TypWeb : profilage de sites Web

Chaîne Typweb 038
Les outils

Equipe Typweb

Valérie Beaudouin[(*)], Serge Fleury[(***)], Benoît Habert[(**,****)],
Gabriel Illouz[(****)], Nicolas Renet [(****)], Helka Folch [(****)],
Cyril Grouin[(****)], Christian Licoppe[(*)], Marie Pasquier[(*)]

**France Télécom R&D [(*)], Université Paris X [(**)],
Université Paris 3 (CLA2T) [(***)], LIMSI [(****)]**

Sommaire

Table des figures

1 Profilage de sites Web

Nous appelons profilage de sites Web l'utilisation d'outils de calibrage donnant des indications sur les contenus et sur les structures de ces sites. Ces outils doivent également permettre de positionner un nouveau site par rapport aux regroupements obtenus sur un base de sites déjà analysé. Ils doivent aussi permettre de mesurer les évolutions des sites. Le projet TypWeb qui associe des chercheurs de France Telecom R&D (DIH/UCE) et de l'équipe TyPText (LIMSI - PARIS X - PARIS 3) propose de fournir un cadre méthodologique et pratique de profilage de sites Web et une typologie fine de ces sites. TyPText développe une architecture de profilage dans une optique inductive qui consiste à faire émerger des textes, considérés comme des agglomérats fonctionnellement cohérents de traits de niveau variés (linguistiques, structurels, typographiques …).

Ce projet vise à décrire l'articulation entre la description formelle et sémantique des sites avec les récits des pratiques des acteurs (concepteurs et visiteurs) : ce dernier point prend appui sur les entretiens réalisés auprès des concepteurs et des visiteurs des sites étudiés et il se déroule en liaison avec l'étude menée dans le cadre d' un projet axé sur l' analyse des parcours de sites. L'examen des sites et de leurs évolutions doit aussi permettre d'étudier les tendances existantes ou à venir dans la mise en œuvre de sites Web : Le mimétisme est-il la règle générale pour la construction de sites Web ? Comment la forme d'un site et son contenu sont-ils cohérents avec le projet de leur auteur ?

La démarche suivie vise à caractériser chaque site par des indicateurs de contenu et de structure qui doivent permettre d'analyser les conditions d'une éventuelle proposition de norme dans cette mise en place progressive des échanges sur l'hypertoile. Les indications permettant de décrire les sites sont enrichies de manière inductive sur la base des premiers résultats produits. Elles sont ensuite utilisées pour caractériser et enrichir la description de nouveaux sites. Ces informations alimentent le « cartouche » descriptif de chaque site analysé : ce cartouche étant conçu pour rester ouvert à toute nouvelle information pertinente capable de l'enrichir.

2 Guide d'utilisation des outils Typweb

2.1 Outils disponibles

Les outils de la chaîne Typweb sont disponibles à l'adresse suivante :

http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/typweb/tools/toolstypweb038.zip

Cette archive contient les programmes suivants :

`webxref-038.pl`	Le programme webxref est le point d'entrée de la chaîne Typweb, il prend en entrée des sites web préalablement aspirés et produit des versions normalisées de ces sites
`webxref-038-homolSF2.pl`	Une seconde version du même programme en cours d'homologation
`countLinks-038.pl`	Le programme prend en entrée un corpus XML construit par webxref-038 et produit en sortie un état statistique d'un certain nombre de lien (interne, externe (http,ftp), et mailto) référencé dans le corpus XML.
`countTagWindow-038.pl`	Le programme prend en entrée un corpus XML produit par webxref et une valeur numérique correspondant à une longueur de TAGs consécutifs, il produit en sortie un fichier contenant les TAGs consécutifs sur l'ensemble du fichier initial
`makeCorpusSelectedTAGForLexico4.pl`	Ce programme prend en entrée entrée un corpus XML produit par webxref et une liste de TAGs HTML sélectionnés, il produit en sortie un corpus de TAGs HTML correspondant à la liste donnée.
`makeCorpusTAGForLexico3.pl`	Ce programme prend en entrée entrée un corpus XML produit par webxref, il produit en sortie un corpus de TAGs HTML prêt pour être analysé par Lexico3
`makeListTagForLexico.pl`	Ce programme établit la liste des TAGs HTML contenu dans le corpus XML produit par webxref.
`makeMatriceLinks-038-pajek.pl`	le programme prend en entrée un corpus XML construit par webxref-038, et produit en sortie des informations sur un certain nombre de lien (interne, externe (http,ftp), et mailto) référencé dans le corpus XML, ces informations sont formatées pour être utilisées par l'outil de représentation de réseaux pajek.
`MakeMatriceLinks-038.pl`	le programme prend en entrée un corpus XML construit par webxref-038, et produit en sortie un tableau d'un certain nombre de lien (interne, externe (http,ftp), et mailto) référencé dans le corpus XML.

`Transcodeur.pl`
`TranscodeurWithLog.pl`
`TranscodeurWithNoLog.pl`
`TableCharacter.txt`
	Ce programme est un programme de transcodage permettant de nettoyer éventuellement les corpus de travail.

Table 1 : Programmes de l'archive ToolsTypweb-038

2.2 Mode d'emploi et syntaxe

2.2.1 Lancement de webxref

Syntaxe : `perl webxref-038.pl -rappspec 0 -html -del -lynx -at Rapports/ sitesWeb1/`

- L'option `-rappspec` suivie d'un chiffre permet de numéroter les répertoires de sortie contenant chacun les rapports d'analyse sur un site.
- L'option `-html` est nécessaire pour le traitement des sites dans le processus Typweb
- L'option `-del` permet de supprimer les rapports html produit en sortie : on ne conserve que le corpus XML d'un site analysé et les états statistiques

- L'option `-lynx` permet de générer des états textuels via Lynx sous Unix
- L'option `-at` permet de rediriger les sorties vers le répertoire donné en argument
- Le dernier élément est le nom du site à analyser

2.2.2 Préparation des matrices de liens

Syntaxe :

```
perl makeMatriceLinks-038.pl corpusTypweb.xml
ou
perl makeMatriceLinks-038-pajek.pl corpusTypweb.xml
```

2.2.3 Préparation des corpus de TAGs

Syntaxe :

```
perl makeCorpusSelectedTAGForLexico4.pl corpusTypweb.xml listeTAG.txt

ou :
```

makeCorpusTAGForLexico3.pl `corpusTypweb.xml`

On trouvera infra d'autres exemples d'utilisation de ces outils.

3 Les outils de la chaîne Typweb : descriptif complet

3.1 Préambule/ Rappel

Deux chaînes de traitements (notées 036 et 038) ont été mises en œuvre. Ces notations sont liées "historiquement" à `webxref`.

1. La version initiale que Calin Mosut a modifiée était la 035.
2. Celle résultant des modifications de Calin est la 036 : cette version permet en particulier la génération de rapports pour chaque page d'un site. A cette version de webxref, on a associé deux programmes (`mktipo` et `ExtAndStat`) chargés de produire un corpus XML et des statistiques élémentaires eux aussi noté 036 (cf programmes présentés *supra*).Globalement, la chaîne 036 correspond à la présentation faite *supra*.
3. La version 038 correspond à l'intégration de nouvelles fonctionnalités à le 036

- les programmes de génération du corpus XML et des statistiques ont été insérés dans le programme `webxref`,
- prise en compte des attributs dans le corpus XML créé et dans les statistiques.

Cette nouvelle version se distingue par rapport à la précédente par la prise en compte des attributs associés aux tags HTML. Dans le fichier de statistiques des éléments présents dans chaque page de ce site, on trouve donc ce décompte supplémentaire (cf exemple supra).

Le schéma de la figure suivante donne l'allure générale de l'architecture mise en œuvre pour le traitement des sites du projet Typweb et les programmes associés. Certains programmes sont présentés infra.

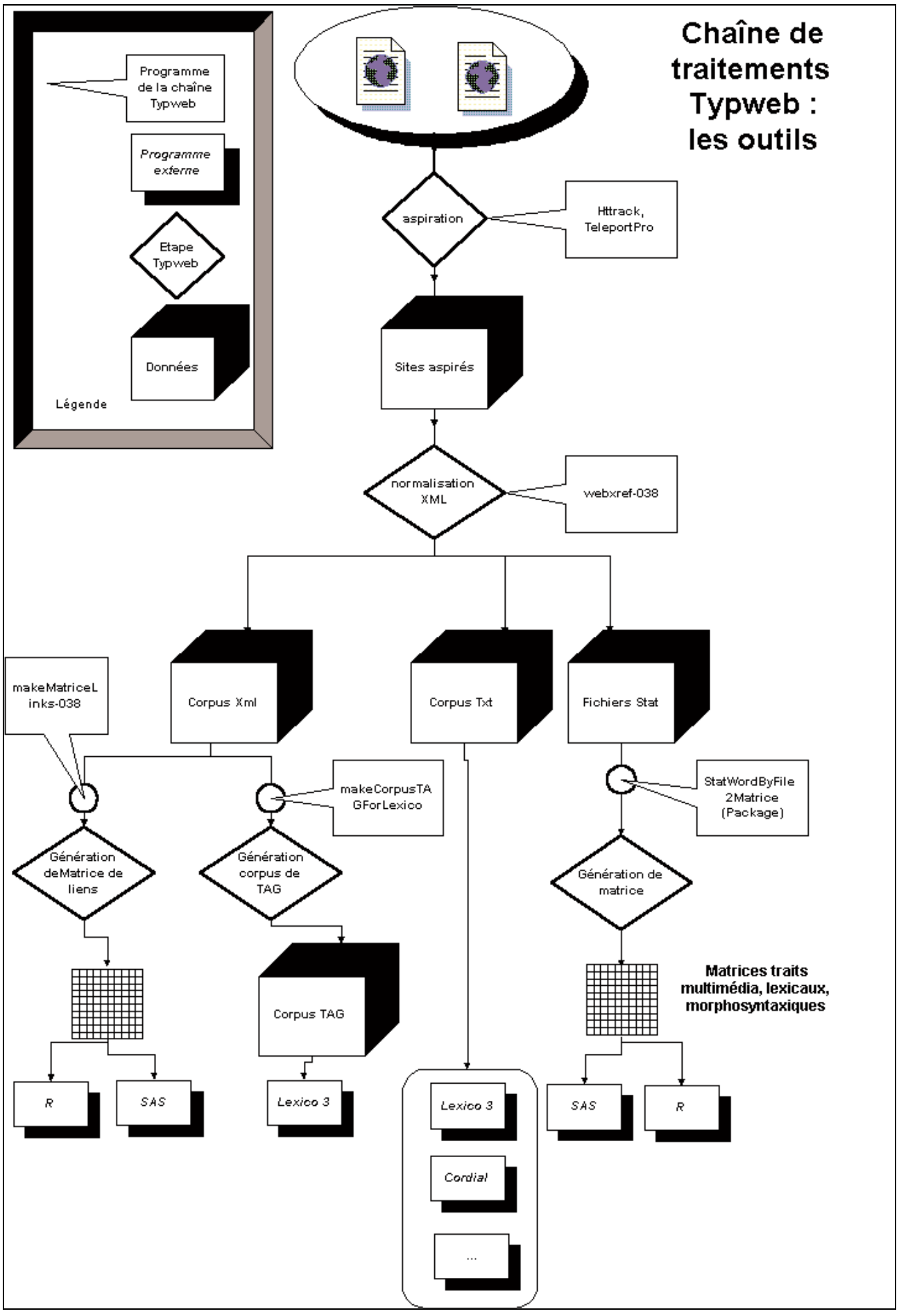


Figure 1 : Chaîne des outils Typweb

3.2 Webxref-038

Le programme `webxref-038` construit en sortie :

- un corpus XML
- un corpus textuel correspondant aux zones textuelles des pages HTML scrutées
- un état statististique pour chaque page du site examiné
- un état statististique global du site

Cette nouvelle version prend en compte le traitement des attributs des éléments HTML rencontrés dans les pages scrutées. Les éléments pris en compte dans ces états statistiques ont été balisés pour permettre ,en aval, de filtrer facilement ces informations.

3.2.1 Schéma du corpus XML 038

Le schéma du corpus relatif à cette version du programme est le suivant :

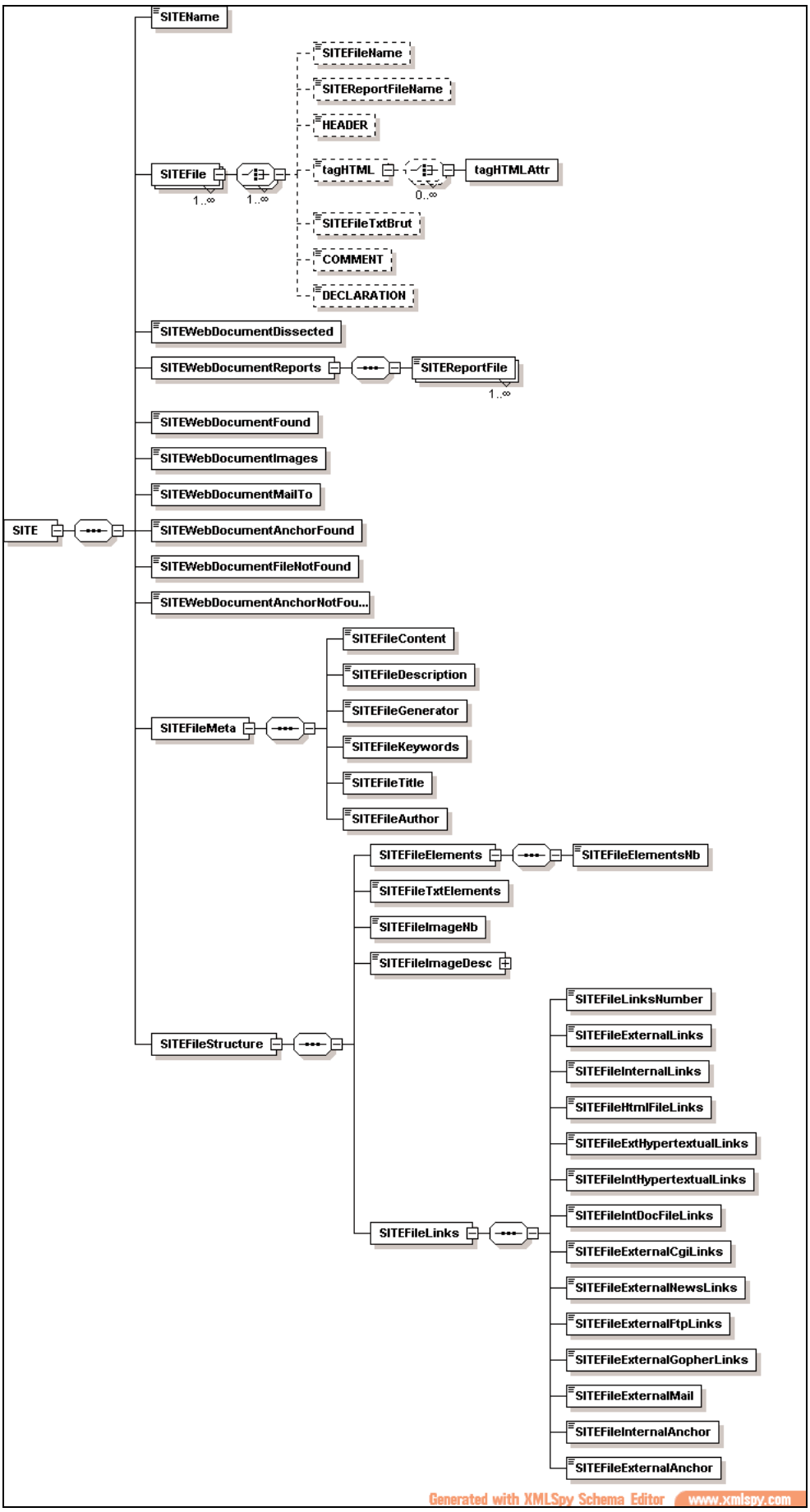


Figure 2 : schéma d'un corpus Typweb (un site)

3.2.2 Fonctionnalités supplémentaires

Le programme `webxref-038` permet un meilleur traitement en amont des parties textuelles des sites web analysés. Les modifications présentées ci-dessous ne sont pas toutes disponibles dans la version 038 de webxref (sous Windows) : des indications supplémentaires sont ajoutées pour préciser la disponibilité de ces nouveautés.

1. Réécriture du parsage des éléments textuels : modification des procédures de nettoyage des données textuelles lues dans les pages HTML scrutées : modifications validées par Calin Mosut et Serge Fleury. (modification disponible dans la version webxref-038-homologation)

2. Intégration de calculs statistiques supplémentaires : statistiques sur des suites d'éléments HTML (succession linéaires de balises, succession linéaire d'attributs, successions linéaires de couples attribut-valeur) autour des zones textuelles. (modification disponible dans la version webxref-038-homologation)

3. Intégration du navigateur lynx pour récupérer une version textuelle "propre" des pages scrutées. (modification disponible dans la version webxref-038 sous Unix)

Une option de webxref, disponible pour le moment dans un environnement Unix, permet de récupérer ces parties textuelles via l'utilisation de lynx avec l'option -dump sur chaque page du site analysé. Le résultat de cette option est l'intégration d'un nouveau nœud dans l'arbre XML construit par webxref. Ce nœud contient pour chaque page le contenu textuel de cette page. On donne ci-dessous le contenu de cette zone du corpus XML produit sur le site démo :

```
<DUMPLYNX>
<FILE>
<FILENAME>/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/siteDemo/index.htm</FILENAME>
<DUMPTEXT>

    Welcome...

    [1]Page 1 [2]Page 2

    [3]Page 3

                                                          ...Bon surf !!

 References

    1.
file://localhost/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/siteDemo/page1.htm
    2.      file://localhost/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/siteDemo/ss-
dossier/page2.htm
    3.
file://localhost/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/siteDemo/page3.htm

</DUMPTEXT>
</FILE>
…
</DUMPLYNX>
```

L'arbre XML du corpus produit a l'allure suivante :

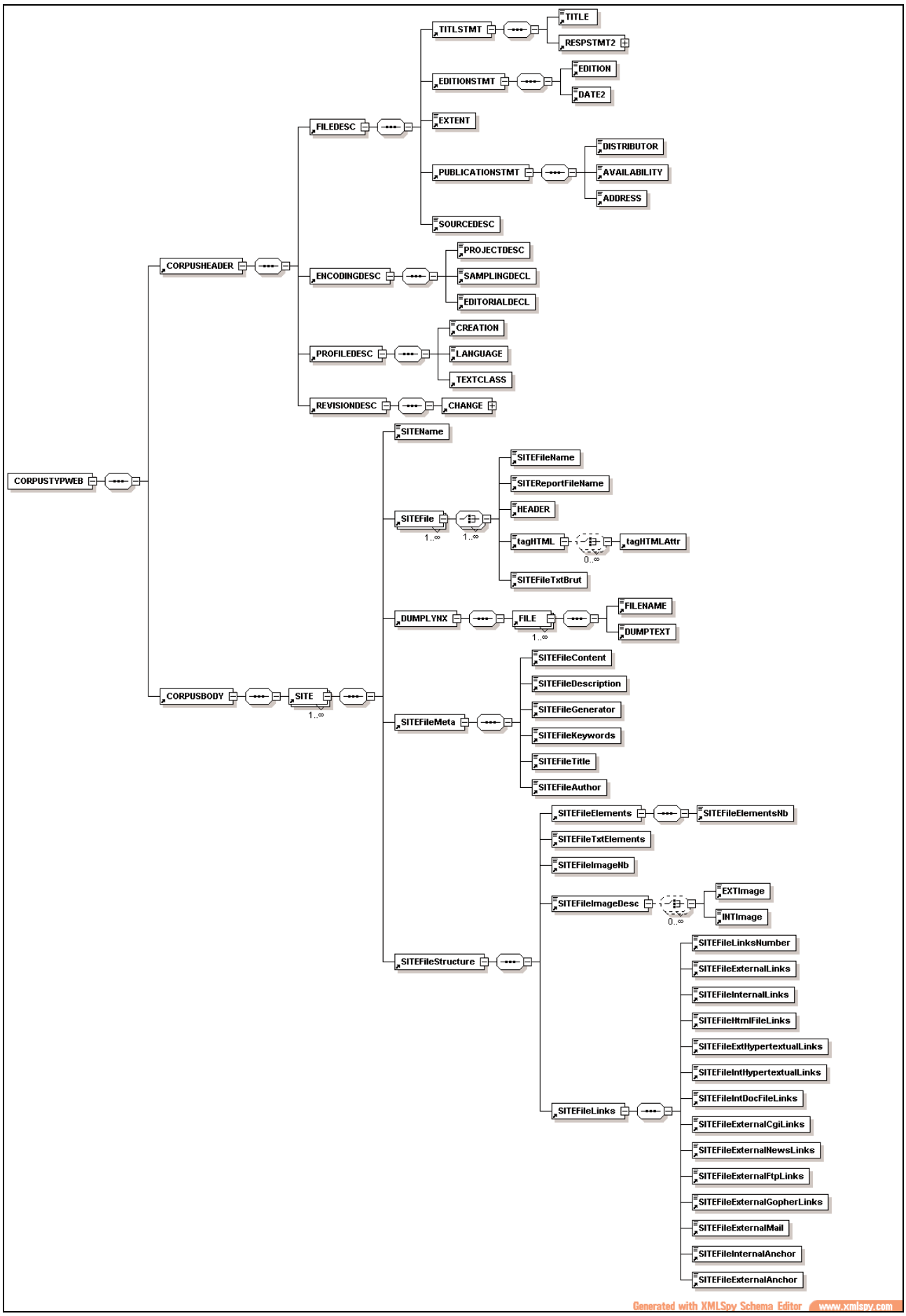


Figure 3 : schéma d'un corpus Typweb avec lynx

3.3 Constitution des corpus d'analyse

3.3.1 countLink-038.pl

- Format d'entrée :

- le programme prend en entrée un corpus XML construit par webxref-038.

- Format de sortie :

- le programme cree en sortie un etat statistique d'un certain nombre de lien (interne, externe (http,ftp), et mailto) référencé dans le corpus XML. Tous les resultats sont balisés de la manière suivante :

```
(<SITE>
<NAME>nom du site</NAME>


(<LINKSBYPAGE>
<PAGE>nom de la page</PAGE>
(1)
</LINKSBYPAGE>)+
....

<LINKSBYSITE>
(2)
</LINKSBYSITE>
</SITE>)+
```

Dans (1) on trouve un état pour les liens actuellement regardés de la page, cet état se décompose ainsi :

```
<INTERNALLINK>
(<LINK>le nom du lien ex: toto.html</LINK><COUNT>le nombre de ce type de lien</COUNT>)+
<TOTALLINKS>nb total de liens internes différents</TOTALLINKS>
<TOTALOCCUR>nb total d'occurrences de liens internes</TOTALOCCUR>
</INTERNALLINK>
<EXTERNALHTTPLINK>
(<LINK>le nom du lien ex: http://toto.html</LINK><COUNT>le nombre de ce type de lien</COUNT>)+
<TOTALLINKS>nb total de liens http-externes différents</TOTALLINKS>
<TOTALOCCUR>nb total d'occurrences de liens http-externes</TOTALOCCUR>
</INTERNALHTTPLINK>
<EXTERNALFTPLINK>
(<LINK>le nom du lien ex: ftp://toto.html</LINK><COUNT>le nombre de ce type de lien</COUNT>)+
<TOTALLINKS>nb total de liens ftp-externes différents</TOTALLINKS>
<TOTALOCCUR>nb total d'occurrences de liens ftp-externes</TOTALOCCUR>
</INTERNALFTPLINK>
<MAILTOLINK>
(<LINK>le nom du lien ex: mailto:monsieurToto</LINK><COUNT>le nombre de ce type de lien</COUNT>)+
<TOTALLINKS>nb total de liens mailto différents</TOTALLINKS>
<TOTALOCCUR>nb total d'occurrences de liens mailto</TOTALOCCUR>
</MAILTOLINK>
<COUNTLINKBYPAGE>
<TOTALLINKS>nombre total de liens différents pour toute la page</TOTALLINKS>
<TOTALOCCUR>nombre total d'occurrences de tous les liens de cette page</TOTALOCCUR>
</COUNTLINKBYPAGE>
```

les séquences <LINKSBYPAGE>...</LINKSBYPAGE> définies pour chaque page s'empilent les unes derrière les autres.

Pour (2), c'est la même chose mais pour tout le site :

comptage des liens internes, http-externe, ftp-externe et mailto sur l'ensemble du site puis sommation du tout.
Avec le même type de balises mais dans le contexte <LINKSBYSITE>

3.3.2 makeMatriceLink-038.pl

Ce programme se comporte comme le précédent mais produit en sortie une matrice décrivant pour chaque page du corpus de site choisi, un état des liens scrutés, la structure de chaque ligne est la suivante :

Site	Page	Lien Interne	Fréq	Lien Externe	Fréq	Lien Mail	Fréq	Lien Ftp	Fréq

Exemple de sorties produites sur le site Démo :

```
<SITE>
<NAME>siteDemo</NAME>
<LINKSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/ind
ex.htm</PAGE>
<INTERNALLINK>
<LINK>SS-
DOSSIER/PAGE2.HTM</LINK><COUNT>1</COUNT>
<LINK>PAGE1.HTM</LINK><COUNT>1</COUNT>
<LINK>PAGE3.HTM</LINK><COUNT>1</COUNT>
<TOTALLINKS>3</TOTALLINKS>
<TOTALOCCUR>3</TOTALOCCUR>
</INTERNALLINK>
<EXTERNALHTTPLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</EXTERNALHTTPLINK>
<EXTERNALFTPLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</EXTERNALFTPLINK>
<MAILTOLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</MAILTOLINK>
<COUNTLINKBYPAGE>
<TOTALLINKS>3</TOTALLINKS>
<TOTALOCCUR>3</TOTALOCCUR>
</COUNTLINKBYPAGE>
</LINKSBYPAGE>
<LINKSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/ss-
dossier/page2.htm</PAGE>
<INTERNALLINK>
<LINK>../PAGE1.HTM</LINK><COUNT>1</COUNT>
<TOTALLINKS>1</TOTALLINKS>
<TOTALOCCUR>1</TOTALOCCUR>
</INTERNALLINK>
<EXTERNALHTTPLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</EXTERNALHTTPLINK>
<EXTERNALFTPLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</EXTERNALFTPLINK>
<MAILTOLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</MAILTOLINK>
<COUNTLINKBYPAGE>
<TOTALLINKS>1</TOTALLINKS>
<TOTALOCCUR>1</TOTALOCCUR>
</COUNTLINKBYPAGE>
</LINKSBYPAGE>
<LINKSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/pag
e1.htm</PAGE>
<INTERNALLINK>
<LINK>PAGE3.HTM</LINK><COUNT>1</COUNT>
<TOTALLINKS>1</TOTALLINKS>
<TOTALOCCUR>1</TOTALOCCUR>
</INTERNALLINK>
<EXTERNALHTTPLINK>
<LINK>HTTP://WWW.NETSCAPE.FR/</LINK><COUNT>1</
COUNT>
<LINK>HTTP://WWW.MICROSOFT.COM/FRANCE/</LINK><
COUNT>1</COUNT>
<LINK>HTTP://WWW.MICROSOIF.COM/</LINK><COUNT>1
</COUNT>
<TOTALLINKS>3</TOTALLINKS>
<TOTALOCCUR>3</TOTALOCCUR>
</EXTERNALHTTPLINK>
<EXTERNALFTPLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</EXTERNALFTPLINK>
<MAILTOLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</MAILTOLINK>
<COUNTLINKBYPAGE>
<TOTALLINKS>4</TOTALLINKS>
<TOTALOCCUR>4</TOTALOCCUR>
</COUNTLINKBYPAGE>
</LINKSBYPAGE>
<LINKSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/pag
e3.htm</PAGE>
<INTERNALLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</INTERNALLINK>
<EXTERNALHTTPLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</EXTERNALHTTPLINK>
<EXTERNALFTPLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</EXTERNALFTPLINK>
<MAILTOLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</MAILTOLINK>
<COUNTLINKBYPAGE>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</COUNTLINKBYPAGE>
</LINKSBYPAGE>
<LINKSBYSITE>
<INTERNALLINK>
<LINK>PAGE3.HTM</LINK><COUNT>2</COUNT>
<LINK>../PAGE1.HTM</LINK><COUNT>1</COUNT>
<LINK>SS-
DOSSIER/PAGE2.HTM</LINK><COUNT>1</COUNT>
<LINK>PAGE1.HTM</LINK><COUNT>1</COUNT>
<TOTALLINKS>4</TOTALLINKS>
<TOTALOCCUR>5</TOTALOCCUR>
</INTERNALLINK>
<EXTERNALHTTPLINK>
<LINK>HTTP://WWW.NETSCAPE.FR/</LINK><COUNT>1</
COUNT>
<LINK>HTTP://WWW.MICROSOFT.COM/FRANCE/</LINK><
COUNT>1</COUNT>
<LINK>HTTP://WWW.MICROSOIF.COM/</LINK><COUNT>1
</COUNT>
<TOTALLINKS>3</TOTALLINKS>
<TOTALOCCUR>3</TOTALOCCUR>
</EXTERNALHTTPLINK>
<EXTERNALFTPLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</EXTERNALFTPLINK>
<MAILTOLINK>
<TOTALLINKS>0</TOTALLINKS>
<TOTALOCCUR>0</TOTALOCCUR>
</MAILTOLINK>
<COUNTLINKBYSITE>
<TOTALLINKS>7</TOTALLINKS>
<TOTALOCCUR>8</TOTALOCCUR>
</COUNTLINKBYSITE>
</LINKSBYSITE>
</SITE>
```

3.3.3 makeMatriceLink-038-pajek.pl

Ce programme est une adaptation du programme précédent qui produit en sortie des données formatées pour l'outil de représentation graphique de réseaux PAJEK .

On présente ci-dessous les résultats produits par ce programme sur le site demo :
Sortie n°1 : réseau global

```
*Vertices     5
1  siteDemo2
2 http://www.netscape.com/page1.htm
3 http://www.netscape.fr/
4 http://www.microsoif.com/
5 http://www.microsoft.com/france/
*Arcs
1 2
1 3
1 4
1 5
```

Sortie n°2 : réseau à partir de la page d'accueil

```
*Vertices     1
1  siteDemo2
*Arcs
```

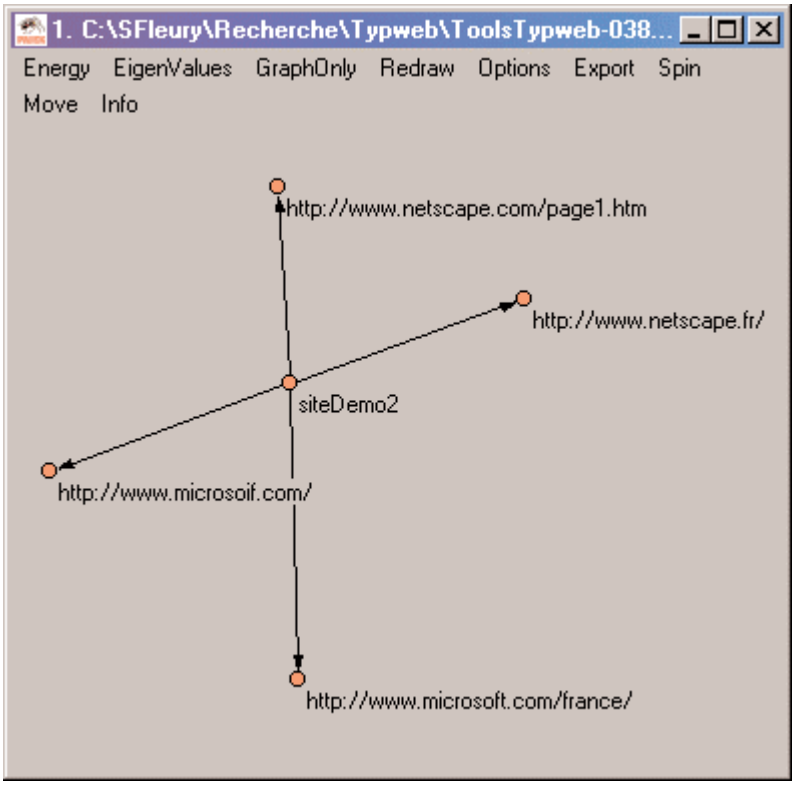


Figure 4 : réseau de lien avec Pajek

3.3.4 countTagWindow-038.pl

- Format d'entrée :

- le programme prend en entrée un corpus XML construit par webxref-038 et une valeur numérique correspondant à une fenêtre de TAG Html à scruter. Ce programme vise en fait à rechercher tous les segments répétés de TAG HTML sur la longueur demandée.

- Format de sortie :

- le programme crée en sortie un état des segments répétés de TAG HTML. Ces résultats sont produits pour chaque page du site scuté et pour l'ensemble du site.

Exemple de sorties produites sur le site Démo :

```
<SITE>
<NAME>siteDemo</NAME>
<TAGSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/index.htm</PAGE>
<WINDOWTAG LENGHT="3">meta-meta-meta</WINDOWTAG><COUNT>3</COUNT>
<WINDOWTAG LENGHT="3">font-table-tr</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">table-tr-td</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">title-body-p</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">body-p-font</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-table</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">br-p-font</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">td-a-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-td-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">td-a-td</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">head-meta-meta</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-a-br</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">meta-title-body</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">tr-td-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">meta-meta-title</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">html-head-meta</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-br-p</WINDOWTAG><COUNT>1</COUNT>
</TAGSBYPAGE>
<TAGSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/ss-dossier/page2.htm</PAGE>
<WINDOWTAG LENGHT="3">meta-meta-meta</WINDOWTAG><COUNT>3</COUNT>
<WINDOWTAG LENGHT="3">br-p-font</WINDOWTAG><COUNT>2</COUNT>
<WINDOWTAG LENGHT="3">font-html-head</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">body-p-font</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">title-body-p</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">head-meta-meta</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">meta-title-body</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">font-a-br</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">meta-meta-title</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">html-head-meta</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-br-p</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-html</WINDOWTAG><COUNT>1</COUNT>
</TAGSBYPAGE>
<TAGSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/page1.htm</PAGE>
<WINDOWTAG LENGHT="3">meta-meta-meta</WINDOWTAG><COUNT>3</COUNT>
<WINDOWTAG LENGHT="3">br-a-br</WINDOWTAG><COUNT>2</COUNT>
<WINDOWTAG LENGHT="3">br-p-font</WINDOWTAG><COUNT>2</COUNT>
<WINDOWTAG LENGHT="3">img-br-a</WINDOWTAG><COUNT>2</COUNT>
<WINDOWTAG LENGHT="3">br-a-img</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">font-html-head</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-img-br</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">body-p-font</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">title-body-p</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-br-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-br-img</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">head-meta-meta</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">meta-title-body</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">font-a-br</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">meta-meta-title</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">html-head-meta</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">br-img-br</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-br-p</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-html</WINDOWTAG><COUNT>1</COUNT>
</TAGSBYPAGE>
<TAGSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/page3.htm</PAGE>
<WINDOWTAG LENGHT="3">meta-meta-meta</WINDOWTAG><COUNT>3</COUNT>
<WINDOWTAG LENGHT="3">font-html-head</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">body-p-font</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">title-body-p</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">br-p-font</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">head-meta-meta</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">meta-title-body</WINDOWTAG><COUNT>1</COUNT>
```

```
<WINDOWTAG LENGHT="3">font-p-font</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">meta-meta-title</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">html-head-meta</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-html</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-p</WINDOWTAG><COUNT>1</COUNT>
</TAGSBYPAGE>
<TAGSBYSITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/page3.htm</PAGE>
<WINDOWTAG LENGHT="3">meta-meta-meta</WINDOWTAG><COUNT>12</COUNT>
<WINDOWTAG LENGHT="3">br-p-font</WINDOWTAG><COUNT>6</COUNT>
<WINDOWTAG LENGHT="3">title-body-p</WINDOWTAG><COUNT>4</COUNT>
<WINDOWTAG LENGHT="3">meta-meta-title</WINDOWTAG><COUNT>4</COUNT>
<WINDOWTAG LENGHT="3">html-head-meta</WINDOWTAG><COUNT>4</COUNT>
<WINDOWTAG LENGHT="3">body-p-font</WINDOWTAG><COUNT>4</COUNT>
<WINDOWTAG LENGHT="3">meta-title-body</WINDOWTAG><COUNT>4</COUNT>
<WINDOWTAG LENGHT="3">head-meta-meta</WINDOWTAG><COUNT>4</COUNT>
<WINDOWTAG LENGHT="3">font-html-head</WINDOWTAG><COUNT>3</COUNT>
<WINDOWTAG LENGHT="3">p-font-html</WINDOWTAG><COUNT>3</COUNT>
<WINDOWTAG LENGHT="3">a-br-p</WINDOWTAG><COUNT>3</COUNT>
<WINDOWTAG LENGHT="3">br-a-br</WINDOWTAG><COUNT>2</COUNT>
<WINDOWTAG LENGHT="3">img-br-a</WINDOWTAG><COUNT>2</COUNT>
<WINDOWTAG LENGHT="3">font-a-br</WINDOWTAG><COUNT>2</COUNT>
<WINDOWTAG LENGHT="3">p-font-a</WINDOWTAG><COUNT>2</COUNT>
<WINDOWTAG LENGHT="3">a-br-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">tr-td-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-br-img</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">td-a-td</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">font-table-tr</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">br-img-br</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-a-br</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">br-a-img</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">font-p-font</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-img-br</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">table-tr-td</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-table</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">td-a-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">a-td-a</WINDOWTAG><COUNT>1</COUNT>
<WINDOWTAG LENGHT="3">p-font-p</WINDOWTAG><COUNT>1</COUNT>
</TAGSBYSITE>
</SITE>
```

3.3.5 makeCorpusTAGForLexico-038.pl

- Format d'entrée :

Le programme `makeCorpusTagForLexico.pl` prend en entrée le corpus XML issu de `webxref` et génère un corpus qui contient uniquement les balises HTML en gardant l'identification du site et de la page concernée.

- Format de sortie :

Le corpus résultant a l'allure suivante :

```
<sitename=jura>
<page=index.html>
HTML HEAD META META END@HEAD BODY P END@P....
END@BODY END@HTML §
<page=index2.html>
HTML HEAD META META END@HEAD BODY P END@P....
END@BODY END@HTML §
....
<sitename=jura2>
....
```

- la séquence "END@" indique une balise HTML de fermeture
- le caractère "§" est une marque de paragraphe virtuelle pour éventuellement utiliser la carte des sections de Lexico

Le but de cette manipulation est de construire des données formatées pour l'outil Lexico[1] ce type de corpus pour y repérer en particulier les segments répétés de TAG pris linéairement dans la page HTML et plus si possible (calcul de spécificités…).

[1] http://www.cavi.unib-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

3.3.6 makeCorpusSelectedTAGForLexico4

- Format d'entrée :

Le programme `makeCorpusSelectedTagForLexico4.pl` prend en entrée le corpus XML issu de `webxref` et une liste préselectionnée de TAGs HTML, il génère un corpus qui contient uniquement les balises HTML sélectionnées en gardant l'identification du site et de la page concernée.

On dispose d'un programme annexe qui permet d'établir la liste des TAGs HTML contenues dans le corpus initial. Ce programme, `makeListTagForLexico.pl`, construit 2 listes : une liste de tous les TAGs et une liste des TAGs accompagnées de leur fréquence. C'est ce deuxième fichier qui peut être utilisé en entrée du programme `makeCorpusSelectedTAGForLexico4.pl`. Pour l'exemple donné infra, ce fichier contenant les TAGs sélectionnés a l'allure suivante :

```
FONT  8
A     8
IMG   2
TABLE 1
```

Ce qui signifie que le corpus produit par le programme ne contiendra que les `TAGs FONT, A, IMG, TABLE.`

- Format de sortie :

Le corpus résultant a l'allure suivante :

```
$ <SITENAME= siteDemo2 >
§ <PAGE=c:/SFleury/Recherche/Typweb/siteDemo2/index.htm>
TRAIT1 TRAIT2 TRAIT3 TRAIT4 TRAIT5 TRAIT1
$ <SITENAME= siteDemo2 >
§ <PAGE=c:/SFleury/Recherche/Typweb/siteDemo2/ss-dossier/page2.htm>
TRAIT1 TRAIT6 TRAIT1
$ <SITENAME= siteDemo2 >
§ <PAGE=c:/SFleury/Recherche/Typweb/siteDemo2/page1.htm>
TRAIT1 TRAIT7 TRAIT8 TRAIT9 TRAIT10 TRAIT11 TRAIT5 TRAIT1
$ <SITENAME= siteDemo2 >
§ <PAGE=c:/SFleury/Recherche/Typweb/siteDemo2/page3.htm>
TRAIT1 TRAIT1
```

Ce corpus de TAG est accompagné d'une table de correspondance des traits codés de manière générique dans le résultat donné supra :

TRAIT4	A#TYPE#HREF#VALUE#SSDOSSIERPAGE2HTM#
TRAIT11	A#TYPE#HREF#VALUE#HTTPWWWMICROSOFTCOMFRANCE#
TRAIT5	A#TYPE#HREF#VALUE#PAGE3HTM#
TRAIT8	IMG#TYPE#HEIGHT#VALUE#31#TYPE#ALT#VALUE#800*600#TYPE#SRC#VALUE#IMAGES800X600GIF#TYPE#WIDTH#VALUE#88#
TRAIT6	A#TYPE#HREF#VALUE#HTTPWWWNETSCAPECOMPAGE1HTM#
TRAIT1	FONT#TYPE#SIZE#VALUE#4#
TRAIT9	A#TYPE#HREF#VALUE#HTTPWWWMICROSOIFCOM#
TRAIT2	TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#
TRAIT3	A#TYPE#HREF#VALUE#PAGE1HTM#
TRAIT7	A#TYPE#HREF#VALUE#HTTPWWWNETSCAPEFR#
TRAIT10	IMG#TYPE#HEIGHT#VALUE#31#TYPE#ALT#VALUE#FRONTPAGE#TYPE#SRC#VALUE#HTTPWWWMICROSOIFCOMIMAGESFPCREATEGIF#TYPE#WIDTH#VALUE#88#

On dispose en sortie d'une état des fréquences des TAGs sélectionnés :

FONT#TYPE#SIZE#VALUE#4#	8
A#TYPE#HREF#VALUE#PAGE3HTM#	2
A#TYPE#HREF#VALUE#SSDOSSIERPAGE2HTM#	1
A#TYPE#HREF#VALUE#HTTPWWWMICROSOFTCOMFRANCE#	1
IMG#TYPE#HEIGHT#VALUE#31#TYPE#ALT#VALUE#800*600#TYPE#SRC#VALUE#IMAGES800X600GIF#TYPE#WIDTH#VALUE#88#	1

A#TYPE#HREF#VALUE#HTTPWWWNETSCAPECOMPAGE1HTM#	1
A#TYPE#HREF#VALUE#HTTPWWWMICROSOIFCOM#	1
TABLE#TYPE#CELLPADDING#VALUE#0#TYPE#BORDER#VALUE#0#TYPE#WIDTH#VALUE#100%#TYPE#CELLSPACING#VALUE#0#	1
A#TYPE#HREF#VALUE#PAGE1HTM#	1
A#TYPE#HREF#VALUE#HTTPWWWNETSCAPEFR#	1
IMG#TYPE#HEIGHT#VALUE#31#TYPE#ALT#VALUE#FRONTPAGE#TYPE#SRC#VALUE#HTTPWWWMICROSOIFCOMIMAGESFPCREATEGIF#TYPE#WIDTH#VALUE#88#	1

3.4 Constitution des matrices

3.4.1 Préambule

Dans les chaines 036 et 038 on obtient un état statistique par page du nombre de `tag` et de mots des pages du corpus regroupé dans un fichier noté `StatWordByFile`.

BH a travaillé à partir de ce fichier pour produire une matrice. Nous avons adopté un format commun pour lire facilement le fichier de stat et pour générer la matrice. Ce format est le suivant :

```
<TAGS>
<SITE>sitename</SITE>                ;;; nom du site
<PAGE>pagename</PAGE>                ;;; nom de la page
<ELEMENTS>
...                                         ;;; liste et fréquence des éléments
</ELEMENTS>
<ELEMENTS_ATTR>
...                                         ;;; liste et fréquence des attributs
</ELEMENTS_ATTR>
<WORDS>
<SITE>res1$jura$speleo1</SITE>
<PAGE>menu_acc</PAGE>
...                                         ;;; liste et fréquence des words
</WORDS>
```

et ce pour chaque page/fichier de chaque site.

(Cet état est produit par le programme `webxref-038` pour la chaîne 038 et par le programme `ExtAndStatMatrice` pour la chaîne 036)

3.4.2 Préparation des matrices

Les matrices sont produites à partir des sorties de type `StatWordByFile.txt` construites par `ExtAndStat` (chaîne 036) et `webxref-038` (chaîne 038). A partir de ce fichier qui donne pour chaque page les éléments et les mots employés, `StatWordByFile2Matrice.pl` produit une matrice dont le contenu peut être paramétré (cf document *infra*).

On reproduit ci-dessous le document de présentation pour la génération des matrices.

3.4.3 Construction des matrices

EtiquetageTyPWeb : Etiquetage du corpus TyPWeb

3.4.3.1 Documentation

3.4.3.1.1 Programmes fournis

```
-rwxr-xr-x   1 habert   habert       3909 Jan  3 11:22
ElimineColonnesLignesDeMatrice.pl*
```

```
-rwxr-xr-x   1 habert   habert       9965 Jan  3 11:15 EmondeProfilMatrice.pl*
-rwxr-xr-x   1 habert   habert       6643 Oct 14 15:23
FournitProfilColonnesLignesMatrice.pl*
-rwxr-xr-x   1 habert   habert       7589 Nov 15 00:31
LignesNomFichierCouplesFrequenceType2Matrice.pl*
-rwxr-xr-x   1 habert   habert       8789 Jan  3 11:00
StatWordByFile2ChoixTraitsPourStatWordByFile2Matrice.pl*
-rwxr-xr-x   1 habert   habert      15891 Jan  3 00:17
StatWordByFile2Matrice.pl*
```

Les programmes suivants sont appelés par d'autres programmes Perl et ne sont pas destinés à être appelés directement :

```
-rwxr-xr-x   1 habert   habert       6643 Oct 14 15:23
FournitProfilColonnesLignesMatrice.pl*
-rwxr-xr-x   1 habert   habert       7589 Nov 15 00:31
LignesNomFichierCouplesFrequenceType2Matrice.pl*
```

3.4.3.1.2 Journal

Le fichier Matrices.Journal sert de mémoire des traitements. Il est mis à jour à chaque appel de :

```
StatWordByFile2ChoixTraitsPourStatWordByFile2Matrice.pl
StatWordByFile2Matrice.pl
EmondeProfilMatrice.pl
ElimineColonnesLignesDeMatrice.pl
```

Ne pas le modifier à la main. Eventuellement le détruire s'il devient trop gros.

3.4.3.1.3 Etapes de travail

- Préparation du filtrage a priori d'une matrice : StatWordByFile2ChoixTraitsPourStatWordByFile2Matrice.pl
- Production d'une matrice de base (éventuellement filtrée a priori) : StatWordByFile2Matrice.pl
- Filtrage a posteriori d'une matrice : EmondeProfilMatrice.pl puis ElimineColonnesLignesDeMatrice.pl

3.4.4 Préparation du filtrage a priori d'une matrice :

`StatWordByFile2ChoixTraitsPourStatWordByFile2Matrice.pl`

- Format d'entrée :

```
StatWordByFile2ChoixTraitsPourStatWordByFile2Matrice.pl -segmentation
page|site [-all O|o] [-words O|o] [-tags O|o] [-elts O|o] [-attrs O|o] [-
attrvalues O|o] -entree <sorties StatWordByFile> -prefixe_sortie <préfixe du
fichier résultat> [-tri frequence|alphabetique]
```

Le format d'entrée est aussi proche que possible de celui de StatWordByFile2Matrice.pl. Trois arguments obligatoires (-segmentation, -entree, -prefixe_sortie). On peut se concentrer sur une partie seulement des traits grâce au jeu des options.

Attention : -prefixe_sortie et non -prefixe_sorties (contrairement à StatWordByFile2Matrice.pl). L'obligation de donner un préfixe de sortie correspond à la volonté de pouvoir constituer pour une matrice donnée plusieurs fichiers de choix de traits.

- Format de sortie :

Un fichier de choix où chaque ligne est de la forme :

`<CHOIX/><trait>\t<fréquence>`

Les lignes, par défaut (ou avec l'option (-tri frequence), sont triées par <fréquence> décroissante d'occurrence des traits, pour faciliter les choix. On peut aussi demander un tri par ordre alphabétique des traits (-tri alphabetique).

Ce fichier a pour préfixe la valeur donnée à -prefixe_sortie et pour extension $ExtensionChoixTraits (".ChoixTraitsPourStatWordByFile2Matrice").

Ce fichier sera modifié à la main pour être donné en argument à StatWordByFile2Matrice.pl. Les traits que l'on voudra conserver devront comporter O, o ou + en tout début de ligne (avant <CHOIX/>).

3.4.5 Production d'une matrice de base

Eventuellement filtrée a priori, en utilisant un résultat de StatWordByFile2ChoixTraitsPourStatWordByFile2Matrice.pl, modifié à la main : StatWordByFile2Matrice.pl

- Format d'entrée :

```
Emploi : StatWordByFile2Matrice.pl -segmentation page|site [-all O|o] [-words O|o] [-tags O|o] [-elts O|o] [-attrs O|o] [-attrvalues O|o] -entree <sorties StatWordByFile> -prefixe_sorties <partie commune des fichiers résultats> [-choix_traits <fichier de choix de traits>]
```

3 couples mot-clé / valeur sont obligatoires :
- segmentation page|site
- entree <sorties StatWordByFile>
- prefixe_sorties <partie commune des fichiers résultats>

On peut sélectionner tous les traits souhaités et toutes les combinaisons. Par ailleurs, si l'on fournit un <fichier de choix de traits> pour le mot-clé -choix_traits, en utilisant un fichier d'extension ".ChoixTraitsPourStatWordByFile2Matrice" engendré précédemment par StatWordByFile2ChoixTraitsPourStatWordByFile2Matrice.pl et modifié à la main (avec O, o ou + en première position pour les traits que l'on veut garder), la matrice engendrée ne comprendra que les traits retenus.

- Sorties produites par SF (StatWordByFile.txt), de la forme :

```
<TAGS>
<SITE>jura.speleo_new</SITE>
<PAGE>E:/sitesPPete99/jura.speleo_new/index.html</PAGE>
<ELEMENTS>
<ITEM>P</ITEM><FRQ>17</FRQ>
<ITEM>FONT</ITEM><FRQ>17</FRQ>
...
</ELEMENTS>
<ELEMENTS_ATTR>
<ITEM>FONT(COLOR)</ITEM><FRQ>10</FRQ>
<ITEM>FONT(FACE)</ITEM><FRQ>9</FRQ>
...
</ELEMENTS_ATTR>
<ELEMENTS_ATTRVALUE>
<ITEM>FONT(FACE=ARIAL)</ITEM><FRQ>7</FRQ>
<ITEM>IMG(ALIGN=BOTTOM)</ITEM><FRQ>4</FRQ>
...
</ELEMENTS_ATTRVALUE>
</TAGS>
<WORDS>
<SITE>jura.speleo_new</SITE>
<PAGE>E:/sitesPPete99/jura.speleo_new/index.html</PAGE>
<ITEM>le</ITEM><FRQ>4</FRQ>
<ITEM>et</ITEM><FRQ>3</FRQ>
...
</WORDS>
```

- Format des sorties :

Une matrice pour traitements statistiques, une correspondance noms / traits, un changeur noms -> traits, le profil des individus et le fichier pour indiquer les choix faits pour les individus, le profil des traits et le fichier pour indiquer les choix faits pour les traits (extensions : ".Matrice", ".Noms2Traits", ".ChangeNoms2Traits.pl", ".ProfilLignes", ".ChoixIndividus", ".ProfilColonnes", ".ChoixTraits").

3.4.6 Filtrage a posteriori d'une matrice

```
EmondeProfilMatrice.pl puis ElimineColonnesLignesDeMatrice.pl
```

3.4.6.1 EmondeProfilMatrice.pl

- Format d'entrée :

```
Usage : EmondeProfilMatrice.pl -type <individus|traits> -partie_commune <partie commune des noms de fichiers de la matrice traitée> [-repartition_plancher <entier positif>][-frequence_plancher <entier positif>][-moyenne_plancher <entier positif>][-ecart_type_plancher <entier positif>][-repartition_plafond <entier positif>][-frequence_plafond <entier positif>][-moyenne_plafond <entier positif>][-ecart_type_plafond <entier positif>]
```

Il s'agit ici de vraies options qui peuvent être dans n'importe quel ordre. Suppose l'existence de fichiers de nom <partie commune des noms de fichiers de la matrice traitée> et d'extension .ProfilColonnes et .ChoixTraits si -type a pour valeur traits ou d'extension .ProfilLignes et .ChoixIndividus si -type a pour valeur individus

Exemple d'appel :
```
EmondeProfilMatrice.pl -type traits -partie_commune essai8 -frequence_plancher 5 -frequence_plafond 12
```

prend en entrée un fichier de choix et un profil de lignes ou de colonnes de matrice, produits par StatWordByFile2Matrice.pl (FournitProfilColonnesLignesMatrice.pl).

De la forme pour les traits :

```
<trait><tabulation><# parties concernées><tabulation><fréquence totale><tabulation><moyenne><tabulation><écart type><tabulation><nom engendré><tabulation><numéro colonne><tabulation><à garder ou non>
```

Exemple :

```
A       4       4584    1146.0  0.0     _aaaaab_        1       O
```

et pour les individus :

```
<individu><tabulation><# parties concernées><tabulation><fréquence totale><tabulation><moyenne><tabulation><écart type><tabulation><individu><tabulation><à garder ou non>
```

Exemple :

```
jura_speleo1    56      23629   363.5   6808.5  jura_speleo1    N
```

- Format de sortie :

le profil d'entrée avec la dernière colonne modifiée (O remplacé par N) en fonction des seuils choisis, dans un fichier d'émondage de suffixe $SuffixeFichierProfilEmonde [.Emondage]. C'est ce fichier d'émondage qui sera utilisé par ElimineColonnesLignesDeMatrice.pl.

3.4.6.2 ElimineColonnesLignesDeMatrice.pl

- Format d'entrée :

ElimineColonnesLignesDeMatrice.pl <matrice produite par StatWordByFile2Matrice.pl><profil de colonnes et statut><profil de lignes et statut>

Exemple d'appel :
```
    ElimineColonnesLignesDeMatrice.pl es7.Matrice es7.ProfilColonnes.Emondage es7.ProfilLignes.Emondage
    ElimineColonnesLignesDeMatrice.pl es7.Matrice es7.ProfilColonnes es7.ProfilLignes.Emondage
```

(dans ce deuxième cas, les profils colonnes, les traits donc, sont laissés à l'identique).

Dans les profils, la dernière colonne est soit O soit N. Si elle n'est pas O, le trait (ou l'individu) sera éliminé de la matrice résultant. Si l'on veut garder soit les traits soit les individus inchangés, il suffit de fournir tel quel le fichier correspondant produit par StatWordByFile2Matrice.pl

- Format de sortie :

Une matrice dans lequel un ou des traits et/ou un ou des individus ont été éliminés. Le nom est celui de la matrice de départ avec le suffixe $SuffixeMatriceEmondee [.Emondee]

3.4.7 Problèmes et tâches

Tous les cas de figure n'ont pas été traités... Les valeurs données aux mots-clés lors des appels par mots-clés ne sont pas systématiquement testées en ce qui concerne leur cohérence.

3.4.8 Tests/exemples

1) Préparation du filtrage a priori d'une matrice

a) Engendrement du fichier de choix

```
StatWordByFile2ChoixTraitsPourStatWordByFile2Matrice.pl -segmentation site -all O -entree SitesAlain.StatWordByFile.txt -prefixe_sortie SitesAlain-all
```

On retient ici tous les traits (-all 0).

Fichier de choix de traits pour StatWordByFile2Matrice.pl : SitesAlain-all.ChoixTraitsPourStatWordByFile2Matrice

Si l'on veut élaguer a priori la matrice issue de SitesAlain.StatWordByFile.txt :

1) Modifier à la main SitesAlain-all.ChoixTraitsPourStatWordByFile2Matrice en ajoutant en début de ligne O, o ou + devant les traits que l'on veut garder

2) Appeler StatWordByFile2Matrice.pl avec les options souhaitées pour les traits

```
StatWordByFile2Matrice.pl -segmentation site <options pour les trits> -entree SitesAlain.StatWordByFile.txt -prefixe_sorties <préfixe sorties> -choix_traits SitesAlain-all.ChoixTraitsPourStatWordByFile2Matrice
```

b) Modification à la main de SitesAlain-all.ChoixTraitsPourStatWordByFile2Matrice en ajoutant en début de ligne O, o ou + devant les traits que l'on veut garder

Par exemple, les lignes :

```
<CHOIX/>INTERNAL (IMAGE)     170
<CHOIX/>dans      164
<CHOIX/>p   162
```

deviennent

```
O<CHOIX/>INTERNAL (IMAGE)    170
O<CHOIX/>dans     164
o<CHOIX/>p  162
```

2) Engendrement d'une matrice de base

a) Sans émondage a priori

```
StatWordByFile2Matrice.pl -segmentation site -all O -entree SitesAlain.StatWordByFile.txt -prefixe_sorties essai1
```

```
ll essai1.*
-rwxr-xr-x  1 habert  habert    195421 Jan  3 12:50 essai1.ChangeNoms2Traits.pl*
-rw-r--r--  1 habert  habert        98 Jan  3 12:50 essai1.ChoixIndividus
-rw-r--r--  1 habert  habert    211156 Jan  3 12:50 essai1.ChoixTraits
-rw-r--r--  1 habert  habert    136113 Jan  3 12:50 essai1.Matrice
-rw-r--r--  1 habert  habert    147356 Jan  3 12:50 essai1.Noms2Traits
-rw-r--r--  1 habert  habert    298617 Jan  3 12:50 essai1.ProfilColonnes
-rw-r--r--  1 habert  habert       218 Jan  3 12:50 essai1.ProfilLignes
```

b) Avec émondage a priori

```
StatWordByFile2Matrice.pl -segmentation site -all 0 -entree SitesAlain.StatWordByFile.txt -prefixe_sorties essai2 -choix_traits SitesAlain-all.ChoixTraitsPourStatWordByFile2Matrice
```

```
ll essai2*
-rwxr-xr-x  1 habert  habert       877 Jan  3 12:54 essai2.ChangeNoms2Traits.pl*
-rw-r--r--  1 habert  habert        98 Jan  3 12:54 essai2.ChoixIndividus
-rw-r--r--  1 habert  habert       728 Jan  3 12:54 essai2.ChoixTraits
-rw-r--r--  1 habert  habert       742 Jan  3 12:54 essai2.Matrice
-rw-r--r--  1 habert  habert       464 Jan  3 12:54 essai2.Noms2Traits
-rw-r--r--  1 habert  habert      1186 Jan  3 12:54 essai2.ProfilColonnes
-rw-r--r--  1 habert  habert       210 Jan  3 12:54 essai2.ProfilLignes
```

3) Filtrage a posteriori d'une matrice

a) Modification à la main des fichiers concernant les individus

Dans essai1.ChoixTraits :

```
<CHOIX/>alain.cf_new
```
devient
```
N<CHOIX/>alain.cf_new
```

b) Engendrement d'un profil modifié pour les individus à partir des choix modifiés à la main

```
EmondeProfilMatrice.pl -type individus -partie_commune essai1
```

```
more essai1.ProfilLignes
alain.bertrand_new      5368    19722   2.5     1644.3  alain.bertrand_new      0
alain.bosmans_new       3318    14513   1.8     1125.3  alain.bosmans_new       0
alain.cf_new    139     285     0.0     29.5    alain.cf_new    0
alain.dubus_new 250     697     0.1     77.4    alain.dubus_new 0

more  essai1.ProfilLignes.Emondage
#Répartition plancher 0  plafond 100000 ; frequence plancher 0  plafond 100000 ; moyenne plancher 0 plafond 100000 ; ecart_type
plancher 0 plafond 100000
alain.bertrand_new      5368    19722   2.5     1644.3  alain.bertrand_new      0
alain.bosmans_new       3318    14513   1.8     1125.3  alain.bosmans_new       0
alain.cf_new    139     285     0.0     29.5    alain.cf_new    N
alain.dubus_new 250     697     0.1     77.4    alain.dubus_new 0
```

c) Engendrement d'un profil modifié pour les colonnes à partir de seuils et de plafonds

```
EmondeProfilMatrice.pl -type traits -partie_commune essai1 -moyenne_plancher 50 -moyenne_plafond 200
```

Les traits correspondant à ces conditions sont les suivants :

```
A       4       230     57.5    107.5   _aaaaqa_        416     0
A(HREF) 4       205     51.2    105.3   _aaaaqb_        417     0
B       3       249     62.2    129.9   _aaaawy_        596     0
BR      2       220     55.0    187.1   _aaaayr_        641     0
FONT    3       367     91.8    204.1   _aaaazn_        663     0
```

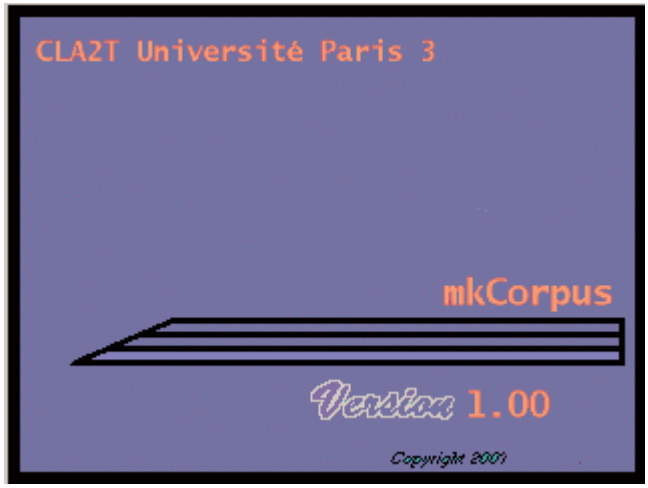
```
d       2       511     127.8   276.4   _aaaecg_        2762    0
des     2       430     107.5   239.7   _aaaefw_        2856    0
du      2       359     89.8    184.3   _aaaenf_        3047    0
en      3       297     74.2    148.8   _aaaevr_        3267    0
et      3       733     183.2   389.2   _aaafbc_        3408    0
l       2       548     137.0   290.9   _aaagti_        4558    0
le      3       522     130.5   265.6   _aaagvt_        4621    0
les     4       549     137.2   303.6   _aaagwu_        4648    0
un      3       246     61.5    126.5   _aaalfw_        7588    0
une     2       253     63.2    130.8   _aaalga_        7592    0
```

d) Engendrement d'une matrice émondée en fonction des profils émondés d'individus et de traits

```
ElimineColonnesLignesDeMatrice.pl  essai1.Matrice  essai1.ProfilColonnes.Emondage
essai1.ProfilLignes.Emondage

ll essai1*Matrice*
-rw-r--r--   1 habert   habert     136113 Jan  3 13:02 essai1.Matrice
-rw-r--r--   1 habert   habert        338 Jan  3 13:27 essai1.Matrice.Emondee
```

4 MKCORPUS



Mkcorpus est un programme de préparation de corpus pour leurs analyses ultérieures via des outils traditionnels du TAL. Il est écrit en Perl/TK.
Ce programme permet :

- de visualiser le corpus,
- de manipuler via des outils idoines le contenu du corpus et de ses éléments pour les formater suivant les contingences imposées par les outils (suppression de balises, nettoyage…).

Cet outil se présente comme un éditeur traditionnel et les menus construits permettent de réaliser des opérations sur les fichiers visualisés dans la zone d'édition ou attachés aux programmes de traitement.

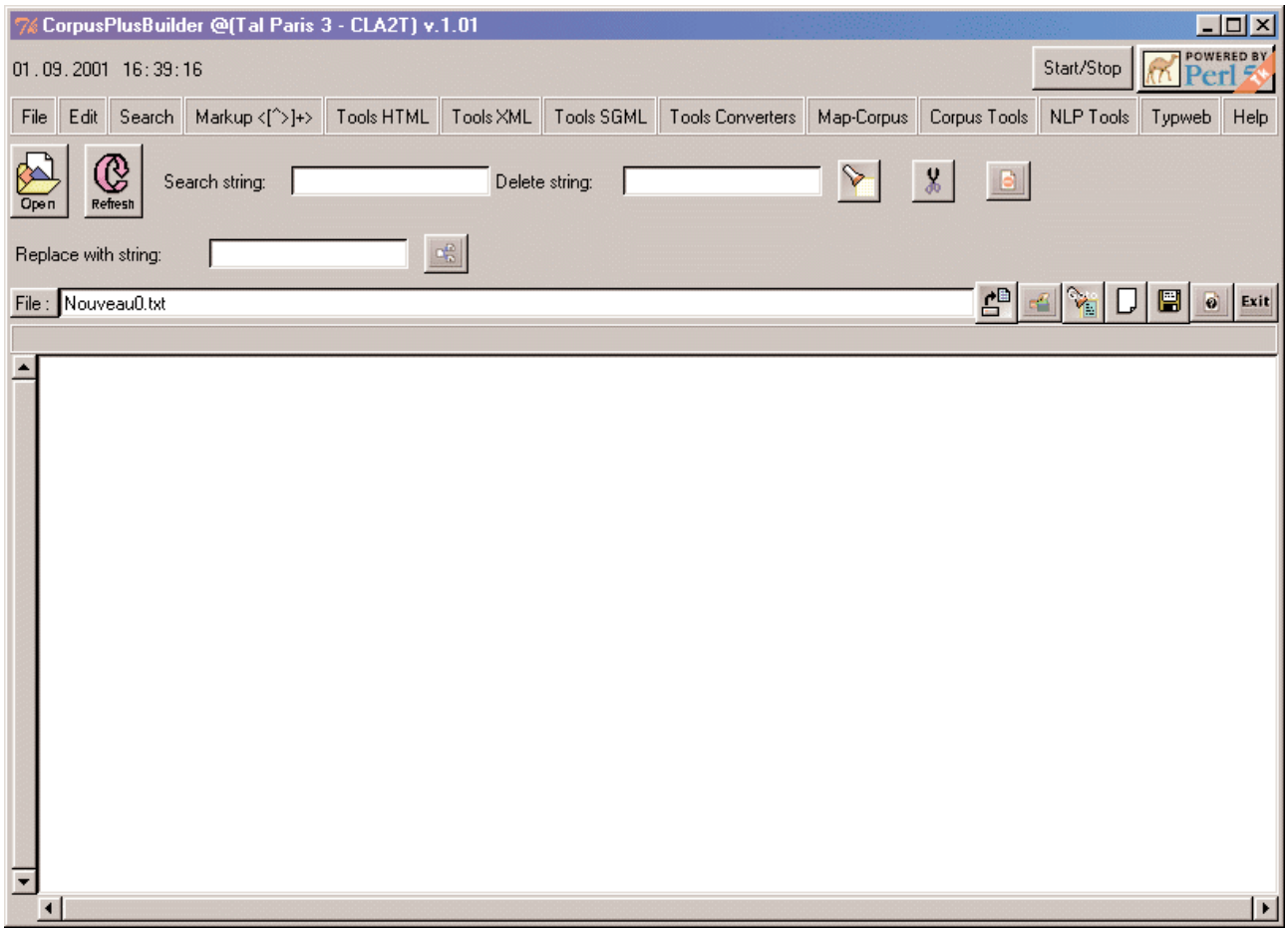


Figure 5 : interface MKCORPUS

- **Typweb** : la chaine Typweb est disponible complètement : version 036 et 038, avec les programmes associés. La phase d'aspiration de site est en cours de mise au point.

On donne ci-dessous une représentation de l'arborescence d'un site issu de webxref via MKCORPUS :

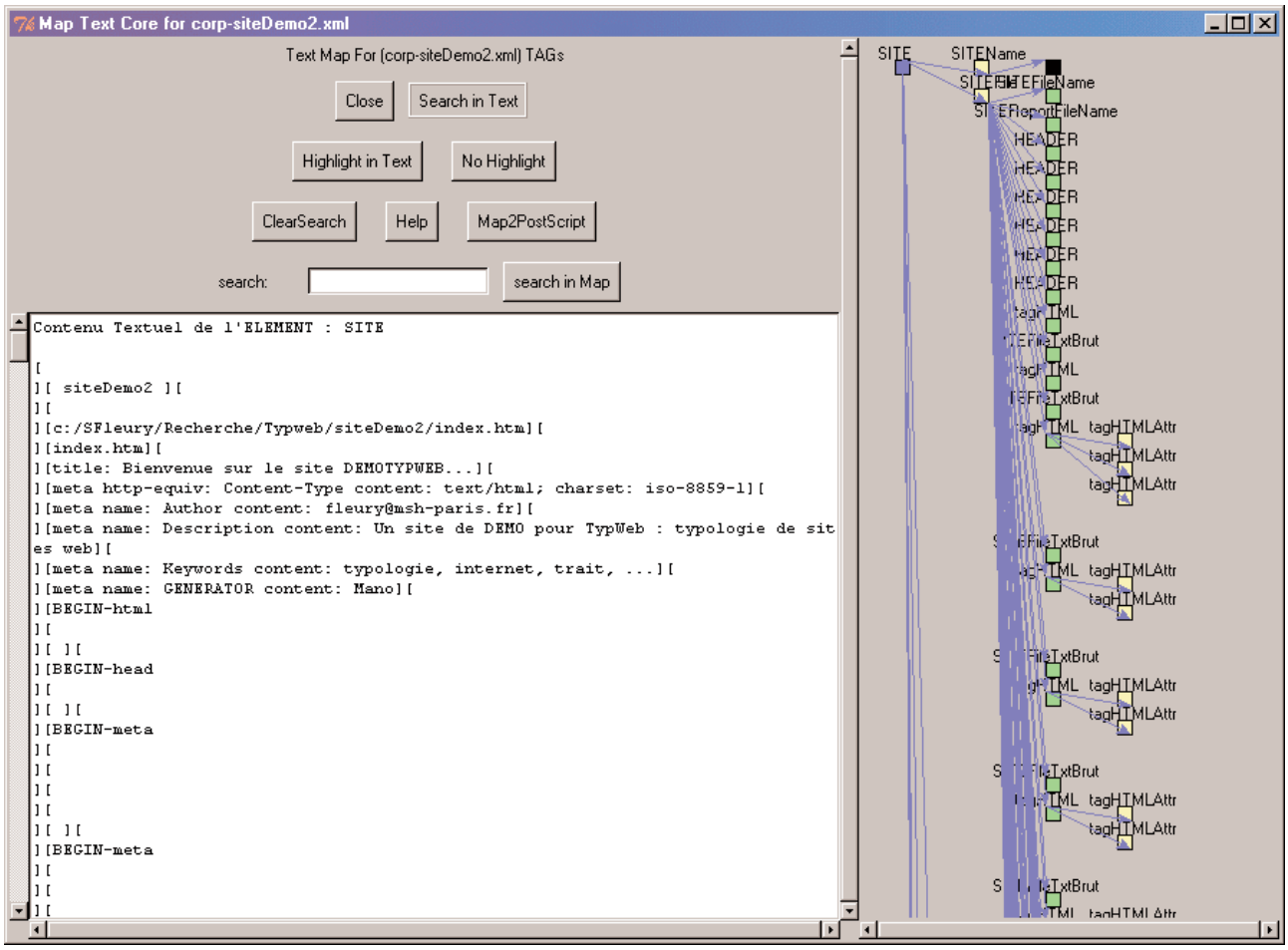


Figure 6 : Arborescence d'un corpus Typweb (un site) avec MKCORPUS

Pour utiliser MKCORPUS, il faut télécharger l'archive disponible sur la page :

www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/mkcorpusProject.htm

- Cette version correspond à l'archive nommée :

http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/typweb/tools/CorpusBuilderLocation.zip

- Cette archive contient toutes les ressources disponibles et le code source complet. Le plus simple pour l'installer est de "dézipper" l'archive dans un répertoire en respectant l'arborescence construite dans l'archive. Cette version a été testée sous Windows et Linux. Pour lancer MKCORPUS, il faut lancer le programme Perl nommé `loadMkCorpus.pl`.

  Pré-requis : il faut disposer de Perl et de Perl/Tk. Il faut aussi installer les modules Perl contenus dans le répertoire `mkCorpusModules` contenu dans l'archive. Pour ces installations suivre les instructions contenues dans les `Readme` de chaque module :il faut en général déposer une bibliothèque dans le répertoire local `lib` de Perl (en général c:\Perl\lib sous windows, sous linux, la procédure d'installation standard fait le travail parfaitement).