	6155	de
"031011"	112323	15098	8362	5362	de
"031012"	83159	12568	6998	4189	de
"031014"	108584	13802	7402	5527	de
"031015"	89974	12988	7173	4692	de
"031016"	87704	13224	7487	4566	de
"031017"	111048	15410	8673	5566	de
"031019"	79706	12460	7018	3804	de
"031021"	108752	14489	7915	5540	de
"031022"	88478	12909	7237	4799	de
"031023"	87386	12622	7144	4271	de
"031024"	113844	15792	8779	5798	de
"031025"	110937	14962	8201	5285	de
"031026"	78315	12242	6875	3933	de
"031028"	102141	13684	7543	5294	de
"031029"	79568	12585	7093	3949	de
"031030"	81861	12422	7028	4190	de
"031031"	104513	14953	8389	5357	de
"031101"	105643	14944	8345	5025	de
"031102"	52445	10058	5924	2804	de
"031104"	104633	13677	7388	5374	de
"031105"	68685	11586	6697	3629	de
"031106"	81071	12225	6834	4172	de
"031107"	111285	15798	8900	5393	de
"031108"	105463	14506	8075	5059	de
"031109"	85302	12804	7195	4203	de
"031111"	102156	13684	7421	5365	de
"031112"	75596	11807	6603	3919	de
"031113"	76967	12120	6963	4029	de
"031114"	98389	14687	8299	5095	de
"031115"	94928	13958	7907	4656	de
"031116"	79363	12293	6986	4015	de
"031118"	116019	14637	7980	5824	de
"031119"	93895	14319	8269	5182	de
"031120"	91477	13681	7683	4637	de
"031121"	113437	15612	8734	5781	de
"031122"	101338	14520	8189	5324	de
"031123"	74021	11946	6728	3822	de
"031125"	103589	13269	7153	5482	de
"031126"	92825	13686	7601	4729	de
"031127"	96573	13996	7866	4567	de
"031128"	119165	15942	8916	5919	de
"031129"	103244	14560	8111	5128	de
"031130"	77452	11781	6574	3975	de
"031202"	107238	14011	7612	5185	de
"031203"	85134	12561	6955	4459	de
"031204"	87123	13786	7815	4534	de
"031205"	202201	21222	11358	9725	de
"031206"	103145	14363	8039	5102	de
"031207"	80957	12462	6924	3924	de
"031209"	99729	13748	7600	4922	de
"031210"	98224	13067	7006	4908	de
"031211"	75428	11958	6777	3890	de
"031212"	127029	16830	9224	6392	de
"031213"	97819	13355	7501	4823	de
"031214"	78778	11931	6685	4015	de
"031216"	115673	14392	7782	5822	de
"031217"	83939	12319	6824	4590	de
"031218"	78089	12139	6878	4423	de
"031219"	114529	15814	8842	5951	de
"031220"	93323	14222	8117	4728	de

"031221"	57030	10482	6163	2858	de
"031223"	53027	9641	5607	2659	de
"031224"	58348	10807	6338	2801	de
"031225"	56074	9748	5627	2948	de
"031226"	48958	8914	5167	2723	de
"031227"	69290	12095	7216	3531	de
"031228"	67983	11292	6555	3634	de
"031230"	52642	9643	5658	2753	de
"031231"	56048	10202	5949	2902	de
"040101"	54360	9808	5697	2819	de
"040102"	43396	8433	4950	2319	de
"040103"	74000	12422	7277	3571	de
"040104"	66877	10903	6181	3385	de
"040106"	101469	13654	7340	5380	de
"040107"	72926	11763	6670	3692	de
"040108"	71578	11586	6539	3795	de
"040109"	116952	16089	8768	6279	de
"040110"	103666	14322	7809	5028	de
"040111"	64191	10673	6092	3295	de
"040113"	94508	12995	7154	4644	de
"040114"	81149	12032	6757	4074	de
"040115"	75868	11762	6740	3984	de
"040116"	108921	15369	8623	5782	de
"040117"	95563	13872	7875	5036	de
"040118"	73335	11544	6397	3728	de
"040120"	112678	14965	8110	5839	de
"040121"	74214	11672	6513	3711	de
"040122"	80671	12290	6922	4166	de
"040123"	110148	15353	8556	5823	de
"040124"	110780	15333	8602	5284	de
"040125"	76958	11980	6659	3998	de

9.1.1.2 Spécifs - Part : DATE Parties sélectionnées : "040125"

On n'a conservé ici qu'un extrait du tableau construit

Forme	Frq. Tot.	Fréquence	Coeff.
banque	2464	106	***
donation	132	35	***
Johnny	604	68	***
Intenses	22	22	***
janvier	6981	150	***
Boudou	29	25	***
lentilles	44	26	***
Hallyday	282	50	***
packages	32	24	***
Song	107	30	48
fiscalement	46	24	47
patriotiques	52	25	47
2003	10899	158	45
Twombly	21	18	42
inexistante	65	24	42
tarifs	747	47	41
profils	128	28	41
Carmignac	76	24	40
Changer	59	22	39
Emergents	72	23	38
QUESTIONSà	67	22	37
exotisme	87	23	36
étudiés	92	23	36
désyndicalisation	23	16	35
inévitables	99	23	35
Eisner	54	20	35
courtiers	146	25	34
Joaquín	15	14	34
Montalvo	19	15	34
Cabrio	13	13	33
Soigneusement	20	15	33
Universal	1093	45	32
banques	2342	62	32
Vaconsin	15	13	31
aviaire	63	19	31
Immobilier	121	22	31
PT	183	25	31
fusions	189	25	31
clients	1442	48	30
Kay	107	21	30
services	5463	87	29
remanie	32	15	29
coûteux	196	24	29
Disney	408	30	29
Eva	200	24	29
matérielle	96	19	28
Frontières	92	19	28
Xiaonian	11	11	28
Culmell	11	11	28
Combien	225	24	28
épidémie	1048	40	27
calculées	49	16	27
Forza	99	19	27
spéculations	212	23	27
Nin	21	13	27
Maître	171	22	27
RS	16	12	27
Huchon	154	21	27
Jaime	45	15	26
Dicos	17	12	26
Sganzerla	10	10	25
grippe	217	22	25
montent	260	23	25

Singe	27	13	25
intempéries	67	16	25

9.1.1.3 Graphes de ventilation de formes

Les formes examinées ici sont : voile, laïcité, croix, kippa

Fréquences relatives :

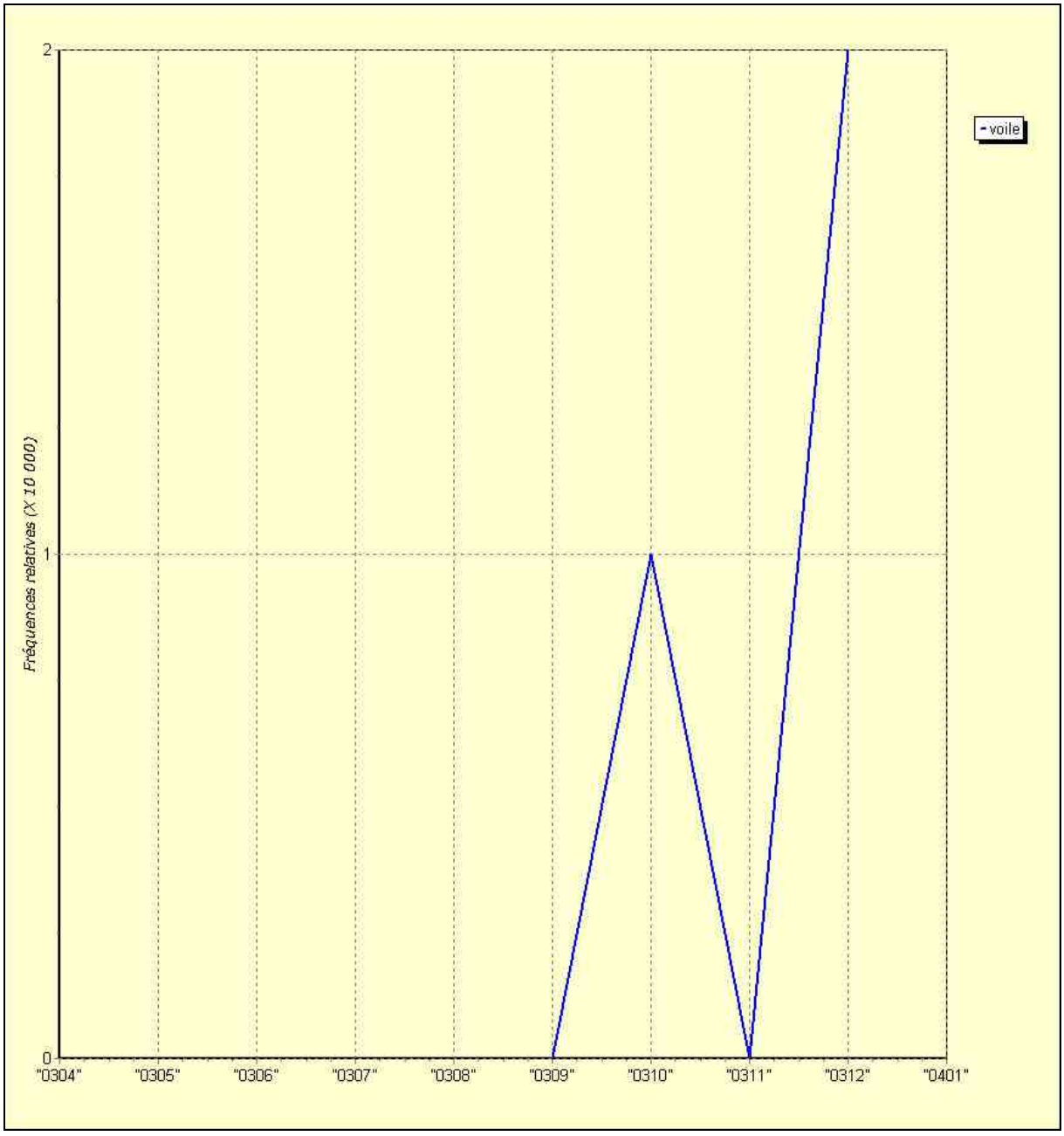


Figure 42 : Graphique de ventilation (voile) 1/2

Fréquences absolues :

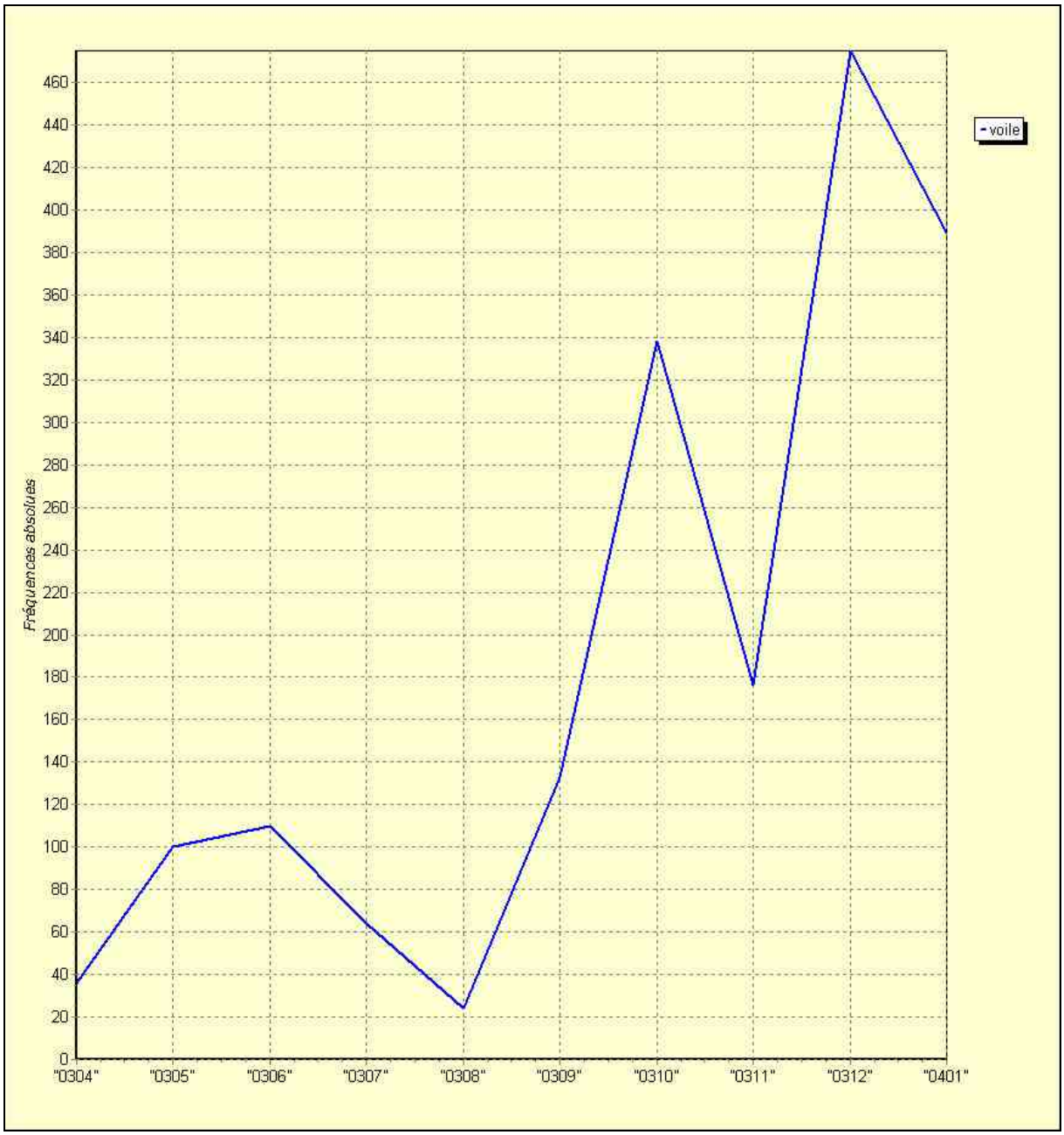


Figure 43 : Graphique de ventilation (voile) 2/2

Fréquences absolues :

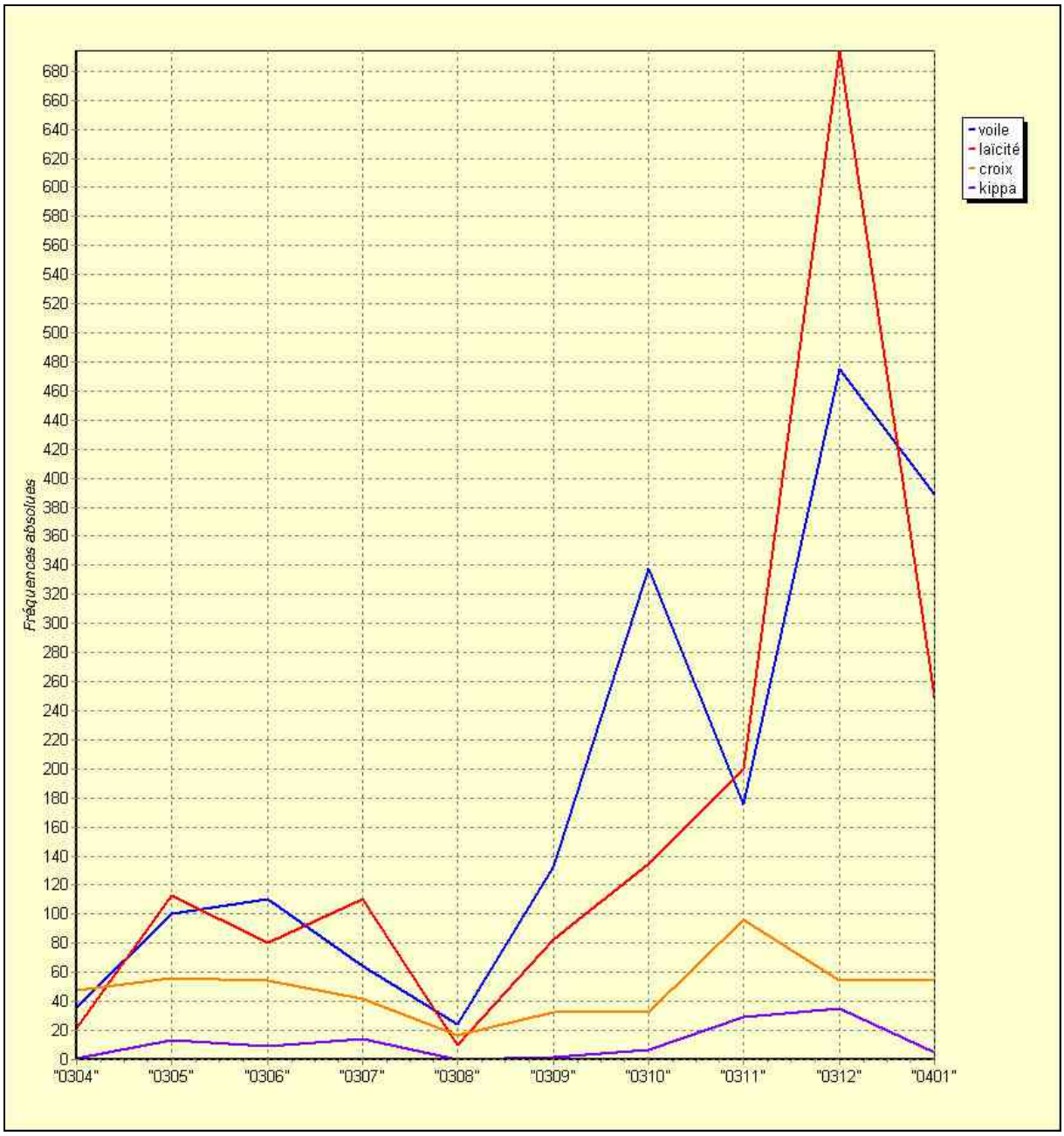


Figure 44 : Graphique de ventilation (voile, laïcité, croix, kippa) 1/2

Spécificités :

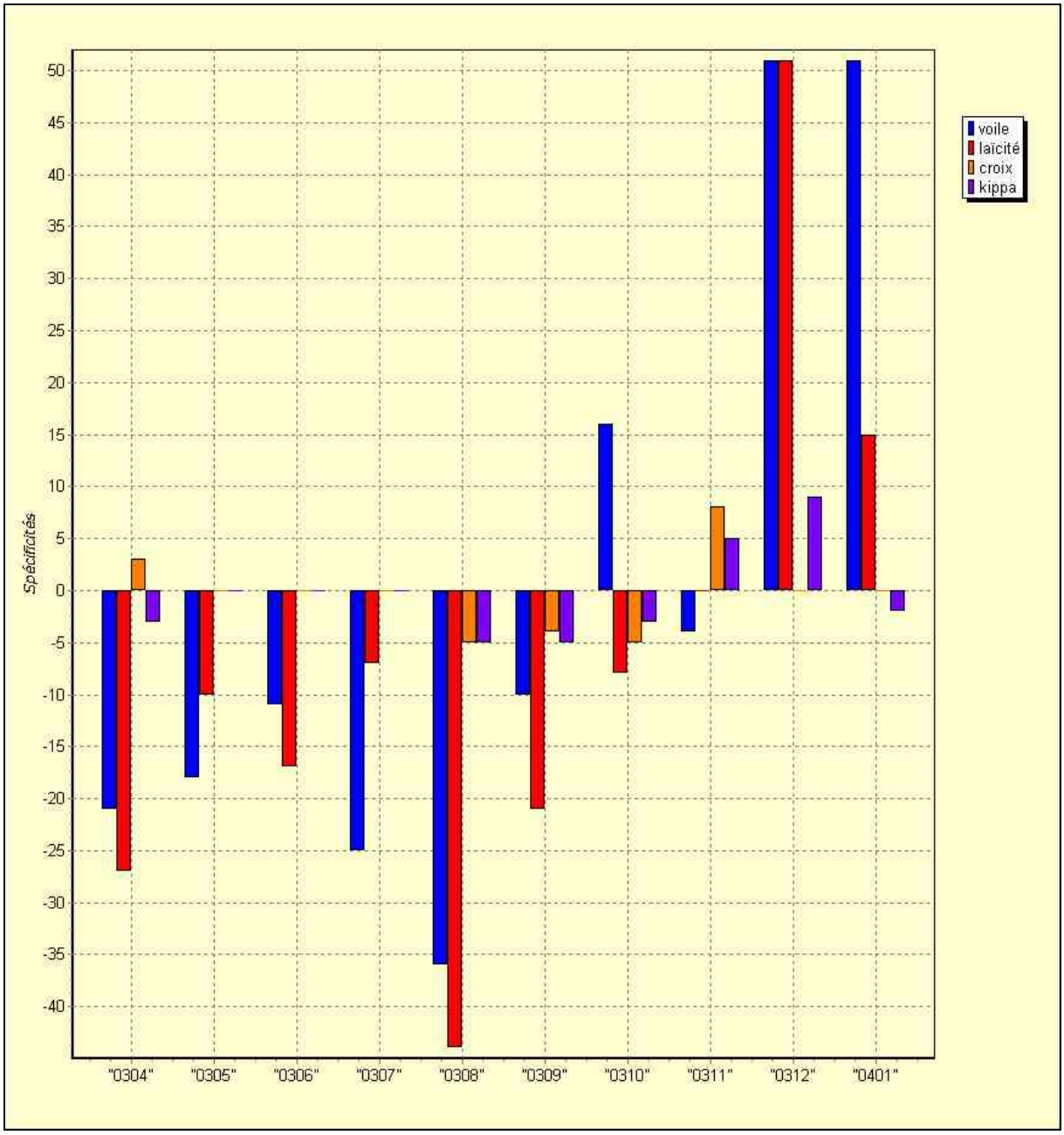


Figure 45 : Graphique de ventilation (voile, laïcité, croix, kippa) 2/2

9.2 Annexe Partie 2

9.2.1 Liens/projets

9.2.1.1 Les nuages de Tags chez Technocrati[18]

Principe :

- *Arranging the words and terms in one paragraph, and*
- *Varying the font-sizes to represent the popularity of a keyword/ tag.*

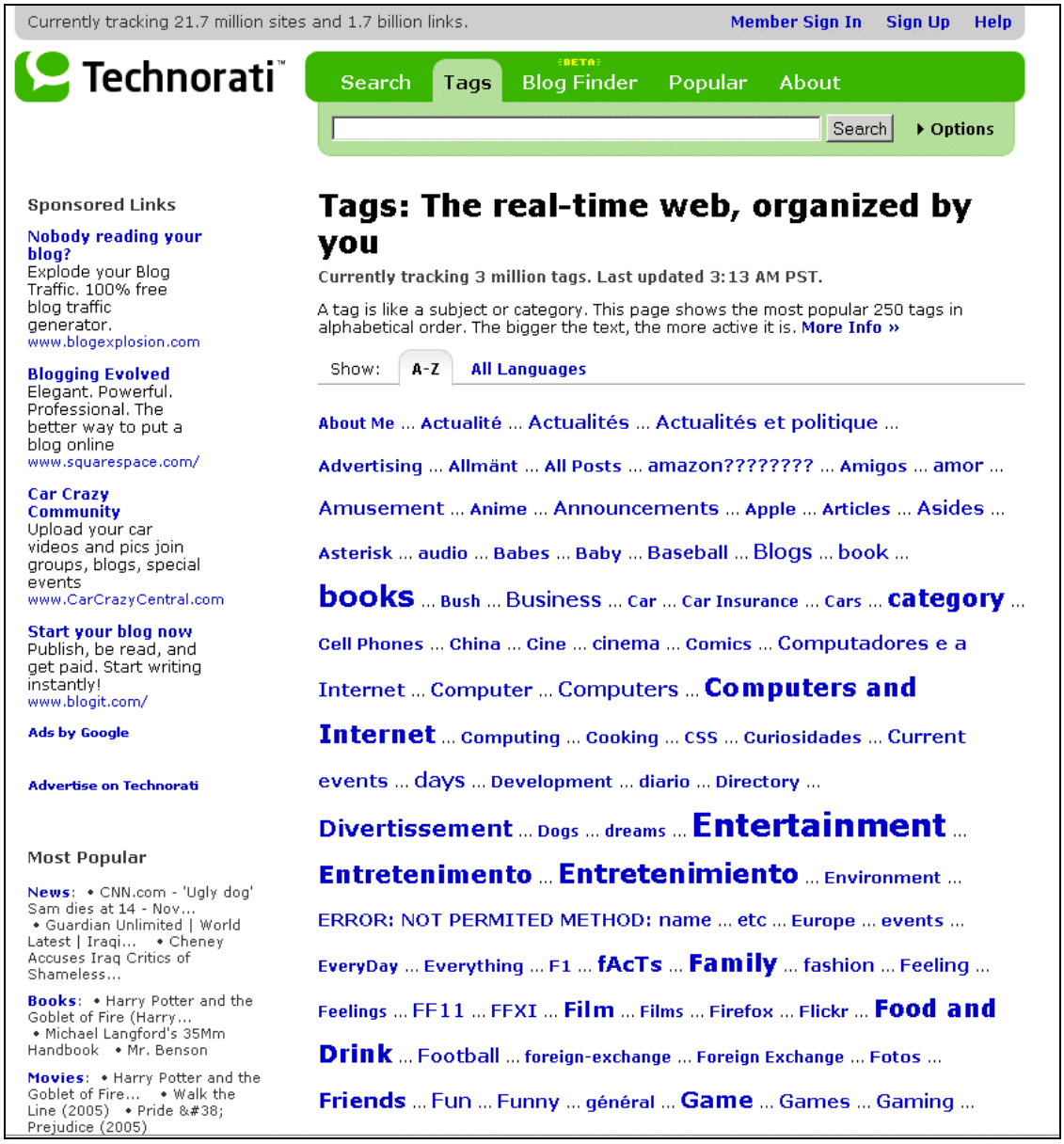


Figure 46 : Nuages de TAG

[18] http://www.technorati.com/tags/

A tag is like a subject or category. This page shows the most popular 250 tags in alphabetical order. The bigger the text, the more active it is.

Plus d'infos : http://www.technorati.com/help/tags.html

9.2.1.2 Annuaire de Fils

LaMoooche.fr[19] est un système d'information en perpétuelle évolution qui récupére périodiquement les actualités issues de plus de **1 000 diffuseurs de contenu** (LeMonde, Libération, Le Nouvel Observateur, Clubic, Jeux Video.com, ...). Ce procédé s'appelle l'agrégation de contenu[20] (procédé de lecture et de stockage d'articles issus de plusieurs fils d'information).

Annuaire Actualités :

http://www.lamoooche.com/2,1,annuaire-rss-actualite.html

9.2.1.3 AlertInfo, un agrégateur RSS de la presse française

Source : http://www.geste.fr/alertinfo/home.html

Communiqué de presse

26 mai 2005

France Télévisions Interactive, Le Monde Interactif, RTL Net, Libération.fr, Les Echos.fr, Le NouvelObs.com, L'Equipe.fr, 01Net, ZDNet, La Tribune.fr, Le Figaro.fr, LExpress.fr, LExpansion.com, LEntreprise.com, BusinessMobiles.fr, tous membres du GESTE, lancent, le 26 mai prochain, AlertInfo, lecteur RSS légal et gratuit des médias d'information français, proposant, dès son installation, près de274 fils d'informations ciblées.

Cette initiative, première mondiale, vise à apporter aux internautes francophones un outil d'information exceptionnel, mis à jour en permanence, promu et réalisé directement par les éditeurs et les rédactions des grands médias électroniques français.

En un seul clic, l'utilisateur aura accès à ses infos préférées : France, International, Business/Entreprises, Communication/Médias/HighTech, Emploi/RH/Métiers, Etudiants/Formations, Solutions IT/Informatique/Matériels, Les Marchés/Investisseurs, Loisirs/Week-end/Culture, Musique, Patrimoine, Sports, Régions, Sciences, Insolite/People, Féminin.

Téléchargeable gratuitement sur le site du GESTE www.geste.fr et sur l'ensemble des sites des éditeurs présents dans le lecteur, AlertInfo permet à chacun de sélectionner ses thématiques et ses sources préférées. Les dernières infos sont présentées par titre et par ordre de mise à jour. Lorsque l'internaute clique sur un titre, le « chapô » se développe dans une partie du lecteur et propose un lien vers le site de l'éditeur pour lire l'article. Pour les articles payants, si l'internaute est abonné, il accède directement à l'article, sans avoir à s'identifier.

Pour être présent dans le lecteur RSS du GESTE, l'éditeur doit être membre du groupement (radio, presse, pure player ou télévision, hors agences de presse), et détenteur des droits des articles qu'il propose. Les textes proposés doivent être des textes d'information et non des textes de promotion.

---

[19] http://www.lamoooche.fr/

[20] http://www.lamoooche.com/definition/agregation.php

Avec AlertInfo, les éditeurs ont souhaité apporter une réponse légale aux attentes des internautes en matière d'information. La réutilisation des fils d'information ainsi mise à la disposition des internautes est soumise aux conditions de chacun des éditeurs.

Le lecteur AlertInfo est le fruit d'un travail d'équipe. Les équipes de développement des echos.fr ont réalisé une version française de la technologie Feedreader (application issue du monde du logiciel libre), en collaboration avec son initiateur, Toomas Toots, à partir de laquelle chacun a apporté sa pierre pour consolider l'édifice (rédaction de l'aide, création d'une charte et de conditions d'utilisation, définition des catégories, etc.).

AlertInfo propose de multiples fonctionnalités, notamment :

- La possibilité d'ajout ou de retrait de fils d'information ;
- La visualisation des « chapôs » (si disponibles) dans une partie de l'écran ;
- la possibilité de trier les fils d'information par catégorie ou par éditeur
- La possibilité d'envoyer l'url d'un article à un ami ;
- et l'envoi d'alertes à chaque nouvelle mise à jour, etc.

Éditeurs présents au lancement d'AlertInfo : lesechos.fr, latribune.fr, lemonde.fr, lefigaro.fr, lentreprise.com, lexpress.fr, lexpansion.com, france2.fr, liberation.fr, rtl.fr, nouvelobs.com, france3.fr, lequipe.fr, 01net.fr, zdnet.fr, businessmobile.fr et, très prochainement, france5.fr et m6.fr

Contacts

Laure de Lataillade : contact@alertinfo.fr

Astrid Flesch : a.flesch@geste.fr

Tél. 01 55 62 00 70

A propos du Geste :

Le GESTE, qui regroupe les principaux éditeurs de contenus sur internet (presse, radios, télévisions, éditeurs indépendants), a pour objet de créer les conditions économiques, législatives et concurrentielles indispensables au développement des services et éditions électroniques. Avec plus d'une centaine de sociétés membres, le GESTE poursuit sa constante progression et s'est imposé comme l'interlocuteur privilégié et incontournable en matière de contenus en ligne.

Pour télécharger l'outil : http://www.geste.fr/alertinfo/telecharger.html

Mode d'emploi : http://www.geste.fr/alertinfo/modedemploi.html

9.2.1.4 Amazon concordance

Concordance

Concordance is an alphabetized list of the most frequently occurring words in a book, excluding common words such as "of" and "it." The font size of a word is proportional to the number of times it occurs in the book. Hover your mouse over a word to see how many times it occurs, or click on a word to see a list of book excerpts containing that word.

Please send your feedback on this feature to sitb-feedback@amazon.com

Sur le site Amazon.com, présentation du livre : "*In the Beginning...was the Command Line*", par Neal Stephenson, 1999 :

Figure 47 : Amazon "In the Beginning...was the Command Line[21]"

En passant la souris sur l'image de la couverture du livre, on accède au menu suivant :

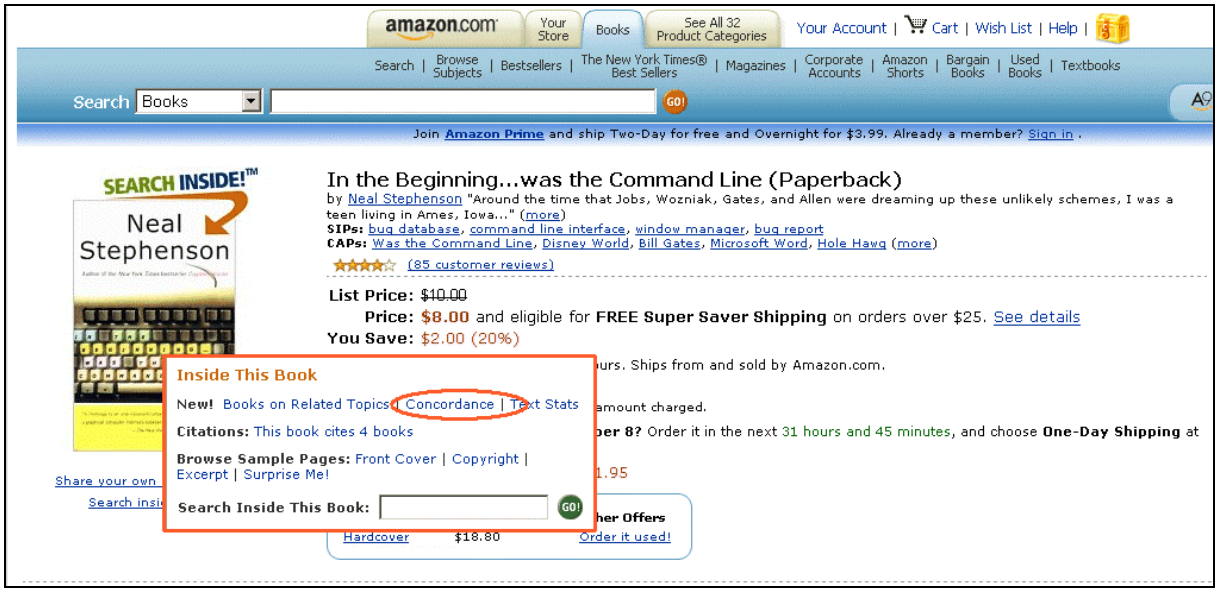


Figure 48 : Menu "concordance" sur Amazon

Ce menu donne accès à un programme « *Concordance* » qui construit dans un premier temps un nuage de mots (les 100 mots les plus fréquents du livre) :

[21] http://www.amazon.com/exec/obidos/tg/detail/-/0380815931/002-1510917-1432801?v=glance

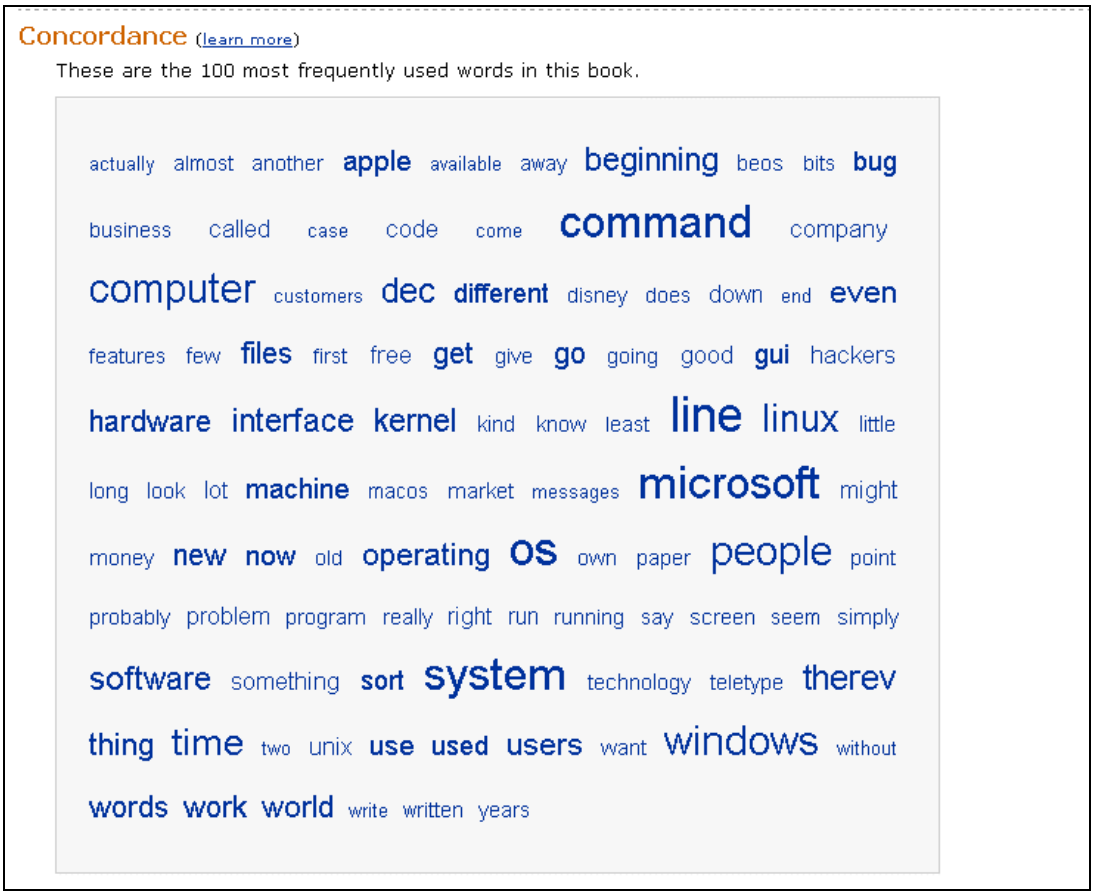
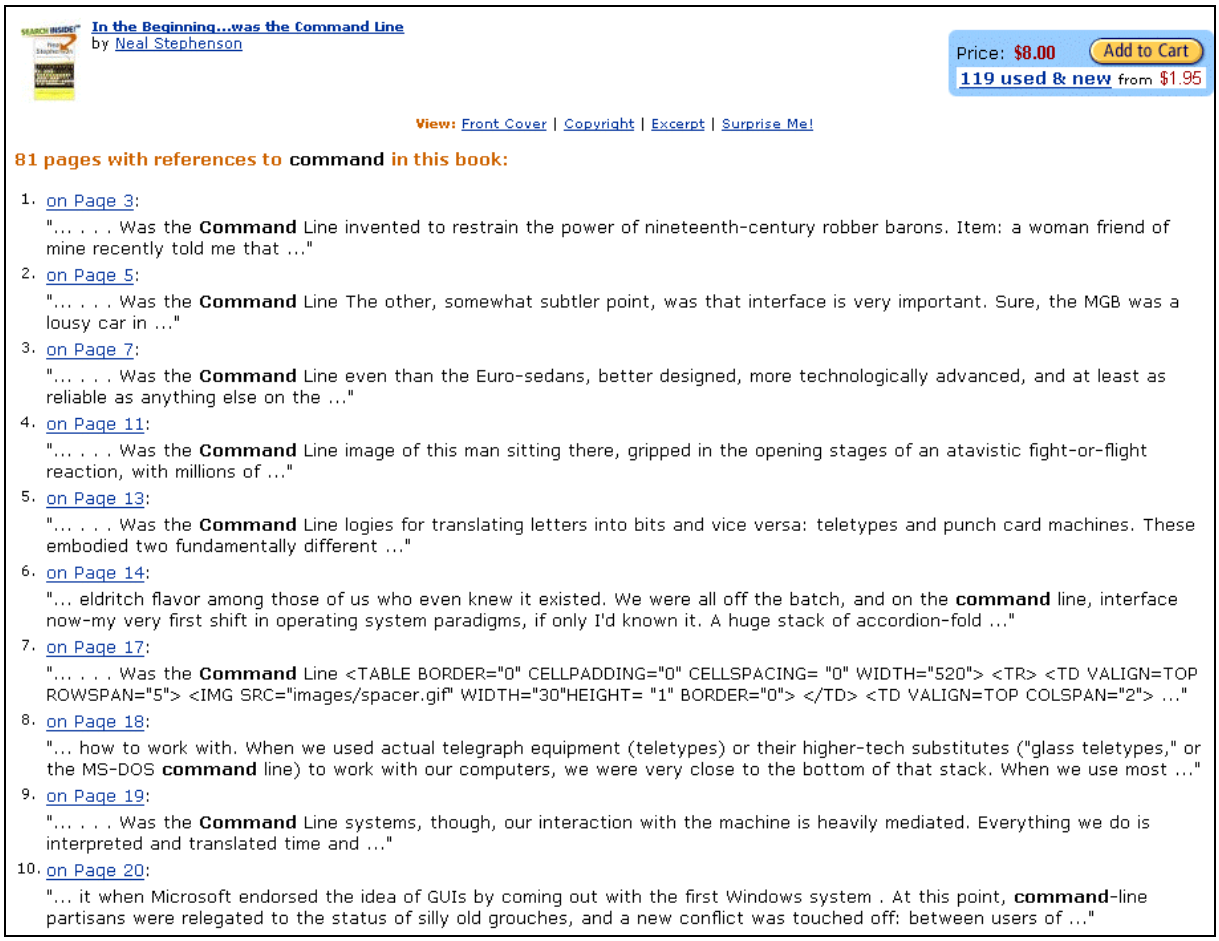


Figure 49 : Nuage de mots "concordance[22]"

chaque mot est ensuite « cliquable » et donne ainsi accès aux contextes du mot visé :

[22] http://www.amazon.com/gp/product/sitb-next/0380815931/ref=sbx_con/002-1510917-1432801?%5Fencoding=UTF8#concordance



In the Beginning...was the Command Line
by Neal Stephenson

Price: $8.00 Add to Cart
119 used & new from $1.95

View: Front Cover | Copyright | Excerpt | Surprise Me!

81 pages with references to command in this book:

1. on Page 3:
"... . . . Was the **Command** Line invented to restrain the power of nineteenth-century robber barons. Item: a woman friend of mine recently told me that ..."
2. on Page 5:
"... . . . Was the **Command** Line The other, somewhat subtler point, was that interface is very important. Sure, the MGB was a lousy car in ..."
3. on Page 7:
"... . . . Was the **Command** Line even than the Euro-sedans, better designed, more technologically advanced, and at least as reliable as anything else on the ..."
4. on Page 11:
"... . . . Was the **Command** Line image of this man sitting there, gripped in the opening stages of an atavistic fight-or-flight reaction, with millions of ..."
5. on Page 13:
"... . . . Was the **Command** Line logies for translating letters into bits and vice versa: teletypes and punch card machines. These embodied two fundamentally different ..."
6. on Page 14:
"... eldritch flavor among those of us who even knew it existed. We were all off the batch, and on the **command** line, interface now-my very first shift in operating system paradigms, if only I'd known it. A huge stack of accordion-fold ..."
7. on Page 17:
"... . . . Was the **Command** Line <TABLE BORDER="0" CELLPADDING="0" CELLSPACING= "0" WIDTH="520"> <TR> <TD VALIGN=TOP ROWSPAN="5"> <IMG SRC="images/spacer.gif" WIDTH="30"HEIGHT= "1" BORDER="0"> </TD> <TD VALIGN=TOP COLSPAN="2"> ..."
8. on Page 18:
"... how to work with. When we used actual telegraph equipment (teletypes) or their higher-tech substitutes ("glass teletypes," or the MS-DOS **command** line) to work with our computers, we were very close to the bottom of that stack. When we use most ..."
9. on Page 19:
"... . . . Was the **Command** Line systems, though, our interaction with the machine is heavily mediated. Everything we do is interpreted and translated time and ..."
10. on Page 20:
"... it when Microsoft endorsed the idea of GUIs by coming out with the first Windows system . At this point, **command**-line partisans were relegated to the status of silly old grouches, and a new conflict was touched off: between users of ..."

Figure 50 : Contextes "Concordance"

9.2.1.5 TagClouds (« Nuage de mots »)

URL du projet : http://tagcloud.com

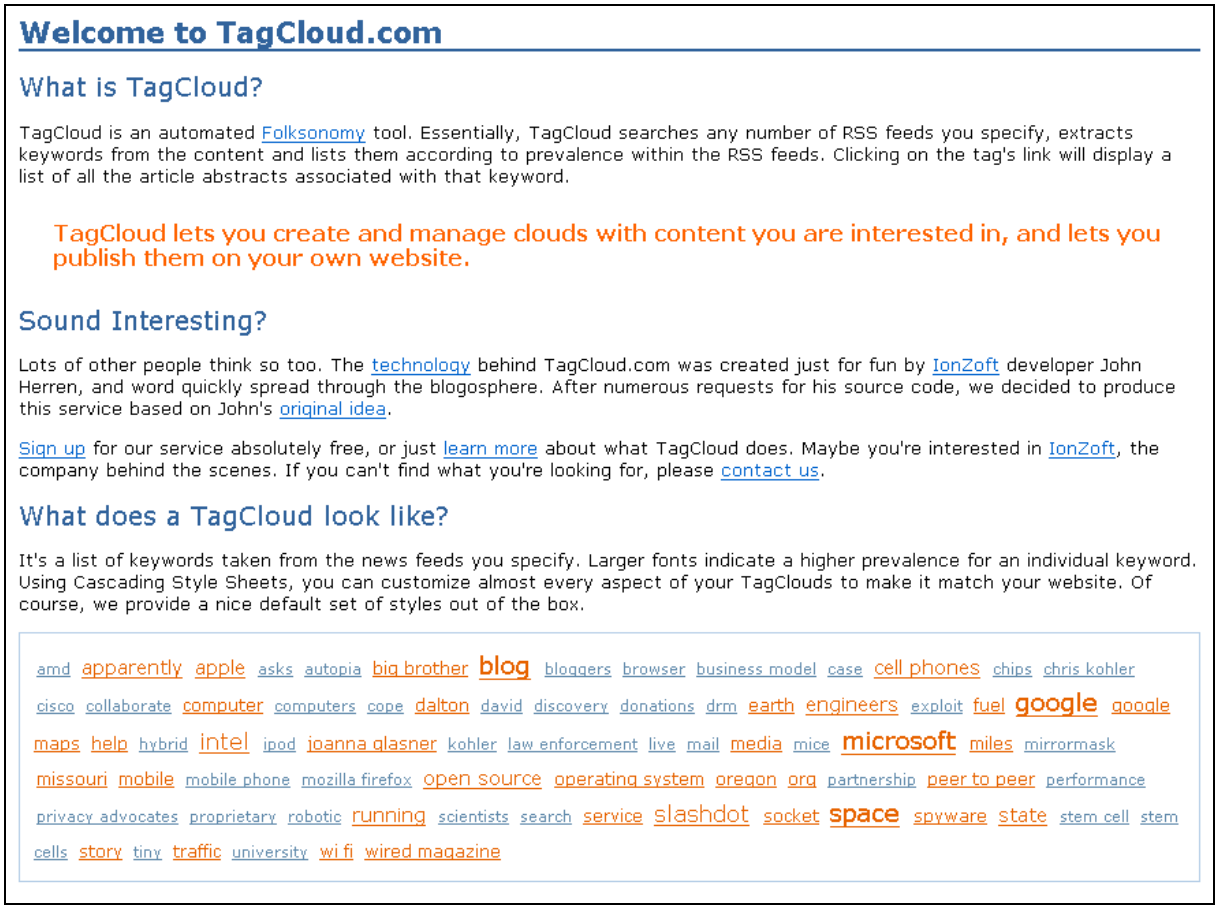


Figure 51 : projet TagCloud

Application : «TagCloud Le Monde »

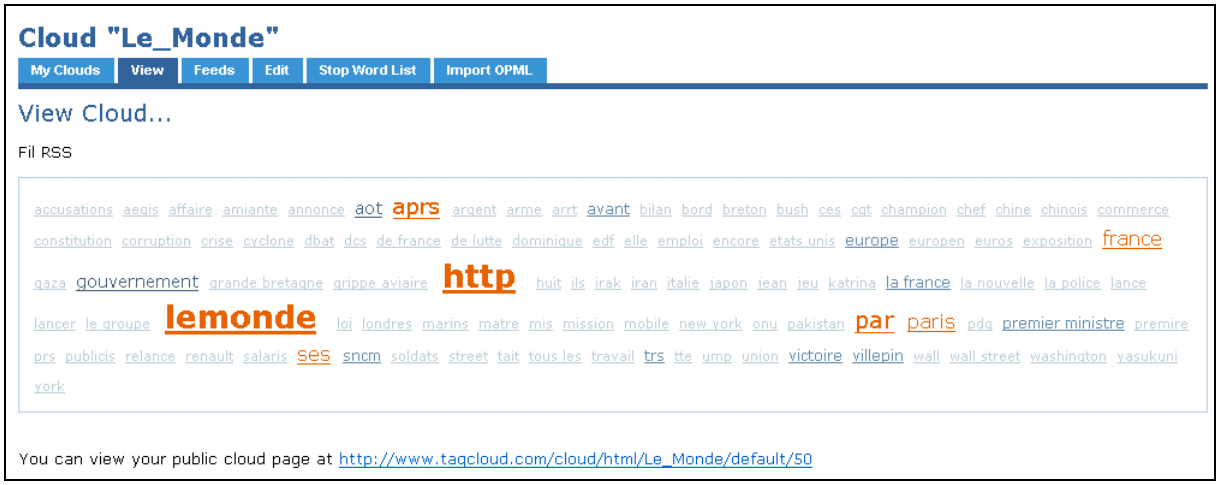


Figure 52 : tagcloud sur Fils du Monde

Les fils utilisés pour construire ce nuage sont paramétrables via l'onglet *Feeds* visible dans la figure précédente :

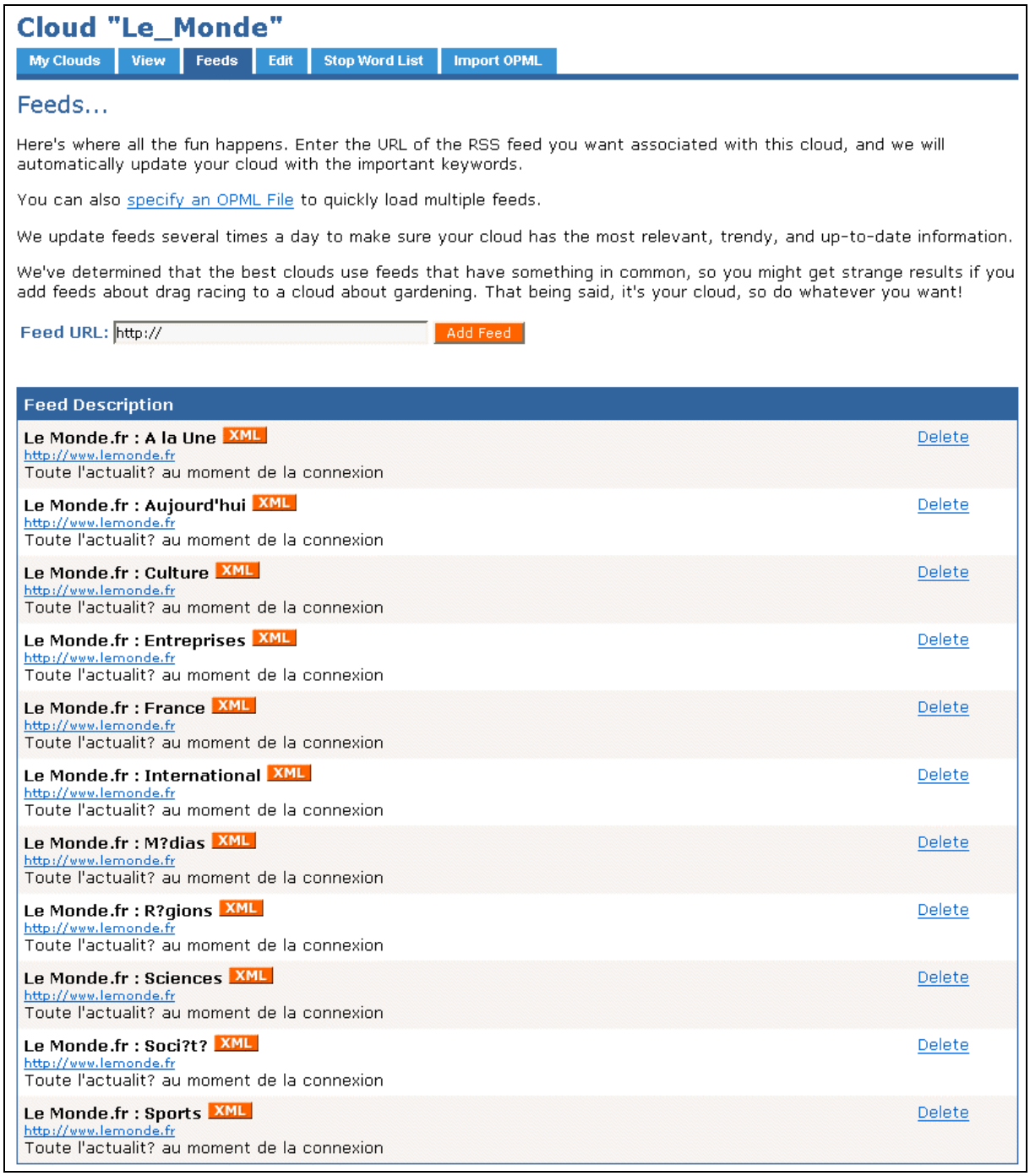


Figure 53 : Paramétrage du tagcloud LeMonde

9.2.1.6 Le filtre Google

http://www.marumushi.com/apps/newsmap/newsmap.cfm

Newsmap is an application that visually reflects the constantly changing landscape of the Google News news aggregator. A treemap visualization algorithm helps display the enormous amount of information gathered by the aggregator. Treemaps are traditionally space-constrained visualizations of information. Newsmap's objective takes that goal a step further and provides a tool to divide information into quickly recognizable bands which, when presented together, reveal underlying patterns in news reporting across cultures and within news segments in constant change around the globe

Newsmap does not pretend to replace the googlenews aggregator. It's objective is to simply demonstrate visually the relationships between data and the unseen patterns in news media. It is not thought to display an unbiased view of the news, on the contrary it is thought to ironically accentuate the bias of it.

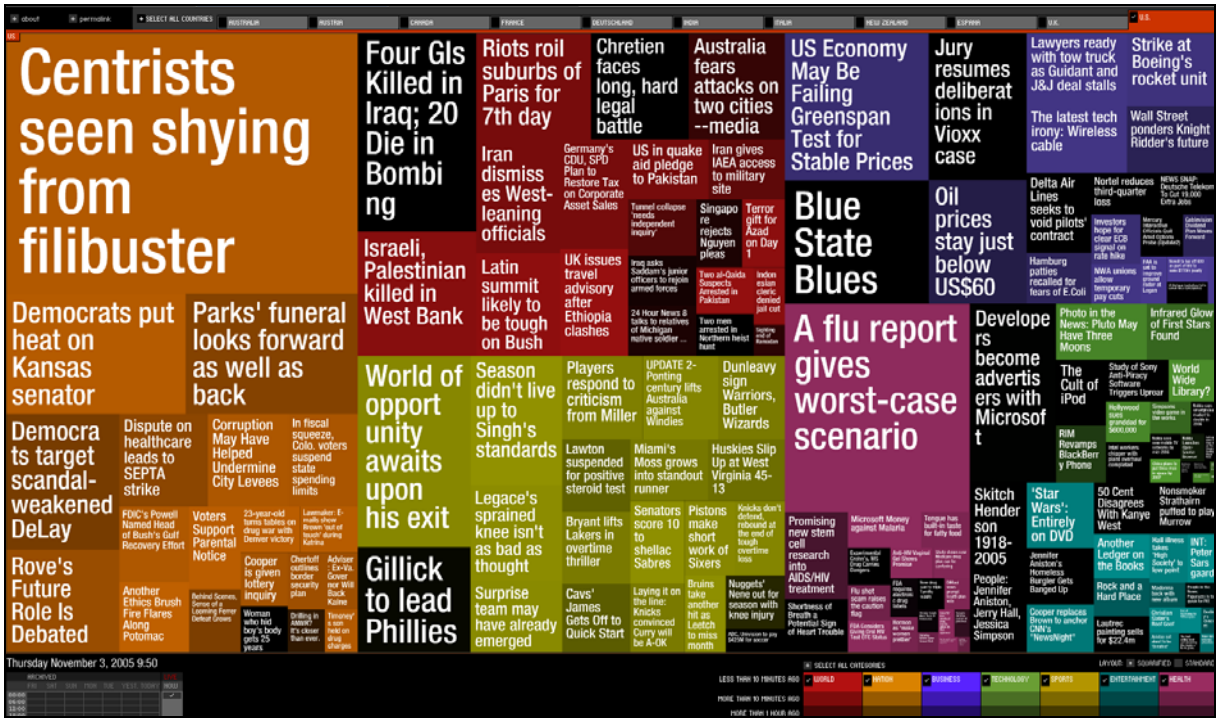


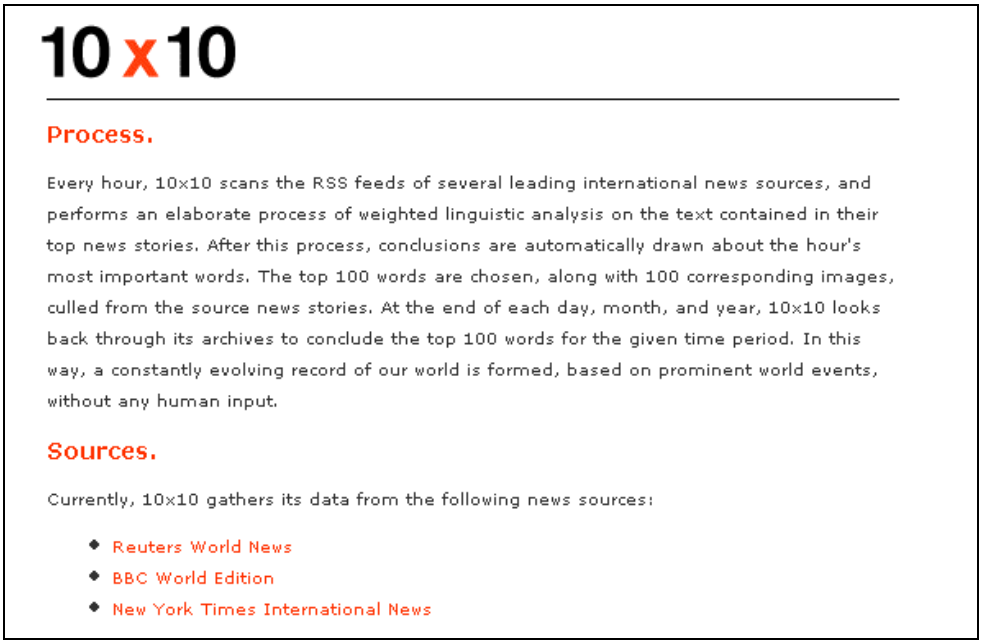
Figure 54 : Projet NewsMap

Figure 55 : Projet NewsMap (france)

9.2.1.7 10x10 : images du monde

Cet outil d'exploration interactive passe au crible les fils RSS de Reuters, de la BBC et du NewYorker pour créer une mosaïque de photos d'actualités.

Lien : http://tenbyten.org/index.html



10 x 10

10x10™ ('ten by ten') is an interactive exploration of the words and pictures that define the time. The result is an often moving, sometimes shocking, occasionally frivolous, but always fitting snapshot of our world. Every hour, 10x10 collects the 100 words and pictures that matter most on a global scale, and presents them as a single image, taken to encapsulate that moment in time. Over the course of days, months, and years, 10x10 leaves a trail of these hourly statements which, stitched together side by side, form a continuous patchwork tapestry of human life.

10x10 is ever-changing, ever-growing, quietly observing the ways in which we live. It records our wars and crises, our triumphs and tragedies, our mistakes and milestones. When we make history, or at least the headlines, 10x10 takes note and remembers.

Each hour is presented as a picture postcard window, composed of 100 different frames, each of which holds the image of a single moment in time. Clicking on a single frame allows us to peer a bit deeper into the story that lies behind the image. In this way, we can dart in and out of the news, understanding both the individual stories and the ways in which they relate to each other.

10x10 runs with no human intervention, autonomously observing what a handful of leading international news sources are saying and showing. 10x10 makes no comment on news media bias, or lack thereof. It has no politics, nor any secret agenda; it simply shows what it finds.

With no human editors and no regulation, 10x10 is open and free, raw and fresh, and consequently a unique way of following world events. In 10x10, we respond instinctively to patterns in the grid, visual indicators of relevance. When we see a frequently repeated image, we know it's important. When we see a picture of a movie star next to a picture of dead bodies, we understand the extremes that exist in our world. Scanning a grid of pictures can be more intuitive than reading headlines, for it lets the news come to life, and everything feels a bit less distant, a bit closer to heart, and maybe, if we're lucky, gives us pause to think. If you'd like to learn more about 10x10, you can read how it works.

Credits.

27 FABRICA

10x10 was designed and developed by Jonathan Harris of Number27, in conjunction with the FABRICA communication research center in Italy.

Special thanks to: Andy Cameron, Joel Gethin Lewis, Francesca Granato, David Towey

Figure 56 : Projet 10x10

9.2.1.8 Projet « Post Remix[23] » (Washington Post[24])

Présentation de ce projet à partir d'un billet publié le 27 novembre 2005 sur le weblog « La feuille[25] » :

Anticiper les usages des lecteurs

http://lafeuille.blogspot.com/2005/11/anticiper-les-usages-des-lecteurs.htm

Récemment, en France, le débat vieux médias/nouveaux médias a été relancé, à l'occasion de l'annonce du basculement de Libération vers un modèle de publication "bi-médias", après la refonte récente des maquettes de ses deux grands concurrents nationaux : Le Monde et Le Figaro, explicitement pensées comme repositionnées par rapport à Internet.
Une autre manière de présenter l'information, une autre information, une plus grande rapidité, un autre ton, une plus grande interactivité ; c'est toujours sous cet aspect que les médias classiques semblent devoir présenter leur stratégie de communication sur Internet. Il me semble pourtant qu'ils manquent l'essentiel, en considérant toujours leurs lecteurs comme...des lecteurs justement, sans jamais se demander ce qu'ils vont bien pouvoir faire des informations auxquelles ils ont accès.

Donner à d'autres la possibilité de faire quelque chose des informations que l'on publie, les autoriser et leur permettre techniquement de construire des services à partir d'un flux d'informations et de remixer ce flux pour offrir des contenus recomposés, c'est ce que vient de faire le Washington Post en ouvrant son service "Post remix[26]". Il s'agit ni plus ni moins de permettre à quiconque de programmer des mashups du Washington Post sur le base d'API fournies par le journal. Un premier mashup permet de créer un flux RSS sur les résultats de recherche[27] par mots-clés sur les résultats du Post, et un autre, très intéressant, représente ces mots-clés en nuage de tags[28], à la Del.icio.us.

C'est assez futé tout de même, car le journal externalise ainsi à bon compte sur sa communauté de lecteurs tous les services qu'il aurait pu développer lui-même. Par ailleurs, la licence d'utilisation[29] vaut le coup d'être lue :

Your efforts must be for personal, and not for commercial, use. You may not sell applications that use or incorporate washingtonpost.com content.

You recognize that Washingtonpost.Newsweek Interactive retains all intellectual property rights in all washingtonpost.com content and you that acquire no such rights by participating in Post Remix.

Washingtonpost.com may incorporate your ideas into future projects it develops.

Apprécions en particulier la dernière clause, assez savoureuse dans le style faut-pas-se-gêner... C'est d'ailleurs tout le problème des remix[30], qui reposent le plus souvent sur des bases de coopération pas claires et totalement déséquilibrées.

---

[23] http://blogs.washingtonpost.com/post_remix/

[24] http://www.washingtonpost.com/?nav=globaltop

[25] http://lafeuille.blogspot.com/

[26] *cf supra*

[27] http://socialistsushi.com/wp/

[28] http://www.revsys.com/newscloud/

[29] http://blogs.washingtonpost.com/post_remix/2005/11/terms_of_use.html

[30] http://www.readwriteweb.com/archives/002829.php

Au delà d'un rapport de force qui devra nécessairement s'équilibrer, l'exemple est quand même inéressant : voilà un grand journal qui commence à repenser sa position dans la chaîne de circulation de l'information et considère davantage ses lecteurs comme des partenaires. A méditer.

L'application « nuage de tags » présentée dans ce billet (**NewsCloud)** donne à voir un processus similaire à celui mis en œuvre dans le projet présenté dans ce document. La page d'accueil du site Newscloud est présentée ci-dessous :

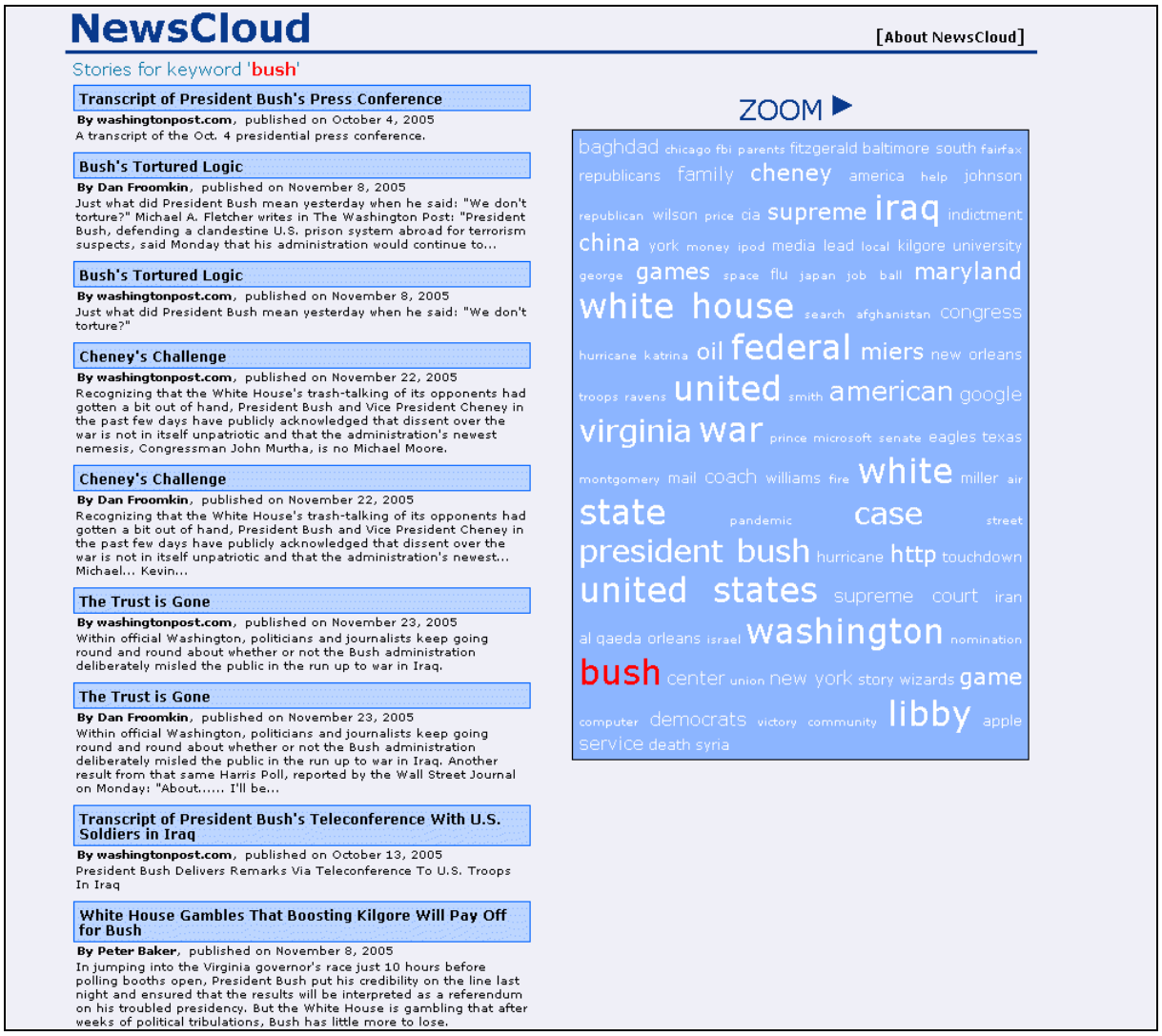


Figure 57 : Projet NewsCloud (Washington Post)

About NewsCloud

NewsCloud is an application that takes all of the RSS feeds from the Washington Post website and builds a blog like tag cloud from the keywords. Each story's full text is pulled from the website and indexed by keywords thses keywords. There are typically around 11,000 news stories and 60,000 keywords being indexed at any given time.

How to use NewsCloud

When you first go to NewsCloud you are seeing the outer most zoom *of the cloud. The outer most level is where the most popular keywords are. The farther you zoom into the cloud the frequency of the keywords is reduced. You can zoom by clicking the arrow to the right of the big* ***ZOOM*** *in the upper right hand corner. Zooming to find less frequent keywords can reveal some interesting topics just below the surface.*

At each zoom level, including the outer most, you will see keywords that are in ***red****. This is the most frequent keyword at the zoom level you are currently on. As you zoom the stories on the left change. These stories are the ones that contain the keyword in* ***red****.*

To view the articles associated with any keyword on the page simply click on the keyword and the articles will be shown on the left.

Technology Used

NewsCloud was written by Frank Wiles as an experiment. It uses a slighlty non-standard LAMP like environment. Typical LAMP application use Linux, Apache, MySQL, and a "P" lanugage such as Perl, PHP, or Python. The following technologies are used in NewsCloud:

Linux

Apache

mod_perl

PostgreSQL

This slightly different LAMP stack is the preferred development environment of Revolution Systems. Please visit the respective technology pages or contact us for more information about how your business can benefit from these Open Source technologies.

9.2.2 Lectures

9.2.2.1 Conversation : De la représentation visuelle à la complexité documentaire

Source : http://affordance.typepad.com/mon_weblog/2005/11/de_la_reprsenta.html

A moins que ce ne soit l'inverse : De la représentation documentaire à la complexité visuelle. Le site Visual Complexity[31] propose, classés par thèmes (biologie, arts, réseaux sociaux, web...), des projets (231 au total) de visualisation de masses complexes d'informations et/ou de documents. Au-delà du côté simplement esthétique de ces représentations de la complexité, au-delà également de l'enjeu technique et (parfois) algorithmique que ces mêmes représentations supposent, elles illustrent parfaitement ce qu'est le principe de "tertiarisation documentaire" : après les documents primaires (ouvrages 'originaux'), après les documents secondaires (ouvrages décrivant le contenu des premiers), voici - au même titre que les cartes heuristiques - les documents tertiaires qui font sens en eux-mêmes (dans la mesure ou ils offrent leurs propres parcours interprétatifs) et/mais n'existent que parce qu'il renvoient (au sens propre et non 'intertextuel') vers d'autres. Ce nouveau genre documentaire m'avait frappé lors de ma découverte des premières cartes du métamoteur Kartoo[32] (on a les illuminations qu'on peut, n'est pas Claudel qui veut ...). Il avait aussi frappé mon collègue Gabriel Gallezot (évoquant un 'Darwinisme documentaire'), à tel point que nous avons commis quelques articles effleurant le sujet. Et le gars Roger, il en pense quoi ?

(Ndt : Roger = Roger Pedauque

cf http://rtp-doc.enssib.fr/rubrique.php3?id_rubrique=13)

9.2.2.2 Blog Technologies du Langage[33] *(par Jean Véronis)*

Plusieurs billets sont consacrés à une thématique proche.

Texte: Chirac sur un nuage

http://aixtal.blogspot.com/2005/11/texte-chirac-sur-un-nuage.html

Blogs: Banlieues dans les nuages

http://aixtal.blogspot.com/2005/11/blogs-banlieues-dans-les-nuages.html

Blogs: Un nuage sur les banlieues

http://aixtal.blogspot.com/2005/11/blogs-un-nuage-sur-les-banlieues.html

Dialogue entre blogues

http://aixtal.blogspot.com/2005/10/rcr-dialogue-entre-blogues.html

---

[31] http://www.visualcomplexity.com/vc/

[32] http://www.kartoo.fr/

[33] http://aixtal.blogspot.com/

9.2.2.3 Bibliothèque 2.0[34]

Article paru sur le blog ***bibliosession***[35] :

On parle souvent de Web 2.0[36]*, et de toutes les usages sociaux qui l'accompagnent. Fred Cazzava fait d'ailleurs le point sur ce sujet*[37] *en prenant pas mal d'exemples et surtout en recentrant le débat sur les usages. Ce qu'il y a peut-être de plus important à comprendre peut se résumer dans ce qui était cité par Hubert Guillaud*[38] *qui citait lui-même Cyberlibris blog*[39]*. En Substance:* "On oublie trop souvent l'utilisateur et on ne ne se préoccupe que du livre, de ses ayants-droit (les maisons d'édition) et de ceux qui aimeraient accéder (par des moyens pas toujours orthodoxes) au copyright des ayants-droit (Google et les autres). Mais, où est donc passé l'utilisateur. Est-il si peu important qu'il n'y a rien à en dire? Je pense qu'il y a là une erreur de perspective fondamentale. La bataille de l'émancipation de la musique et de l'image a été gagnée par les utilisateurs (et les pressions plus ou moins hardies qu'ils ont exercées). Il en va de même du livre. Lorsque l'on interroge les utilisateurs (ce que nous avons fait), que souhaitent-ils vraiment à propos du livre? Trois choses principales:

Pertinence: *L'utilisateur veut pouvoir accéder aux livres dont il a besoin. Malheureusement, cette demande est loin d'être satisfaite par les circuits existants. Une librairie, si vaste soit-elle, ne peut stocker tous les livres. Très souvent, elle ne stocke que ce qui se vend. Pour passer des journées entières dans les catalogues d'éditeurs, je suis tout à la fois admiratif de la richesse de l'esprit humain et consterné que si peu en soit visible.*

Immédiateté: *L'utilisateur a besoin du contenu "maintenant", c'est-à-dire au moment où son besoin d'information s'exprime. Il ne s'agit pas d'avoir une réponse demain. L'utilisateur est prêt à payer cette instantanéité de réponse.*

Ubiquité: *L'utilisateur souhaite obtenir une réponse à ses besoins d'information où qu'il se trouve. Il est prêt à payer cette ubiquité documentaire.*

Si l'on rassemble ces trois exigences à l'instar d'un portrait chinois, on découvre le format approprié à les satisfaire: il s'agit d'une ***bibliothèque digitale."***

A quoi peut donc ressembler cette bibliothèque digitale?

Et bien cet article[40] cité par l'excellent blog Librarian in Black[41] explique que suite à l'Internet Librarian Conférence[42] qui s'est tenue récemment aux Etats-Unis, un groupe d'une centaine de bibliothécaires à souhaité se réunir pour réflechir à la prise en compte des usages permis par le Web 2.0 dans les bibliothèques. Ce groupe s'est désigné tout logiquement **library 2.0**.

They hope that the Library 2.0 "movement" will break librarians out of brick-and-mortar establishments and get them to interact with patrons through blog comments, IM and Wiki entries.

Alors qu'est-ce que ça donne? Un des exemples est le site de la bibliothèque de la Thomas Ford Memorial library[43] qui se veut "orienté usager" et qui n'a rien de spectaculaire si ce n'est sa grande clarté et l'intégration d'un blog ainsi que la possibilité d'agrandir les caractères. Plus de détails ici[44] et là[45]. Il est également intéressant de voir combien les bibliothécaires américains commencent à intégrer les logiciels de messageries instantanées[46] comme moyen de communication avec leur usagers.

---

[34] http://bibliobsession.over-blog.com/article-1246121.html

[35] http://bibliobsession.over-blog.com/

[36] http://fr.wikipedia.org/wiki/Web_2.0

[37] http://www.fredcavazza.net/index.php?2005/11/20/951-web-20-le-putsch-des-utilisateurs

[38] http://lafeuille.blogspot.com/2005/11/bibliothques-numriques-pertinence.html

[39] http://cyberlibris.typepad.com/blog/2005/11/je_viens_de_lir.html

[40] http://www.publish.com/article2/0,1895,1881893,00.asp

[41] http://librarianinblack.typepad.com/librarianinblack/

[42] http://www.infotoday.com/il2005/

[43] http://www.fordlibrary.org/foundation/

[44] http://www.flickr.com/photos/aaronschmidt/64966024/in/photostream/

[45] http://www.walkingpaper.org/

[46] http://walkingpaper.org/181

D'autres innovations plus spectaculaires permettent d'utiliser des Tags pour visualiser des collections de bibliothèques.

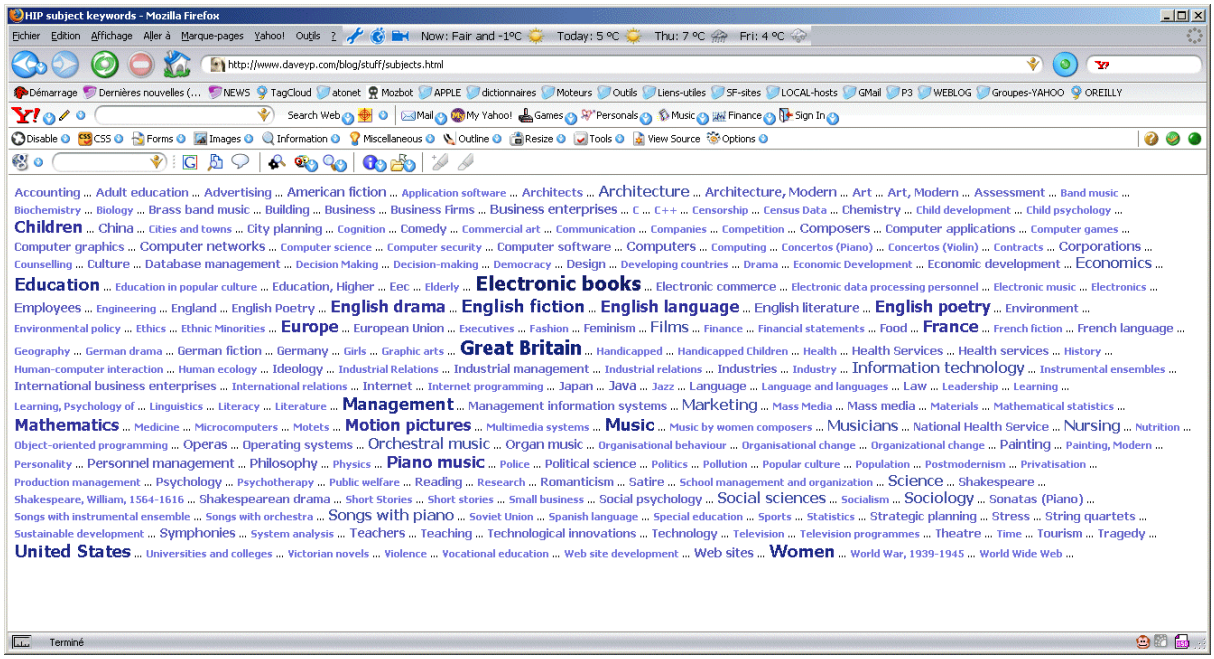


Figure 58 : des Tags pour visualiser des collections de bibliothèques

Ces tags ne relèvent pas du chaos de l'indexation "incontrôlée" (l'absence d'autorités-matières comme sur del.icio.us[47] et bien d'autres). Non ces Tags là sont une manière nouvelle de porter un regard sur la collection et sont reliés à une indexation matière tout ce qu'il y a de plus traditionnel. L'idée a été reprise à partir du travail fait par Jenny Levine[48] et son "prototype" Mock up[49]. L'auteur de cette expérimentation propose même le script[50] pour tout ceux qui ont le SIGB Horizon/HIP.

Coté innovations toujours, cet exemple d'interface d'interrogation[51] du catalogue réalisé par Casey Bisson de l'université de Plymouth (son blog ici[52]) qui fonctionne comme google suggest[53]. Toutes ces expérimentations préfigurent ce que pourront proposer les futurs bibliothèques digitales, à condition que tout cela soit accompagné d'une réflexion sur les pratiques et les usages, avant les prouesses technologiques. Pour finir, ne manquez pas cet article de Tim O'Reilly[54] et celui de Paul Miller[55] cité sur le blog Culture et TIC[56]…

47 http://del.icio.us/

48 http://www.theshiftedlibrarian.com/archives/2005/11/06/anybody_going_to_blog_these_library_20_events.html

49 http://flickr.com/photos/shifted/60728682/

50 http://www.daveyp.com/blog/index.php/archives/47/

51 http://www.plymouth.edu/library/bibinfo/suggest.html

52 http://www.maisonbisson.com/blog/

53 http://www.google.com/webhp?complete=1&hl=en

54 http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

55 http://www.ariadne.ac.uk/issue45/miller/

56 http://culturetic.canalblog.com/archives/2005/11/22/1025082.html

9.2.3 Liens et développements autour de RSS

Wiki XWiki

http://ceclines.xwiki.com/xwiki/bin/view/Main/Fils+RSS
Sur cette page de wiki, vous trouverez à peu près tout sur le RSS : La norme, comment utiliser les fils RSS, des articles, une sélection d'agrégateurs, mixer des fils RSS entre eux, mesurer l'activité des fils RSS, créer un fil RSS....

Fiche de l'ADBS consacrée au RSS

http://www.adbs.fr/site/repertoires/outils/rss.php

RSS et la publication simultanée dans Internet

http://www.culturelibre.ca/rss/

par Olivier Charbonneau

Le concept de diffusion simultanée ou syndication n'est pas nouveau. Dès le milieu du 19e siècle, certains grands quotidiens aux États-Unis employaient des mécanismes de diffusion simultanée grâce aux technologies disponibles á l'époque. Entre autres, ils offraient aux quotidiens régionaux des feuilles «pré-imprimés» où figuraient des articles plus pérennes, des annonces ainsi que des illustrations satyriques. C'est ainsi que les bandes dessinées de la grande presse Nord-Américaine pris son essor…

RSS bidirectionnel

http://affordance.typepad.com/mon_weblog/2005/11/rss_bidirection.html

(L'utilisateur peut ne plus "simplement" se contenter de recevoir des informations mais en ajouter et/ou modifier directement le flux. En anglais dans le texte cela s'appelle "*bidirectional, asynchronous replication*".)

Feed for Thought

http://www.burningdoor.com/feedburner/archives/001518.html

« *How feeds will change the way content is distributed, valued, and consumed* **»,** sur le weblog FeedBurner (Posted by Dick at November 21, 2005 10:47 AM)

9.3 Expressions régulières[57]

Les expressions régulières (ou "expressions rationnelles", en anglais "regular expression" ou "regexes", "regexp", etc.) permettent de manipuler (recherche et remplacement) du texte en utilisant des caractères spéciaux, qui valent pour une classe de cas et non littéralement. Le langage de script Perl en offre le meilleur support, mais des éditeurs de textes courants permettent également de les utiliser.

Les regexes ne sont cependant pas un langage : elles ne peuvent pas faire de boucle, ou agir de façon conditionnelle. Le motif est générique mais statique. Elles sont surtout utiles dans des cas prévisibles et normalisés : paragraphe, ponctuation, encodage...

Il existe de nombreuses excellentes introductions aux expressions régulières et des documentations dans les logiciels qui les implémentent, dont une sélection est proposée ci-dessous ; on se limite ici à des introductions puis à des exemples avec trois environnements : Emacs (Windows / Linux), Textpad (Windows), et Perl. La syntaxe des expressions régulières peut varier très légèrement d'un éditeur ou d'un programme à l'autre, mais les fonctions restent les mêmes. Word n'est pas un éditeur de texte mais implémente un sous-ensemble des expressions régulières.

9.3.1 Liens et tutoriaux

Perl

Pour commencer : « Perl regular expressions quick start » :

http://Perldoc.com/Perl5.8.0/pod/Perlrequick.html

Le plus complet pour utiliser les regexes avec Perl :

"Perl regular expressions tutorial" :

http://Perldoc.com/Perl5.8.0/pod/Perlretut.html

Pour les références complètes, la page de manuel :

http://www.Perldoc.com/Perl5.8.0/pod/Perlre.html

"Perl Regular Expressions FAQ" :

http://Perldoc.com/Perl5.8.0/pod/Perlfaq6.html

Les chapitres "Expressions rationnelles" et "recherches de correspondances" de Programmation en Perl, Larry Wall et al., O'Reilly.

Emacs

La page d'aide d'Emacs « Syntax of Regular Expressions » est une très claire introduction

Traduite en français :

http://www.linux-france.org/article/appli/Emacs/manuel/html/regexps.html

Textpad

L'aide du logiciel est satisfaisante si l'on a déjà une connaissance du principe des regexes.

Tutoriaux généraux

regexes orientées textes électroniques :

http://etext.lib.virginia.edu/helpsheets/regex.html

---

[57] Ce texte est extrait d'une page écrite sur le web par sloiseau@u-paris10.fr (27 Octobre 2003)

Regexes sur une page :

http://www.ciser.cornell.edu/info/regex.html

9.3.2 Présentation générale

Une expression régulière est un motif qui peut correspondre à une classe de chaîne de caractères et pas seulement à une chaîne de caractères littéralement. C'est elles que l'on utilise par exemple pour exprimer "tout fichier finissant par doc" avec *.doc.

Pour exprimer ces éléments génériques certains caractères sont réservés. (le * ci-dessus exprime ici "toute chaîne de n'importe quelle longueur", non un astérisque)

Les éléments réservés doivent être précédés d'un signe spécial pour être employés dans leur sens littéral. Par exemple, Pour rechercher un astérisque, il faut l'exprimer *.

"\" est donc lui-même un caractère réservé, servant à exprimer le statut littéral ou pas de ce qui suit. Pour le chercher littéralement, il faut donc utiliser \\.

Les expressions régulières contiennent notamment des éléments pour exprimer, (1) des caractères littéralement, (2) des classes de caractères, (3) des quantificateurs qui peuvent qualifier les deux premiers groupes, et (4) des assertions positives ou négatives. Exemple pour chacune des catégories :

Tous les caractères sauf les caractères réservés représentent eux-mêmes, notamment les caractères alphanumériques. Les caractères sont a priori sensibles à la casse. L'expression Abc recherche littéralement la chaîne "Abc".

Des caractères entre crochets droits expriment une liste de choix : [aeiouy] = "un caractère parmi les 6 caractères a e i o u ou y". a vaut bien toujours littéralement pour "a", mais dans une classe constituée également d'autres éléments en relation d'alternative. [aeiouy]b trouvera donc "ab", "ib", mais pas "cb". Pour éviter des listes de caractères trop longues, on peut utiliser un tiret : [a-z] = tout caractères entre "a" et "z" (tout alphabétique minuscule), [0-9] tout numérique, [A-Za-z0-9] tout alphanumérique. Ces listes sont cumulables : [0-35-9] recherche tout numérique sauf "4".

Les quantificateurs s'appliquent à une classe ou une chaîne littérale. Par exemple [aeiouy]+ signifie "une ou plusieurs voyelles de suite", [aeiouy]* zéro ou plusieurs, [aeiouy]? zéro ou une ; [aeiouy]{5} signifie "exactement 5 voyelles de suite". Ainsi [aeiouy]+c peut correspondre à "aiec", "ac", mais pas à "adc".

Des assertions peuvent modifier des classes ou ancrer des chaînes. Par exemple le circonflexe permet de nier le contenu entre crochets droits : [^aeiouy]+ signifie "un ou plusieurs caractères de suite qui n'est pas a e i o u ou y". [ae]$ signifie un a ou un e (minuscule) en fin de paragraphe (juste avant un retour chariot).

9.3.3 Présentation complémentaire

1. *Recherche littérale* — Les expressions régulières permettent de rechercher des caractères littéralement : les caractères sont désignés par le symbole correspondant, a recherche le caractère "a", etc. Des caractères "non visibles" comme le retour chariot, la tabulation, etc. peuvent être désignés littéralement par un caractère spécial : \n pour le retour chariot et \t pour la tabulation. L'anti-slash forme un seul symbole avec le caractère qui suit, qui ne vaut plus littéralement pour un "n" ou un "t". L'anti-slash lui-même n'est pas l'équivalent littéral du caractère "\". Pour désigner un anti-slash littéralement, il faut le faire précéder lui même d'un anti-slash : \\ recherche "\". Les "méta-caractères" de cette sorte sont généralement \, ., *, ?, +, ^, $, {, |, (,), et }, dont on

voit ci-dessous le sens spécial ; pour les utiliser littéralement il faut donc utiliser \\, \., *, \?, \+, \^, \$, \{, \|, \(, et \). L'anti-slash permet d'interpréter un méta-caractère comme un caractère littéral, mais également de donner un sens spécial à des caractères littéraux, comme \n ou \t.

2. *Classes d'éléments* — Les expressions régulières permettent également de ne pas faire correspondre les élément un à un, mais de faire correspondre une liste de possibilités à un élément.

a/ Une classe de caractères en alternative est désignée entre crochets : [aeiouy] recherche une occurrence d'une voyelle : elle trouvera deux éléments dans "soir". — Pour ne pas écrire de longue suite de caractères, on peut spécifier des plages avec le tiret : [0-9] recherche tout caractère numérique, [a-z] toute lettre en minuscule, [A-Za-z0-9] recherche tout caractère alphanumérique. Il faut noter que les plages de caractères sont basées sur l'ASCII, il faut donc ajouter les caractères qui n'en font pas partie, comme les caractères accentués : [a-zéèàùâêîôûäëïöüçA-ZÉÈÀÙÂÊÎÔÛÄËÏÖÜÇ0-9] pour être (presque) exhaustif. Mais, par chance, on peut alors utiliser le point . qui vaut pour tout élément sauf le retour chariot (\n). [aeiouy]. trouvera "az", "aa", "a!" mais pas "a" ni "ba".
Pour mettre un tiret littéralement parmi les éléments en alternative, il faut le mettre en premier dans la liste : [-,;] recherche un tiret, une virgule ou un point virgule (pour "]" utiliser l'antislash : [-\]]). — Les méta-caractères perdent leur sens particulier dans une expression entre crochets.

b/ Des symboles spéciaux permettent de désigner les classes les plus souvent utilisées, par exemple en Perl \w = *w*ord, tout caractère de mot (alphanumériques plus '_'), \d = *d*igit ([0-9]), \s tout caractère blanc (= [\t\n\r\f], soit espace, tabulation, saut de ligne, retour chariot, saut de page) ; lesquels peuvent être inversés \W = tout caractère non-mot, \D = tout non numérique, \S tout caractère autre que blanc, etc. (cf la page de manuel Perl : http://www.Perldoc.com/Perl5.8.0/pod/Perlre.html#Regular-Expressions) — Autres familles de classes, du type : [:alnum:] par exemple pour tout *al*pha*num*érique ([A-Za-z0-9])

c/ L'alternative entre des éléments de plusieurs caractères est exprimée par |, et s'applique aux éléments entre parenthèses : (oui|non) permet de rechercher "oui" ou "non". Avec les crochets, le contenu est évalué comme une liste de caractères, ainsi l'alternative | est inutile : [truc|tric] = [truci] — le caractère ^ en début de plage permet d'indiquer une assertion négative : [^a-z] recherche tout ce qui n'est pas un alphabétique minuscule.

3. *Quantifieurs* — Possibilité de préciser le nombre d'occurrences d'un élément (un caractère littéral, ou une classe d'éléments en alternative, ou une chaîne de caractères entre parenthèses). Les principaux quantificateurs sont ? (0 ou 1 fois), + (1 ou plusieurs fois), * (0 ou plusieurs fois). On peut également spécifier une plage d'occurrence : [aeiouy]{3,5} désigne toute suite de entre 3 et 5 voyelles. Si le second caractère n'est pas spécifié, il vaut pour l'infini : [0-9]{5,} recherche au moins 5 numériques de suite. On voit que ? = {0,1}, + = {1,} et * = {0,}. Une valeur seule entre accolades exige une valeur exacte : a{5} recherche exactement 5 "a" de suite.
Les quantificateurs s'appliquent à la dernière unité syntaxique les précédents. non{5} trouvera "nonnnnn" mais non pas "nonnonnonnonnon" pour cela il faut écrire (non){5}
Note : Les expressions rationnelles sont par défaut avides : dans "abbb", ab+ trouvera "abbb" et non "ab". Si on veut éviter ce comportement, on peut utiliser la négation : <[^>]+> par exemple pour ne pas prendre le début d'une balise et la fin de la dernière balise sur la même ligne. Avec Perl, on peut également préciser au quantificateur de chercher la séquence minimale en ajoutant

"?" : +?, *? et {1,5}? cherche donc la première correspondance. D'autre part une expression rationnelle cherche l'expression dès que possible, et s'arrête dès que les conditions sont satisfaites même si une correspondance plus loin serait plus complète. Aussi y* trouve un motif vrai dès le premier caractère de "xyz", puisqu'il peut correspondre à une chaîne nulle.

4. *Référence* — Les parenthèses permettent de hiérarchiser les membres d'une alternative (cf. 2), et de regrouper une chaîne de caractères en un élément (cf. 3), mais servent également à mettre en mémoire le résultat. Ce qui permet par la suite de faire référence au contenu d'une expression entre parenthèses à l'aide de \1 (pour désigner le contenu d'une première parenthèse dans l'expression de recherche), \2 (pour désigner le contenu d'une seconde), etc. Ainsi ([^]+) \1 recherche un mot répété : un blanc, toute suite d'au moins un caractère qui n'est pas un blanc, un blanc, et la même suite de caractères. Les références de cette sorte sont également utiles pour remplacer : faire référence dans l'expression de remplacement au contenu trouvé dans l'expression de recherche : remplacer ([0-9]+)([a-zA-Z]+) par \1 \2 permet par exemple d'ajouter un espace entre une suite de numériques et une suite d'alphabétiques. Remplacer ([a-z])([;:\?!]) par \1 \2 ; permet d'ajouter un espace entre la fin d'un mot et les signes de ponctuation qui demandent cet espace. (anti-slash avant le ? pour lui faire rechercher le caractère "?", puisque c'est un caractère réservé).

5. *Assertion* — De nombreuses expressions permettent d'inclure des assertions, qui ne correspondent pas à un élément à rechercher mais indiquent par exemple que l'expression commence en début de ligne (^) ou en fin de ligne ($), etc. Ces assertions sont la partie la plus variable des expressions régulières entre les différents logiciels, il faut se reporter à la documentation correspondante.