TypWeb : profilage de sites Web

Lot 1 : Analyse de la structure formelle et sémantique des sites

Document de travail

Equipe Typweb

Valérie Beaudouin[(*)], Serge Fleury[(***)], Benoît Habert[(**)(****)], Christian Licoppe[(*)], Marie Pasquier[(*)]

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

1 Sommaire

Table des figures

2 Préambule

Le projet TyPWeb, dans le cadre d'une collaboration avec le CNET, vise à adapter pour le traitement de sites Web l'architecture mise en œuvre dans le projet TyPTex et va nous conduire à passer du prototype actuel à une architecture générique de profilage. Ce projet vise à fournir un cadre méthodologique et pratique de profilage de sites Web et une typologie fine de ces sites. La démarche suivie vise à caractériser chaque site par des indicateurs de contenu et de structure. Il s'agit dans un premier temps de définir puis d'enrichir la description de sites par des indicateurs décrivant la forme et le contenu des sites visés : ces informations alimentent le cartouche descriptif des sites analysés, ce cartouche est conçu pour rester ouvert à toute nouvelle information pertinente capable de l'enrichir. TyPWeb doit conduire ensuite à proposer une typologie des contenus (en utilisant des index thématiques prédéfinis ou en constituant de nouvelles catégories de contenu en suivant la démarche inductive propre à l'architecture). L'analyse visée doit passer par un croisement de la structure formelle et des typologies de contenus produites. Il s'agit aussi de décrire l'articulation entre la description formelle et sémantique des sites avec les récits des pratiques des acteurs (concepteurs et visiteurs). Cette démarche vise en particulier à analyser la mise en place progressive de règles implicites d'échanges sur l'hypertoile.

3 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.

4 Etapes de travail : présentation générale

Le schéma de la figure suivante donne l' allure générale de l' architecture mise en œuvre pour le traitement des sites du projet Typweb :

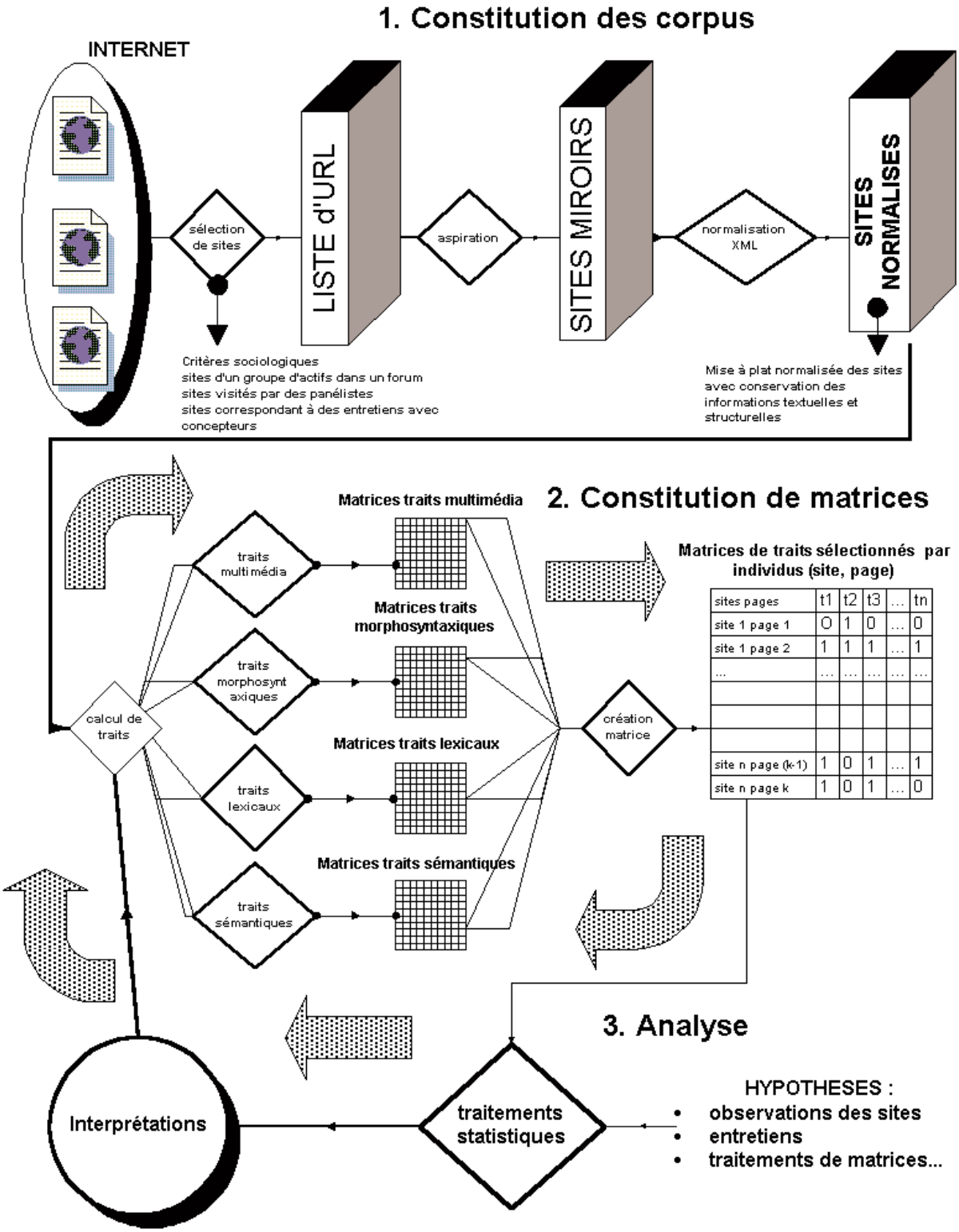


Figure 0 : architecture Typweb

4.1 Constitution d'un corpus de sites sur l'hypertoile

La première tâche de ce projet consiste à la constitution d'un corpus de sites "prélevés" sur le Web. Les sites aspirés sont sauvegardés localement après une aspiration via un outil idoine ; les aspirateurs utilisés sont les suivants :

HTTrack[1] développé par Xavier Roche et Yann Philippot et TELEPORT PRO 1.29. (Maisondieu & Kuncova 2000) présente un comparatif des outils réalisés pour constituer des sites miroirs via ce genre d'outil.

Notre corpus de sites est constitué de la manière suivante :

Pages personnelles

- 541 pages perso Wanadoo : aspiration de sites à partir d'un site leader visités (point d'entrée Frachon, leader du forum : http://perso.wanadoo.fr/jura.speleo) , aspiration aléatoire, aspiration de sites personnels.
- 584 sites Wanadoo visités en mars par le panel Netvalue.

Sites commerciaux

La seconde partie de ce corpus est constitué d'une centaine de sites marchands

Le rapport établi par Aude Maisondieu et Andréa Kuncova (2000) donne une présentation complète de cette phase de travail : démarche suivie, outils utilisés et testés, listes des sites traités.

4.2 Difficultés techniques pour la constitution du corpus de sites

Il faut souligner que l'aspiration des sites pose des problèmes techniques (Maisondieu & Kuncova 2000):

- pour gérer cette aspiration : les temps de traitements consacrés à l'aspiration peuvent être relativement longs suivant la taille des sites visités ; cette aspiration peut aussi comporter des difficultés techniques liées aux architectures mises en place pour construire les sites (scripts, programmes…) ;
- pour gérer le stockage des données aspirées, des données modifiées ou générées.

De plus le croisement de l'étude des sites (sur le plan du contenu et de la structure) et des récits des acteurs impose de fait une limitation dans le nombre de sites étudiés

4.3 Du corpus des sites au corpus normalisé des sites

Parallèlement à l'aspiration des sites traités dans ce projet, nous avons mis en place une chaîne de traitement pour normaliser le corpus de sites à soumettre à l'analyse. Cette chaîne de traitements vise à organiser les informations contenues dans les pages HTML des sites (éléments de structure et éléments de contenu).

Cette phase de travail a d'abord mené à la réalisation d'un "outil de dédossage" des éléments utilisés pour la constitution de pages HTML dans des sites Web : nous entendons par "désossage" la description formelle des éléments composant une page HTML ; ce travail réalisé sur un site complet vise donc à produire pour toutes les pages du site une analyse de ces pages et la production d'un rapport détaillant l'ensemble des éléments HTML présents dans chaque page. L'outil construit est l'adaptation du programme *webxref035*[2], écrit par Rick Jansen en juin 1995. La mise à jour réalisée sur cet outil permet de parcourir plusieurs sites, appliquant aux documents trouvés une fonction qui extrait de manière récursive les attributs HTML et les éléments qu'ils enchâssent en vue de traitements différenciés. **`WebXref-VersionTypweb`** produit des tableaux contenant les références des documents trouvés, URLs, ancres, images et fichiers afférents. La fonction *DissectFile* utilisée dans ce programme prend appui sur les algorithmes utilisés dans le script *dissectsite.hts*[3] disponible à l'adresse suivante : **http://worldwidemart.com/scripts/**.

4.4 Constitution d'un modèle de représentation des sites

La troisième étape de travail actuellement en cours doit conduire à la mise en place d'un modèle de document pour représenter les sites. Ces modèles doivent tenir compte à la fois de la cartographie interne du site (les arêtes du réseau

[1] "HTTrack est un aspirateur de sites web. Il vous permet de transférer un site web d'Internet vers votre disque dur, en construisant récursivement toute la structure, récupérant html, images et fichiers du serveur vers votre ordinateur. Les liens sont reconstruits de manière relative, de façon à pouvoir browser librement le site local via votre butineur habituel. Vous pouvez transférer (miroir) plusieurs sites ensembles de façon à pouvoir passer de l'un à l'autre librement. Vous pouvez également mettre à jour (update) un site existant, ou continuer un transfert interrompu. Le robot est entièrement configurable, avec une aide intégrée. WinHTTrack est la version Windows95/98/NT/2K de HTTrack (Hypertoile : `httrack.free.fr`)."

[2] *Webxref035* est distribué par le site `http://www.perl/com`. C'est un programme Perl conçu pour vérifier rapidement un ensemble local de pages HTML et mettre au jour les dysfonctionnements possibles : liens pointant vers des fichiers ou ancres manquants, etc.

[3] Ce script est utilisable dans un navigateur, il prend en argument une URL (en ligne) et génère des variables stockant l'information sur les en-têtes et les éléments envoyés par le serveur (" text objects " ou " tag objects "). Les résultats produits sont présentés au format HTML dans une nouvelle fenêtre du navigateur.

de liens entre les pages du site), du contenu multimédia des pages traités (les nœuds du réseau). Chaque nœud du réseau (les pages) étant eux-mêmes des micros-réseaux de liens : une page peut être constituée d'ancre pour faciliter la navigation. Pour réaliser cette normalisation dans la représentation des sites, nous prenons appui sur les rapports produits par ***`WebXref-VersionTypweb`***. Ces rapports sont transcodés au format XML : l' ensemble des rapports pour toutes les pages d' un site est regroupé dans un format qui tient compte des éléments propres au site global (structure et contenu) et des éléments propres à chaque page (idem). Ce transcodage produit en sortie pour chaque site un état du site au format XML qui contient donc des indicateurs de nature structurelle et de contenu.

Les outils utilisés dans cette chaîne de traitements seront présentés infra.

4.5 "Désossage" et normalisation d'un site avant analyse : démonstration

4.5.1 Le site démo

Pour illustrer notre travail nous allons travailler sur un site construit de toute pièce et dont on peut donner la composition suivante :

Une page d' index :

- Avec un lien vers une page (page1.htm) dans le même répertoire physique et contenant quelques liens internes, externes, images...
- Avec un lien vers une page (page2.htm) dans un sous-répertoire (sous-dossier) du répertoire de départ

Avec enfin, un lien existe entre la page 2 et/vers la page 1. Les figures qui suivent donnent une présentation du site via le browser de navigation.

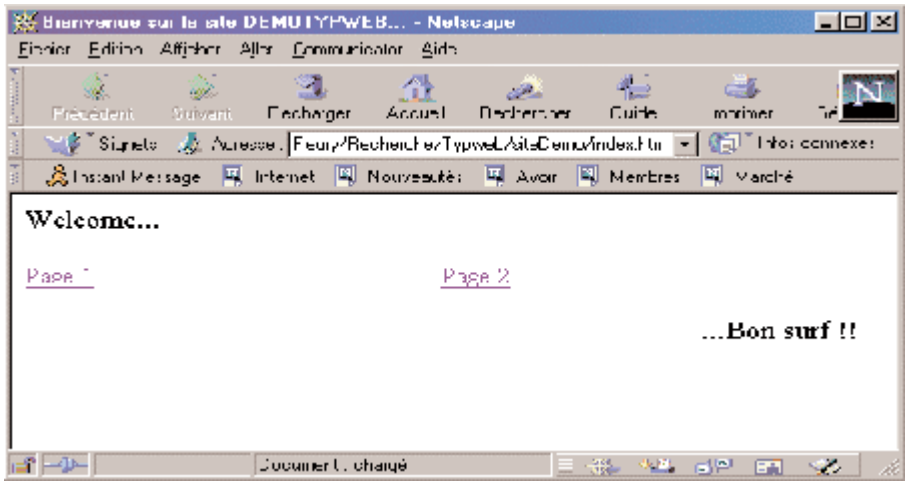


Figure 1 : page d'accueil du site DEMO

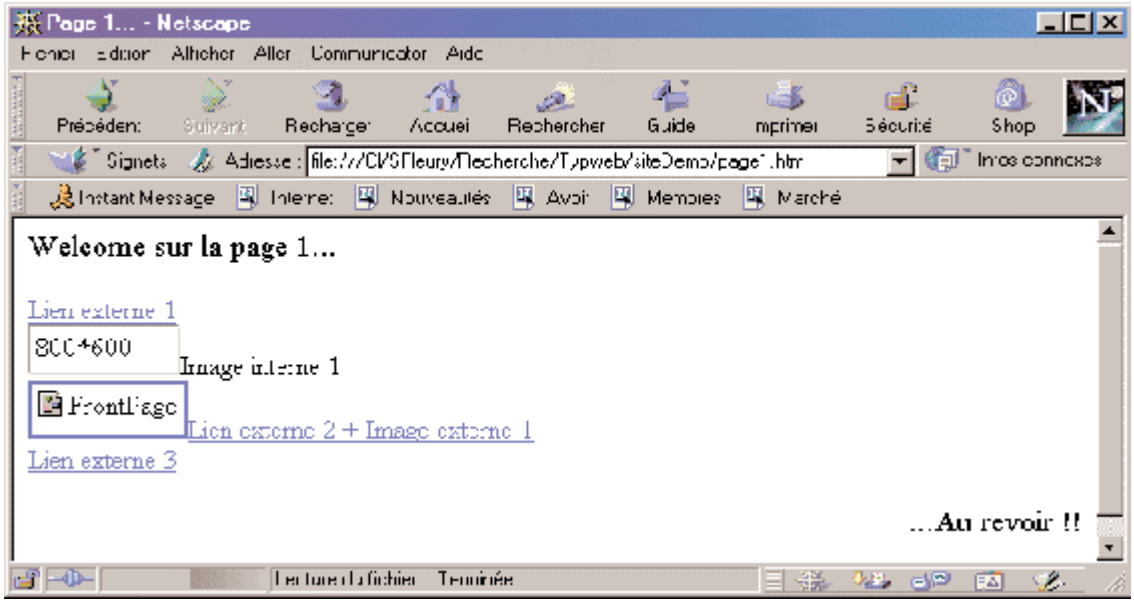


Figure 2 : page 1 du site DEMO

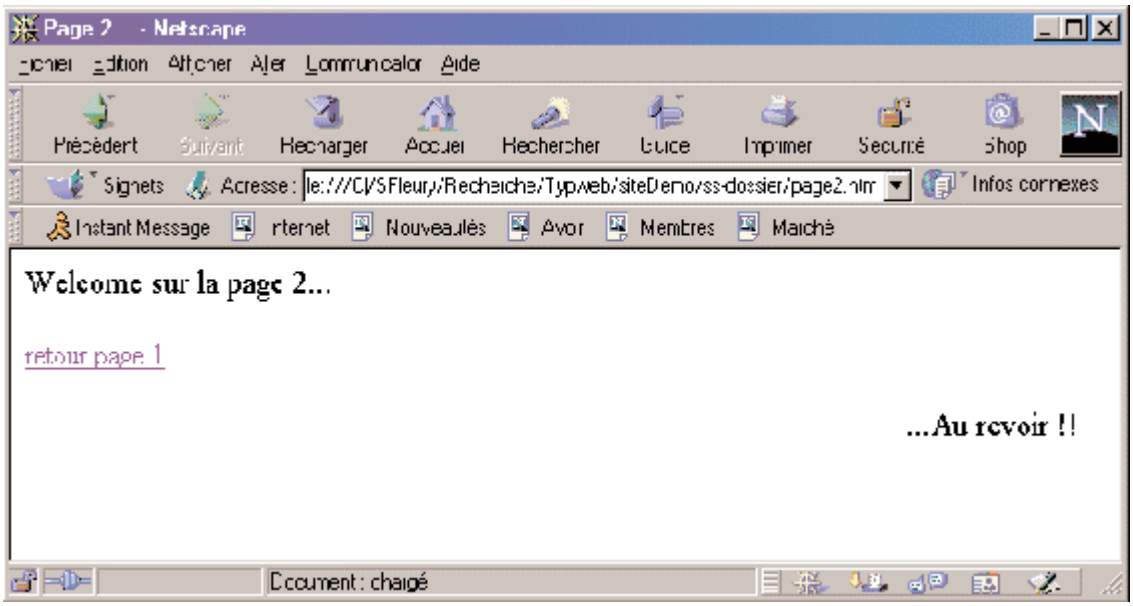


Figure 3 : page 2 du site DEMO

On peut construire une représentation graphique de ce site via l'outil « Astra Site Manager » (Mercury Interactive Corporation, 1996-1999). Ce programme construit une carte d 'un site web en ligne ou disponible localement. Cet outil est conçu initialement pour évaluer les structures et les évolutions des sites étudiés. Cet outil construit une représentation graphique du site en "descendant" les liens à partir du point d'entrée (le lien entre la page 2 et la page 1) n'est pas directement visible ici : pour le visualiser il faut explicitement demander de décrire les liens sortants de la page 2.

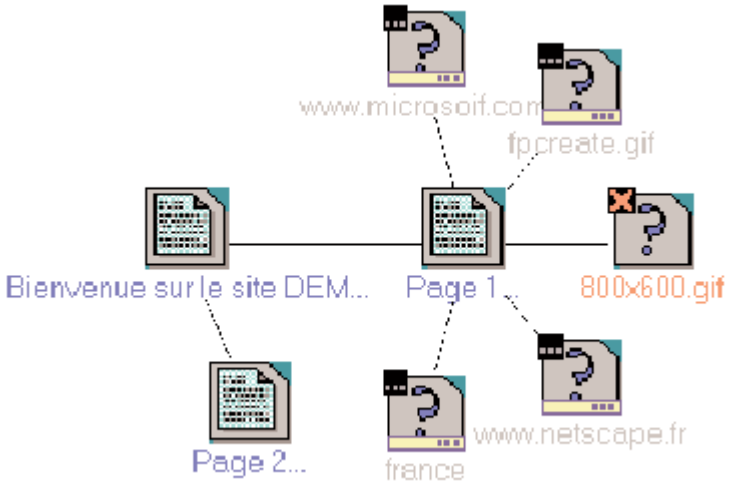


Figure 4 : représentation graphique (partielle) du site DEMO

4.5.2 `WebXref-VersionTypweb` sur le site Démo

Le programme **`WebXref-VersionTypweb`** produit une présentation de la structure formelle du site qui lui est soumis : composants structurels par exemple : éléments HTML avec attributs, éléments textuels… La figure suivant présente le rapport général sur le site : ce rapport donne la liste des fichiers du site pour lesquels le programme a construit un rapport ; il contient aussi des indicateurs généraux sur le site (nombre d'images, de liens externes…).

Site Analysis Results
Webxref version 0.3.5, 13-Mar-97 by Rick Jansen
Updated 12-Apr-2000 as part of the ENS/France Telecom project

---

Generated report files: 4

Report Files	Initial Files
rep1(index).html	index.htm
rep2(page2).html	ss-dossier/page2.htm
rep3(page1).html	page1.htm
rep4(page3).html	page3.htm

Web documents found: 4

Files	Referenced by
index.htm	--webxref--
page1.htm	index.htm ss-dossier/page2.htm
page3.htm	index.htm page1.htm
ss-dossier/page2.htm	index.htm

Directories: 1

Directories	Referenced by
/mnt/windows/SFleury/Recherche/Typweb/siteDemo/	--webxref--

External URLs: 4

URL	Referenced by

http://www.microsoft.com/france/	page1.htm
http://www.microsoif.com/	page1.htm
http://www.microsoif.com/Images/fpcreate.gif	page1.htm
http://www.netscape.fr/	page1.htm

```
------------------
Files not found: 1

------------------
```

Files	Referenced by
Images/800x600.gif	page1.htm

Figure 5 : rapport `WebXref-VersionTypweb` sur le site DEMO

Le rapport global contient aussi des liens vers les rapports construits pour chaque page. On présente ci-dessous le contenu d'un tel rapport. Celui-ci donne une description complète de tous les éléments composants la page décrite.

Document Analysis Results

Analysis of: /mnt/windows/SFleury/Recherche/Typweb/siteDemo/index.htm
by Calin MOSUT

Formatted MIME Headers from: index.htm

HEADER_1 title: Bienvenue sur le site DEMOTYPWEB...
HEADER_2 meta http-equiv: Content-Type content: text/html; charset: iso-8859-1
HEADER_3 meta name: Author content: fleury@msh-paris.fr
HEADER_4 meta name: Description content: Un site de DEMO pour TypWeb : typologie de sites web
HEADER_5 meta name: Keywords content: typologie, internet, trait, ...
HEADER_6 meta name: GENERATOR content: Mano

Elements of: index.htm

TAG OBJECT *ELEMENT:* html *NUMBER OF ATTRIBUTES:* 0	**No Attribute-Value Pairs**

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* head *NUMBER OF ATTRIBUTES:* 0	**No Attribute-Value Pairs**

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* meta *NUMBER OF ATTRIBUTES:* 3	**ATTRIBUTES AND VALUES:**	
	Attribute_1: charset	
	Value_1: iso-8859-1	
	Attribute_2: content	
	Value_2: text/html;	
	Attribute_3: http-equiv	
	Value_3: Content-Type	

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* meta *NUMBER OF ATTRIBUTES:* 2	**ATTRIBUTES AND VALUES:**	
	Attribute_1: content	
	Value_1: fleury@msh-paris.fr	
	Attribute_2: name	

Value_2:	Author	

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* meta *NUMBER OF ATTRIBUTES:* 2	**ATTRIBUTES AND VALUES:**		
	Attribute_1:	content	
	Value_1:	Un site de DEMO pour TypWeb : typologie de sites web	
	Attribute_2:	name	
	Value_2:	Description	

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* meta *NUMBER OF ATTRIBUTES:* 2	**ATTRIBUTES AND VALUES:**		
	Attribute_1:	content	
	Value_1:	typologie, internet, trait, ...	
	Attribute_2:	name	
	Value_2:	Keywords	

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* meta *NUMBER OF ATTRIBUTES:* 2	**ATTRIBUTES AND VALUES:**		
	Attribute_1:	content	
	Value_1:	Mano	
	Attribute_2:	name	
	Value_2:	GENERATOR	

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* title *NUMBER OF ATTRIBUTES:* 0	**No Attribute-Value Pairs**

TEXT OBJECT	`Bienvenue sur le site DEMOTYPWEB...`

TAG OBJECT *ELEMENT:* /title

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* /head

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* body *NUMBER OF ATTRIBUTES:* 2	**ATTRIBUTES AND VALUES:**	
	Attribute_1:	bgcolor
	Value_1:	#FFFFFF
	Attribute_2:	bgproperties
	Value_2:	fixed

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* p *NUMBER OF ATTRIBUTES:* 0	**No Attribute-Value Pairs**

TAG OBJECT *ELEMENT:* font *NUMBER OF ATTRIBUTES:* 1	**ATTRIBUTES AND VALUES:**	
	Attribute_1:	size
	Value_1:	4

TEXT OBJECT	Welcome...

TAG OBJECT *ELEMENT:* /font

TAG OBJECT *ELEMENT:* /p

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* table *NUMBER OF ATTRIBUTES:* 4	**ATTRIBUTES AND VALUES:**	
	Attribute_1:	cellpadding
	Value_1:	0
	Attribute_2:	border
	Value_2:	0
	Attribute_3:	width
	Value_3:	100%
	Attribute_4:	cellspacing
	Value_4:	0

TEXT OBJECT	**Blank space**

TAG OBJECT	**No Attribute-Value Pairs**

ELEMENT: tr *NUMBER OF ATTRIBUTES:* 0	

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* td *NUMBER OF ATTRIBUTES:* 2	**ATTRIBUTES AND VALUES:**		
	Attribute_1:	valign	
	Value_1:	bottom	
	Attribute_2:	rowspan	
	Value_2:	2	

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* a *NUMBER OF ATTRIBUTES:* 1	**ATTRIBUTES AND VALUES:**		
	Attribute_1:	href	
	Value_1:	page1.htm	

TEXT OBJECT	Page 1

TAG OBJECT *ELEMENT:* /a

TAG OBJECT *ELEMENT:* /td

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* td *NUMBER OF ATTRIBUTES:* 2	**ATTRIBUTES AND VALUES:**		
	Attribute_1:	valign	
	Value_1:	bottom	
	Attribute_2:	nowrap	
	Value_2:	nowrap	

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* a *NUMBER OF ATTRIBUTES:* 1	**ATTRIBUTES AND VALUES:**	
	Attribute_1:	href
	Value_1:	ss-dossier/page2.htm

TEXT OBJECT	Page 2

TAG OBJECT *ELEMENT:* /a

TAG OBJECT *ELEMENT:* /td

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* /tr

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* /table

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* a *NUMBER OF ATTRIBUTES:* 1	**ATTRIBUTES AND VALUES:**	
	Attribute_1:	href
	Value_1:	page3.htm

TEXT OBJECT	Page 3

TAG OBJECT *ELEMENT:* /a

TAG OBJECT *ELEMENT:* br *NUMBER OF ATTRIBUTES:* 0	**No Attribute-Value Pairs**

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* p *NUMBER OF ATTRIBUTES:* 1	**ATTRIBUTES AND VALUES:**		
	Attribute_1:	align	
	Value_1:	right	

TAG OBJECT *ELEMENT:* font *NUMBER OF ATTRIBUTES:* 1	**ATTRIBUTES AND VALUES:**		
	Attribute_1:	size	
	Value_1:	4	

TEXT OBJECT	...Bon surf !!

TAG OBJECT *ELEMENT:* /font

TAG OBJECT *ELEMENT:* /p

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* /body

TEXT OBJECT	**Blank space**

TAG OBJECT *ELEMENT:* /html

Links from: index.htm

LINK_1 INTERNAL (HTMLFILE) a href: ss-dossier/page2.htm
LINK_2 INTERNAL (HTMLFILE) a href: page1.htm
LINK_3 INTERNAL (HTMLFILE) a href: page3.htm

Figure 6 : Rapport détaillé du fichier index du site DEMO

4.5.3 Normalisation XML du site Démo

Un deuxième programme prend les rapports produits à l ’étape précédente et construit une représentation XML de tous les sites étudiés :

```
<?xml version="1.0"?>
<!DOCTYPE SITE SYSTEM "typweb.dtd">
<SITE>
<SITEName>res1-siteDemo</SITEName>
<SITEWebDocumentDissected> 4</SITEWebDocumentDissected>
<SITEWebDocumentReports>
<SITEReportFile NUM="1">index.html</SITEReportFile>
<SITEReportFile NUM="2">page2.html</SITEReportFile>
<SITEReportFile NUM="3">page1.html</SITEReportFile>
<SITEReportFile NUM="4">page3.html</SITEReportFile>
</SITEWebDocumentReports>
<SITEWebDocumentFound> 4</SITEWebDocumentFound>
<SITEWebDocumentUrls> 4</SITEWebDocumentUrls>
<SITEWebDocumentFileNotFound> 1</SITEWebDocumentFileNotFound>
<SITEFile>
<SITEFileName>res1(siteDemo)/rep1(index).html</SITEFileName>
<SITEFileMeta>
<SITEFileContent> text/html; charset: iso-8859-1</SITEFileContent>
<SITEFileDescription> Un site de DEMO pour TypWeb : typologie de sites web</SITEFileDescription>
<SITEFileGenerator> Mano</SITEFileGenerator>
<SITEFileKeywords> typologie, internet, trait, ...</SITEFileKeywords>
<SITEFileTitle> Bienvenue sur le site DEMOTYPWEB...</SITEFileTitle>
<SITEFileAuthor> fleury@msh-paris.fr</SITEFileAuthor>
</SITEFileMeta>
<SITEFileStructure>
<SITEFileElements>
<SITEFileElementsNb>21</SITEFileElementsNb>
<SiteFileElementDesc type="meta">5</SiteFileElementDesc>
<SiteFileElementDesc type="a">3</SiteFileElementDesc>
<SiteFileElementDesc type="font">2</SiteFileElementDesc>
<SiteFileElementDesc type="p">2</SiteFileElementDesc>
<SiteFileElementDesc type="td">2</SiteFileElementDesc>
<SiteFileElementDesc type="title">1</SiteFileElementDesc>
<SiteFileElementDesc type="body">1</SiteFileElementDesc>
<SiteFileElementDesc type="tr">1</SiteFileElementDesc>
<SiteFileElementDesc type="head">1</SiteFileElementDesc>
<SiteFileElementDesc type="table">1</SiteFileElementDesc>
<SiteFileElementDesc type="br">1</SiteFileElementDesc>
<SiteFileElementDesc type="html">1</SiteFileElementDesc>
</SITEFileElements>
<SITEFileTxtElements>27</SITEFileTxtElements>
<SITEFileImageNb>0</SITEFileImageNb>
<SITEFileImageDesc>0<EXTImage>0</EXTImage><INTImage>0</INTImage></SITEFileImageDesc>
<SITEFileLinks>
<SITEFileLinksNumber>3</SITEFileLinksNumber>
<SITEFileExternalLinks>0</SITEFileExternalLinks>
<SITEFileInternalLinks>0</SITEFileInternalLinks>
<SITEFileHtmlFileLinks>3</SITEFileHtmlFileLinks>
```

```
<SITEFileExtHypertextualLinks>0</SITEFileExtHypertextualLinks>
<SITEFileIntHypertextualLinks>0</SITEFileIntHypertextualLinks>
<SITEFileIntDocFileLinks>0</SITEFileIntDocFileLinks>
<SITEFileExternalCgiLinks>0</SITEFileExternalCgiLinks>
<SITEFileExternalNewsLinks>0</SITEFileExternalNewsLinks>
<SITEFileExternalFtpLinks>0</SITEFileExternalFtpLinks>
<SITEFileExternalGopherLinks>0</SITEFileExternalGopherLinks>
<SITEFileExternalMail>0</SITEFileExternalMail>
<SITEFileInternalAnchor>0</SITEFileInternalAnchor>
<SITEFileExternalAnchor>0</SITEFileExternalAnchor>
</SITEFileLinks>
</SITEFileStructure>
<SITEFileTxtBrut>
Welcome...
Page 1
Page 2
Page 3
...Bon surf !!

</SITEFileTxtBrut>
<SITEFileTxtAndTagContent>
<tagHTML type="html">begin_html</tagHTML>
<tagHTML type="head">begin_head</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="title">begin_title</tagHTML>
<tagHTML type="head">end_head</tagHTML>
<tagHTML type="body">begin_body</tagHTML>
<tagHTML type="p">begin_p</tagHTML>
<tagHTML type="font">begin_font</tagHTML>
Welcome...
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="p">end_p</tagHTML>
<tagHTML type="table">begin_table</tagHTML>
<tagHTML type="tr">begin_tr</tagHTML>
<tagHTML type="td">begin_td</tagHTML>
<tagHTML type="a">begin_a</tagHTML>
Page 1
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="td">end_td</tagHTML>
<tagHTML type="td">begin_td</tagHTML>
<tagHTML type="a">begin_a</tagHTML>
Page 2
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="td">end_td</tagHTML>
<tagHTML type="tr">end_tr</tagHTML>
<tagHTML type="table">end_table</tagHTML>
<tagHTML type="a">begin_a</tagHTML>
Page 3
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="br">begin_br</tagHTML>
<tagHTML type="p">begin_p</tagHTML>
<tagHTML type="font">begin_font</tagHTML>
...Bon surf !!
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="p">end_p</tagHTML>
<tagHTML type="body">end_body</tagHTML>
<tagHTML type="html">end_html</tagHTML>

</SITEFileTxtAndTagContent>
</SITEFile>
<SITEFile>
<SITEFileName>res1(siteDemo)/rep2(page2).html</SITEFileName>
<SITEFileMeta>
<SITEFileContent> text/html; charset: iso-8859-1</SITEFileContent>
<SITEFileDescription> Un site de DEMO pour TypWeb : typologie de sites web</SITEFileDescription>
<SITEFileGenerator> Mano</SITEFileGenerator>
<SITEFileKeywords> typologie, internet, trait, ...</SITEFileKeywords>
<SITEFileTitle> Page 2...</SITEFileTitle>
<SITEFileAuthor> fleury@msh-paris.fr</SITEFileAuthor>
</SITEFileMeta>
<SITEFileStructure>
<SITEFileElements>
<SITEFileElementsNb>15</SITEFileElementsNb>
<SiteFileElementDesc type="meta">5</SiteFileElementDesc>
<SiteFileElementDesc type="p">2</SiteFileElementDesc>
<SiteFileElementDesc type="font">2</SiteFileElementDesc>
```

```
<SiteFileElementDesc type="title">1</SiteFileElementDesc>
<SiteFileElementDesc type="body">1</SiteFileElementDesc>
<SiteFileElementDesc type="a">1</SiteFileElementDesc>
<SiteFileElementDesc type="head">1</SiteFileElementDesc>
<SiteFileElementDesc type="br">1</SiteFileElementDesc>
<SiteFileElementDesc type="html">1</SiteFileElementDesc>
</SITEFileElements>
<SITEFileTxtElements>17</SITEFileTxtElements>
<SITEFileImageNb>0</SITEFileImageNb>
<SITEFileImageDesc>0<EXTImage>0</EXTImage><INTImage>0</INTImage></SITEFileImageDesc>
<SITEFileLinks>
<SITEFileLinksNumber>1</SITEFileLinksNumber>
<SITEFileExternalLinks>0</SITEFileExternalLinks>
<SITEFileInternalLinks>0</SITEFileInternalLinks>
<SITEFileHtmlFileLinks>1</SITEFileHtmlFileLinks>
<SITEFileExtHypertextualLinks>0</SITEFileExtHypertextualLinks>
<SITEFileIntHypertextualLinks>0</SITEFileIntHypertextualLinks>
<SITEFileIntDocFileLinks>0</SITEFileIntDocFileLinks>
<SITEFileExternalCgiLinks>0</SITEFileExternalCgiLinks>
<SITEFileExternalNewsLinks>0</SITEFileExternalNewsLinks>
<SITEFileExternalFtpLinks>0</SITEFileExternalFtpLinks>
<SITEFileExternalGopherLinks>0</SITEFileExternalGopherLinks>
<SITEFileExternalMail>0</SITEFileExternalMail>
<SITEFileInternalAnchor>0</SITEFileInternalAnchor>
<SITEFileExternalAnchor>0</SITEFileExternalAnchor>
</SITEFileLinks>
</SITEFileStructure>
<SITEFileTxtBrut>
Welcome sur la page 2...
retour page 1
...Au revoir !!

</SITEFileTxtBrut>
<SITEFileTxtAndTagContent>
<tagHTML type="html">begin_html</tagHTML>
<tagHTML type="head">begin_head</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="title">begin_title</tagHTML>
<tagHTML type="head">end_head</tagHTML>
<tagHTML type="body">begin_body</tagHTML>
<tagHTML type="p">begin_p</tagHTML>
<tagHTML type="font">begin_font</tagHTML>
Welcome sur la page 2...
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="p">end_p</tagHTML>
<tagHTML type="a">begin_a</tagHTML>
retour page 1
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="br">begin_br</tagHTML>
<tagHTML type="p">begin_p</tagHTML>
<tagHTML type="font">begin_font</tagHTML>
...Au revoir !!
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="p">end_p</tagHTML>
<tagHTML type="body">end_body</tagHTML>
<tagHTML type="html">end_html</tagHTML>

</SITEFileTxtAndTagContent>
</SITEFile>
<SITEFile>
<SITEFileName>res1(siteDemo)/rep3(page1).html</SITEFileName>
<SITEFileMeta>
<SITEFileContent> text/html; charset: iso-8859-1</SITEFileContent>
<SITEFileDescription> Un site de DEMO pour TypWeb : typologie de sites web</SITEFileDescription>
<SITEFileGenerator> Mano</SITEFileGenerator>
<SITEFileKeywords> typologie, internet, trait, ...</SITEFileKeywords>
<SITEFileTitle> Page 1...</SITEFileTitle>
<SITEFileAuthor> fleury@msh-paris.fr</SITEFileAuthor>
</SITEFileMeta>
<SITEFileStructure>
<SITEFileElements>
<SITEFileElementsNb>24</SITEFileElementsNb>
<SiteFileElementDesc type="br">5</SiteFileElementDesc>
<SiteFileElementDesc type="meta">5</SiteFileElementDesc>
<SiteFileElementDesc type="a">4</SiteFileElementDesc>
<SiteFileElementDesc type="font">2</SiteFileElementDesc>
```

```
<SiteFileElementDesc type="p">2</SiteFileElementDesc>
<SiteFileElementDesc type="img">2</SiteFileElementDesc>
<SiteFileElementDesc type="body">1</SiteFileElementDesc>
<SiteFileElementDesc type="title">1</SiteFileElementDesc>
<SiteFileElementDesc type="head">1</SiteFileElementDesc>
<SiteFileElementDesc type="html">1</SiteFileElementDesc>
</SITEFileElements>
<SITEFileTxtElements>25</SITEFileTxtElements>
<SITEFileImageNb>2</SITEFileImageNb>
<SITEFileImageDesc>2<EXTImage>1</EXTImage><INTImage>1</INTImage></SITEFileImageDesc>
<SITEFileLinks>
<SITEFileLinksNumber>4</SITEFileLinksNumber>
<SITEFileExternalLinks>3</SITEFileExternalLinks>
<SITEFileInternalLinks>0</SITEFileInternalLinks>
<SITEFileHtmlFileLinks>1</SITEFileHtmlFileLinks>
<SITEFileExtHypertextualLinks>0</SITEFileExtHypertextualLinks>
<SITEFileIntHypertextualLinks>0</SITEFileIntHypertextualLinks>
<SITEFileIntDocFileLinks>0</SITEFileIntDocFileLinks>
<SITEFileExternalCgiLinks>0</SITEFileExternalCgiLinks>
<SITEFileExternalNewsLinks>0</SITEFileExternalNewsLinks>
<SITEFileExternalFtpLinks>0</SITEFileExternalFtpLinks>
<SITEFileExternalGopherLinks>0</SITEFileExternalGopherLinks>
<SITEFileExternalMail>0</SITEFileExternalMail>
<SITEFileInternalAnchor>0</SITEFileInternalAnchor>
<SITEFileExternalAnchor>0</SITEFileExternalAnchor>
</SITEFileLinks>
</SITEFileStructure>
<SITEFileTxtBrut>
Welcome sur la page 1...
Lien externe 1
Image interne 1
Lien externe 2 + Image externe 1
Lien externe 3
vers page 3
...Au revoir !!

</SITEFileTxtBrut>
<SITEFileTxtAndTagContent>
<tagHTML type="html">begin_html</tagHTML>
<tagHTML type="head">begin_head</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="title">begin_title</tagHTML>
<tagHTML type="head">end_head</tagHTML>
<tagHTML type="body">begin_body</tagHTML>
<tagHTML type="p">begin_p</tagHTML>
<tagHTML type="font">begin_font</tagHTML>
Welcome sur la page 1...
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="p">end_p</tagHTML>
<tagHTML type="a">begin_a</tagHTML>
Lien externe 1
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="br">begin_br</tagHTML>
<tagHTML type="img">begin_img</tagHTML>
Image interne 1
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="br">begin_br</tagHTML>
<tagHTML type="a">begin_a</tagHTML>
<tagHTML type="img">begin_img</tagHTML>
Lien externe 2 + Image externe 1
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="br">begin_br</tagHTML>
<tagHTML type="a">begin_a</tagHTML>
Lien externe 3
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="br">begin_br</tagHTML>
<tagHTML type="a">begin_a</tagHTML>
vers page 3
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="br">begin_br</tagHTML>
<tagHTML type="p">begin_p</tagHTML>
<tagHTML type="font">begin_font</tagHTML>
...Au revoir !!
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="p">end_p</tagHTML>
<tagHTML type="body">end_body</tagHTML>
```

```
<tagHTML type="html">end_html</tagHTML>

</SITEFileTxtAndTagContent>
</SITEFile>
<SITEFile>
<SITEFileName>res1(siteDemo)/rep4(page3).html</SITEFileName>
<SITEFileMeta>
<SITEFileContent> text/html; charset: iso-8859-1</SITEFileContent>
<SITEFileDescription> Un site de DEMO pour TypWeb : typologie de sites web</SITEFileDescription>
<SITEFileGenerator> Mano</SITEFileGenerator>
<SITEFileKeywords> typologie, internet, trait, ...</SITEFileKeywords>
<SITEFileTitle> Page 3...</SITEFileTitle>
<SITEFileAuthor> fleury@msh-paris.fr</SITEFileAuthor>
</SITEFileMeta>
<SITEFileStructure>
<SITEFileElements>
<SITEFileElementsNb>13</SITEFileElementsNb>
<SiteFileElementDesc type="meta">5</SiteFileElementDesc>
<SiteFileElementDesc type="p">2</SiteFileElementDesc>
<SiteFileElementDesc type="font">2</SiteFileElementDesc>
<SiteFileElementDesc type="title">1</SiteFileElementDesc>
<SiteFileElementDesc type="body">1</SiteFileElementDesc>
<SiteFileElementDesc type="head">1</SiteFileElementDesc>
<SiteFileElementDesc type="html">1</SiteFileElementDesc>
</SITEFileElements>
<SITEFileTxtElements>15</SITEFileTxtElements>
<SITEFileImageNb>0</SITEFileImageNb>
<SITEFileImageDesc>0<EXTImage>0</EXTImage><INTImage>0</INTImage></SITEFileImageDesc>
<SITEFileLinks>
<SITEFileLinksNumber>0</SITEFileLinksNumber>
<SITEFileExternalLinks>0</SITEFileExternalLinks>
<SITEFileInternalLinks>0</SITEFileInternalLinks>
<SITEFileHtmlFileLinks>0</SITEFileHtmlFileLinks>
<SITEFileExtHypertextualLinks>0</SITEFileExtHypertextualLinks>
<SITEFileIntHypertextualLinks>0</SITEFileIntHypertextualLinks>
<SITEFileIntDocFileLinks>0</SITEFileIntDocFileLinks>
<SITEFileExternalCgiLinks>0</SITEFileExternalCgiLinks>
<SITEFileExternalNewsLinks>0</SITEFileExternalNewsLinks>
<SITEFileExternalFtpLinks>0</SITEFileExternalFtpLinks>
<SITEFileExternalGopherLinks>0</SITEFileExternalGopherLinks>
<SITEFileExternalMail>0</SITEFileExternalMail>
<SITEFileInternalAnchor>0</SITEFileInternalAnchor>
<SITEFileExternalAnchor>0</SITEFileExternalAnchor>
</SITEFileLinks>
</SITEFileStructure>
<SITEFileTxtBrut>
Welcome sur la page 3...
...Au revoir !!

</SITEFileTxtBrut>
<SITEFileTxtAndTagContent>
<tagHTML type="html">begin_html</tagHTML>
<tagHTML type="head">begin_head</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="meta">begin_meta</tagHTML>
<tagHTML type="title">begin_title</tagHTML>
<tagHTML type="head">end_head</tagHTML>
<tagHTML type="body">begin_body</tagHTML>
<tagHTML type="p">begin_p</tagHTML>
<tagHTML type="font">begin_font</tagHTML>
Welcome sur la page 3...
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="p">end_p</tagHTML>
<tagHTML type="p">begin_p</tagHTML>
<tagHTML type="font">begin_font</tagHTML>
...Au revoir !!
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="p">end_p</tagHTML>
<tagHTML type="body">end_body</tagHTML>
<tagHTML type="html">end_html</tagHTML>

</SITEFileTxtAndTagContent>
</SITEFile>
</SITE>
```

La figure qui suit est celle obtenue par la lecture de ce corpus dans un éditeur de documents XML.

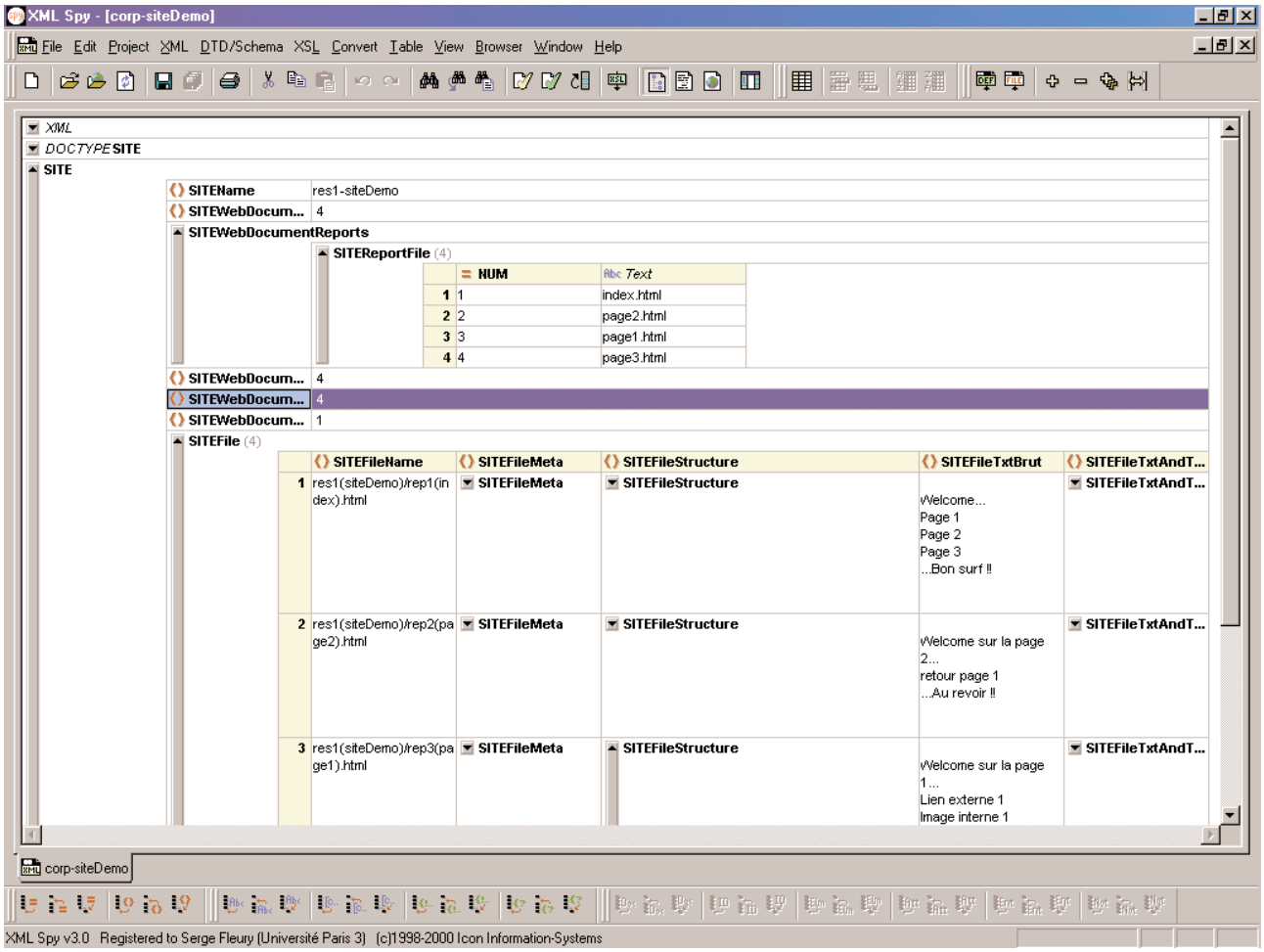


Figure 7 : représentation XML du site DEMO

5 Les outils utilisés

La chaîne de traitements mise en œuvre sur le corpus de sites est constituée de plusieurs outils :

1. Phase1: `Webxref-versionTypweb`

Génération de rapports (au format html ou txt) de tous les composants (textuels et structurels) de toutes les pages web d'un site donné.

2. Phase2 : `Mktipo`

Pour chaque site passé dans la phase1, génération d'un corpus au format XML regroupant toutes les informations d'un site donné

3. Phase 3 : `ExtAndStatFrCorpTwp`

Les corpus construits dans la phase 2 peuvent être soumis à cet outil pour produire un état détaillé de tous les composants du site visé (index de mots, d'éléments structurels).

4. Phase 4 : intégration des phases 1 à 3

On verra infra que la chaîne de traitements précédente se réalisent par deux séries de programmes :

1. La première notée 036 comprend les programmes : `webxref-036`, `mktipo-036` et `ExtAndStatFrCorpTwp-036`. Ces trois scripts perl produisent les résultats présentés *supra*.

2. La deuxième série de programme est en fait une intégration des différents outils de la série précédente en un seul programme, `webxref-038`, et une prise en compte d'éléments supplémentaire dans les pages web analysées. Cette version permet en effet de traiter les attributs des tags HTML utilisés dans les pages scrutées.

Les principes généraux des outils de la chaîne 036 présentés ci-dessous sont identiques à la version 038. On reviendra plus loin sur les éléments supplémentaires pris en compte dans la version 038.

5. MkCorpus : un préparateur de corpus

L'outil `mkCorpus` est un logiciel pour le traitement de corpus balisés. Ce logiciel est actuellement en cours de développement, il intègre en particulier tous les outils développés dans ce projet (Phase 1-3).

5.1 WebXref-VersionTypweb

Le programme `webxref-versionTypweb` est présenté de manière complète par l'auteur (Mosut 2000). On se reportera au document en question donné en annexes.

5.2 Mktipo

Le programme `mktipo` vise à construire un corpus normalisé à partir des rapports construits par le programme `webxref` sur un site donné. Ce programme est disponible sur la page http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/typweb.htm .

A l'issue du traitement, ce programme produit un fichier XML regroupant toutes les informations associées à la description des éléments structurels et textuels du site visé. Cette étape de normalisation permet de structurer les informations à analyser. Elle a aussi permis de délimiter les zones textuelles des sites étudiés et de tester les outils d'analyse textuelle sur ces textes. Chaque fichier XML (représentant un site) est associé à la DTD représentée ci-dessous suivie de l'arbre des éléments composant cette DTD :

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v3.0 (http://www.xmlspy.com) by Serge Fleury (UniversitÃ© Paris 3) -->
<!--DTD generated by XML Spy v3.0 (http://www.xmlspy.com)-->
<!ELEMENT EXTImage (#PCDATA)>
<!ELEMENT INTImage (#PCDATA)>
<!ELEMENT SITE (SITEName, SITEWebDocumentDissected, SITEWebDocumentReports, SITEWebDocumentFound,
SITEWebDocumentImages, SITEWebDocumentMailTo, SITEWebDocumentUrls?, SITEWebDocumentAnchorFound,
SITEWebDocumentFileNotFound, SITEWebDocumentAnchorNotFound, SITEFile+)>
<!ELEMENT SITEFile (SITEFileName, SITEFileMeta, SITEFileStructure, SITEFileTxtBrut,
SITEFileTxtAndTagContent)>
<!ELEMENT SITEFileAuthor (#PCDATA)>
<!ELEMENT SITEFileContent (#PCDATA)>
<!ELEMENT SITEFileDescription (#PCDATA)>
<!ELEMENT SITEFileElements (SITEFileElementsNb, SiteFileElementDesc+)>
<!ELEMENT SITEFileElementsNb (#PCDATA)>
<!ELEMENT SITEFileExtHypertextualLinks (#PCDATA)>
<!ELEMENT SITEFileExternalAnchor (#PCDATA)>
<!ELEMENT SITEFileExternalCgiLinks (#PCDATA)>
<!ELEMENT SITEFileExternalFtpLinks (#PCDATA)>
<!ELEMENT SITEFileExternalGopherLinks (#PCDATA)>
<!ELEMENT SITEFileExternalLinks (#PCDATA)>
<!ELEMENT SITEFileExternalMail (#PCDATA)>
<!ELEMENT SITEFileExternalNewsLinks (#PCDATA)>
<!ELEMENT SITEFileGenerator (#PCDATA)>
<!ELEMENT SITEFileHtmlFileLinks (#PCDATA)>
<!ELEMENT SITEFileImageDesc (#PCDATA | EXTImage | INTImage)*>
<!ELEMENT SITEFileImageNb (#PCDATA)>
<!ELEMENT SITEFileIntDocFileLinks (#PCDATA)>
<!ELEMENT SITEFileIntHypertextualLinks (#PCDATA)>
<!ELEMENT SITEFileInternalAnchor (#PCDATA)>
<!ELEMENT SITEFileInternalLinks (#PCDATA)>
<!ELEMENT SITEFileKeywords (#PCDATA)>
<!ELEMENT SITEFileLinks (SITEFileLinksNumber, SITEFileExternalLinks, SITEFileInternalLinks,
SITEFileHtmlFileLinks, SITEFileExtHypertextualLinks, SITEFileIntHypertextualLinks,
SITEFileIntDocFileLinks, SITEFileExternalCgiLinks, SITEFileExternalNewsLinks,
SITEFileExternalFtpLinks, SITEFileExternalGopherLinks, SITEFileExternalMail, SITEFileInternalAnchor,
SITEFileExternalAnchor)>
<!ELEMENT SITEFileLinksNumber (#PCDATA)>
```

```
<!ELEMENT SITEFileMeta (SITEFileContent, SITEFileDescription, SITEFileGenerator, SITEFileKeywords,
SITEFileTitle, SITEFileAuthor)>
<!ELEMENT SITEFileName (#PCDATA)>
<!ELEMENT SITEFileStructure (SITEFileElements, SITEFileTxtElements, SITEFileImageNb,
SITEFileImageDesc, SITEFileLinks)>
<!ELEMENT SITEFileTitle (#PCDATA)>
<!ELEMENT SITEFileTxtAndTagContent (#PCDATA | tagHTML)*>
<!ELEMENT SITEFileTxtBrut (#PCDATA)>
<!ELEMENT SITEFileTxtElements (#PCDATA)>
<!ELEMENT SITEName (#PCDATA)>
<!ELEMENT SITEReportFile (#PCDATA)>
<!ATTLIST SITEReportFile
       NUM CDATA #REQUIRED
>
<!ELEMENT SITEWebDocumentDissected (#PCDATA)>
<!ELEMENT SITEWebDocumentFileNotFound (#PCDATA)>
<!ELEMENT SITEWebDocumentFound (#PCDATA)>
<!ELEMENT SITEWebDocumentReports (SITEReportFile+)>
<!ELEMENT SITEWebDocumentUrls (#PCDATA)>
<!ELEMENT SITEWebDocumentAnchorFound (#PCDATA)>
<!ELEMENT SITEWebDocumentAnchorNotFound (#PCDATA)>
<!ELEMENT SITEWebDocumentImages (#PCDATA)>
<!ELEMENT SITEWebDocumentMailTo (#PCDATA)>

<!ELEMENT SiteFileElementDesc (#PCDATA)>
<!ATTLIST SiteFileElementDesc
       type CDATA #REQUIRED
>
<!ELEMENT tagHTML (#PCDATA)>
<!ATTLIST tagHTML
       type CDATA #REQUIRED
>
```

Figure 8 : DTD pour site Typweb

```
--------------------------------------------------------------------------------
                                    SITE
--------------------------------------------------------------------------------
SITE
 |_(sitename,
 |   |_(#PCDATA)
 |
 |__sitewebdocumentdissected,
 |   |_(#PCDATA)
 |
 |__sitewebdocumentreports,
 |   |_(sitereportfile+)
 |       |_(#PCDATA)
 |
 |
 |__sitewebdocumentfound,
 |   |_(#PCDATA)
 |
 |__sitewebdocumentimages,
 |   |_(#PCDATA)
 |
 |__sitewebdocumentmailto,
 |   |_(#PCDATA)
 |
 |__sitewebdocumenturls?,
 |   |_(#PCDATA)
 |
 |__sitewebdocumentanchorfound,
 |   |_(#PCDATA)
 |
 |__sitewebdocumentfilenotfound,
 |   |_(#PCDATA)
 |
 |__sitewebdocumentanchornotfound,
 |   |_(#PCDATA)
 |
 |__sitefile+)
     |_(sitefilename,
     |   |_(#PCDATA)
     |
     |__sitefilemeta,
     |   |_(sitefilecontent,
     |   |   |_(#PCDATA)
     |   |
```

```
      |   |__sitefiledescription,
      |   |   |_(#PCDATA)
      |   |
      |   |__sitefilegenerator,
      |   |   |_(#PCDATA)
      |   |
      |   |__sitefilekeywords,
      |   |   |_(#PCDATA)
      |   |
      |   |__sitefiletitle,
      |   |   |_(#PCDATA)
      |   |
      |   |__sitefileauthor)
      |       |_(#PCDATA)
      |
      |
      |__sitefilestructure,
      |   |_(sitefileelements,
      |   |   |_(sitefileelementsnb,
      |   |   |   |_(#PCDATA)
      |   |   |
      |   |   |__sitefileelementdesc+)
      |   |       |_(#PCDATA)
      |   |
      |   |
      |   |__sitefiletxtelements,
      |   |   |_(#PCDATA)
      |   |
      |   |__sitefileimagenb,
      |   |   |_(#PCDATA)
      |   |
      |   |__sitefileimagedesc,
      |   |   |_(#PCDATA |
      |   |   |__extimage |
      |   |   |   |_(#PCDATA)
      |   |   |
      |   |   |__intimage)*
      |   |       |_(#PCDATA)
      |   |
      |   |
      |   |__sitefilelinks)
      |       |_(sitefilelinksnumber,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileexternallinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileinternallinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefilehtmlfilelinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileexthypertextuallinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileinthypertextuallinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileintdocfilelinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileexternalcgilinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileexternalnewslinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileexternalftplinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileexternalgopherlinks,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileexternalmail,
      |       |   |_(#PCDATA)
      |       |
      |       |__sitefileinternalanchor,
      |       |   |_(#PCDATA)
      |       |
```

```
     |       |__sitefileexternalanchor)
     |           |_(#PCDATA)
     |
     |
     |
     |__sitefiletxtbrut,
     |   |_(#PCDATA)
     |
     |__sitefiletxtandtagcontent)
         |_(#PCDATA |
         |__taghtml)*
             |_(#PCDATA)
```

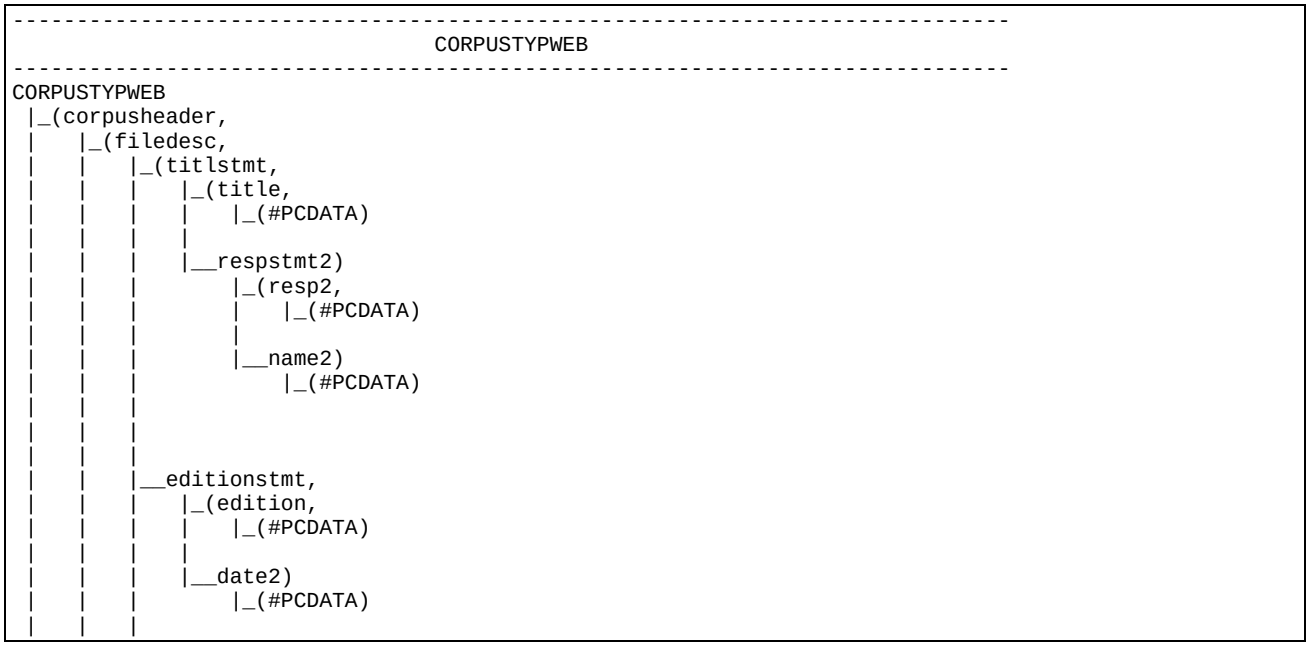
Figure 9 : arbre des éléments d'un site normalisé

Notre corpus doit *in fine* être composé de plusieurs centaines de sites. A l'issue du traitement de normalisation, tous ces sites normalisés peuvent être concaténés et regroupés en un seul corpus XML dont on donne ci-dessous la DTD puis l'arbre des éléments :

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v3.0 (http://www.xmlspy.com) by Serge Fleury (UniversitÃ© Paris 3) -->
<!--DTD generated by XML Spy v3.0 (http://www.xmlspy.com)-->
<!ELEMENT ADDRESS (#PCDATA)>
<!ELEMENT AVAILABILITY (#PCDATA)>
<!ATTLIST AVAILABILITY
       STATUS CDATA #REQUIRED
>
<!ELEMENT CHANGE (DATE, RESPSTMT, ITEMOND)>
<!ELEMENT CORPUSBODY (SITE+)>
<!ELEMENT CORPUSHEADER (FILEDESC, ENCODINGDESC, PROFILEDESC, REVISIONDESC)>
<!ELEMENT CORPUSTYPWEB (CORPUSHEADER, CORPUSBODY)>
<!ELEMENT CREATION (#PCDATA)>
<!ELEMENT DATE (#PCDATA)>
<!ELEMENT DATE2 (#PCDATA)>
<!ELEMENT DISTRIBUTOR (#PCDATA)>
<!ELEMENT EDITION (#PCDATA)>
<!ATTLIST EDITION
       N CDATA #REQUIRED
>
<!ELEMENT EDITIONSTMT (EDITION, DATE2)>
<!ELEMENT EDITORIALDECL (#PCDATA)>
<!ELEMENT ENCODINGDESC (PROJECTDESC, SAMPLINGDECL, EDITORIALDECL)>
<!ELEMENT EXTENT (#PCDATA)>
<!ELEMENT EXTImage (#PCDATA)>
<!ELEMENT FILEDESC (TITLSTMT, EDITIONSTMT, EXTENT, PUBLICATIONSTMT, SOURCEDESC)>
<!ELEMENT INTImage (#PCDATA)>
<!ELEMENT ITEMOND (#PCDATA)>
<!ELEMENT LANGUAGE (#PCDATA)>
<!ELEMENT NAME (#PCDATA)>
<!ELEMENT NAME2 (#PCDATA)>
<!ELEMENT PROFILEDESC (CREATION, LANGUAGE, TEXTCLASS)>
<!ELEMENT PROJECTDESC (#PCDATA)>
<!ELEMENT PUBLICATIONSTMT (DISTRIBUTOR, AVAILABILITY, ADDRESS)>
<!ELEMENT RESP (#PCDATA)>
<!ELEMENT RESP2 (#PCDATA)>
<!ELEMENT RESPSTMT (NAME, RESP)>
<!ELEMENT RESPSTMT2 (RESP2, NAME2)>
<!ELEMENT REVISIONDESC (CHANGE)>
<!ELEMENT SAMPLINGDECL (#PCDATA)>
<!ELEMENT SITE (SITEName, SITEWebDocumentDissected, SITEWebDocumentReports, SITEWebDocumentFound,
SITEWebDocumentImages, SITEWebDocumentMailTo, SITEWebDocumentUrls?, SITEWebDocumentAnchorFound,
SITEWebDocumentFileNotFound, SITEWebDocumentAnchorNotFound, SITEFile+)>
<!ELEMENT SITEFile (SITEFileName, SITEFileMeta, SITEFileStructure, SITEFileTxtBrut,
SITEFileTxtAndTagContent)>
<!ELEMENT SITEFileAuthor (#PCDATA)>
<!ELEMENT SITEFileContent (#PCDATA)>
<!ELEMENT SITEFileDescription (#PCDATA)>
<!ELEMENT SITEFileElements (SITEFileElementsNb, SiteFileElementDesc+)>
<!ELEMENT SITEFileElementsNb (#PCDATA)>
<!ELEMENT SITEFileExtHypertextualLinks (#PCDATA)>
<!ELEMENT SITEFileExternalAnchor (#PCDATA)>
<!ELEMENT SITEFileExternalCgiLinks (#PCDATA)>
<!ELEMENT SITEFileExternalFtpLinks (#PCDATA)>
<!ELEMENT SITEFileExternalGopherLinks (#PCDATA)>
<!ELEMENT SITEFileExternalLinks (#PCDATA)>
<!ELEMENT SITEFileExternalMail (#PCDATA)>
<!ELEMENT SITEFileExternalNewsLinks (#PCDATA)>
```

```
<!ELEMENT SITEFileGenerator (#PCDATA)>
<!ELEMENT SITEFileHtmlFileLinks (#PCDATA)>
<!ELEMENT SITEFileImageDesc (#PCDATA | EXTImage | INTImage)*>
<!ELEMENT SITEFileImageNb (#PCDATA)>
<!ELEMENT SITEFileIntDocFileLinks (#PCDATA)>
<!ELEMENT SITEFileIntHypertextualLinks (#PCDATA)>
<!ELEMENT SITEFileInternalAnchor (#PCDATA)>
<!ELEMENT SITEFileInternalLinks (#PCDATA)>
<!ELEMENT SITEFileKeywords (#PCDATA)>
<!ELEMENT SITEFileLinks (SITEFileLinksNumber, SITEFileExternalLinks, SITEFileInternalLinks,
SITEFileHtmlFileLinks, SITEFileExtHypertextualLinks, SITEFileIntHypertextualLinks,
SITEFileIntDocFileLinks, SITEFileExternalCgiLinks, SITEFileExternalNewsLinks,
SITEFileExternalFtpLinks, SITEFileExternalGopherLinks, SITEFileExternalMail, SITEFileInternalAnchor,
SITEFileExternalAnchor)>
<!ELEMENT SITEFileLinksNumber (#PCDATA)>
<!ELEMENT SITEFileMeta (SITEFileContent, SITEFileDescription, SITEFileGenerator, SITEFileKeywords,
SITEFileTitle, SITEFileAuthor)>
<!ELEMENT SITEFileName (#PCDATA)>
<!ELEMENT SITEFileStructure (SITEFileElements, SITEFileTxtElements, SITEFileImageNb,
SITEFileImageDesc, SITEFileLinks)>
<!ELEMENT SITEFileTitle (#PCDATA)>
<!ELEMENT SITEFileTxtAndTagContent (#PCDATA | tagHTML)*>
<!ELEMENT SITEFileTxtBrut (#PCDATA)>
<!ELEMENT SITEFileTxtElements (#PCDATA)>
<!ELEMENT SITEName (#PCDATA)>
<!ELEMENT SITEReportFile (#PCDATA)>
<!ATTLIST SITEReportFile
       NUM CDATA #REQUIRED
>
<!ELEMENT SITEWebDocumentAnchorFound (#PCDATA)>
<!ELEMENT SITEWebDocumentAnchorNotFound (#PCDATA)>
<!ELEMENT SITEWebDocumentDissected (#PCDATA)>
<!ELEMENT SITEWebDocumentFileNotFound (#PCDATA)>
<!ELEMENT SITEWebDocumentFound (#PCDATA)>
<!ELEMENT SITEWebDocumentImages (#PCDATA)>
<!ELEMENT SITEWebDocumentMailTo (#PCDATA)>
<!ELEMENT SITEWebDocumentReports (SITEReportFile+)>
<!ELEMENT SITEWebDocumentUrls (#PCDATA)>
<!ELEMENT SOURCEDESC (#PCDATA)>
<!ELEMENT SiteFileElementDesc (#PCDATA)>
<!ATTLIST SiteFileElementDesc
       type CDATA #REQUIRED
>
<!ELEMENT TEXTCLASS EMPTY>
<!ELEMENT TITLE (#PCDATA)>
<!ELEMENT TITLSTMT (TITLE, RESPSTMT2)>
<!ELEMENT tagHTML (#PCDATA)>
<!ATTLIST tagHTML
       type CDATA #REQUIRED
>
```

Figure 10 : DTD pour corpus Typweb



```
|    |    |
|    |    |__extent,
|    |    |   |_(#PCDATA)
|    |    |
|    |    |__publicationstmt,
|    |    |   |_(distributor,
|    |    |   |   |_(#PCDATA)
|    |    |   |
|    |    |   |__availability,
|    |    |   |   |_(#PCDATA)
|    |    |   |
|    |    |   |__address)
|    |    |       |_(#PCDATA)
|    |    |
|    |    |
|    |    |__sourcedesc)
|    |        |_(#PCDATA)
|    |
|    |
|    |__encodingdesc,
|    |   |_(projectdesc,
|    |   |   |_(#PCDATA)
|    |   |
|    |   |__samplingdecl,
|    |   |   |_(#PCDATA)
|    |   |
|    |   |__editorialdecl)
|    |       |_(#PCDATA)
|    |
|    |
|    |__profiledesc,
|    |   |_(creation,
|    |   |   |_(#PCDATA)
|    |   |
|    |   |__language,
|    |   |   |_(#PCDATA)
|    |   |
|    |   |__textclass)
|    |       |_EMPTY
|    |
|    |
|    |__revisiondesc)
|        |_(change)
|            |_(date,
|            |   |_(#PCDATA)
|            |
|            |__respstmt,
|            |   |_(name,
|            |   |   |_(#PCDATA)
|            |   |
|            |   |__resp)
|            |       |_(#PCDATA)
|            |
|            |
|            |__itemond)
|                |_(#PCDATA)
|
|
|
|
|__corpusbody)
    |_(site+)
        |_(sitename,
        |   |_(#PCDATA)
        |
        |__sitewebdocumentdissected,
        |   |_(#PCDATA)
        |
        |__sitewebdocumentreports,
        |   |_(sitereportfile+)
        |       |_(#PCDATA)
        |
        |
        |__sitewebdocumentfound,
        |   |_(#PCDATA)
        |
        |__sitewebdocumentimages,
        |   |_(#PCDATA)
        |
        |__sitewebdocumentmailto,
```

```
          |   |_(#PCDATA)
          |
          |__sitewebdocumenturls?,
          |   |_(#PCDATA)
          |
          |__sitewebdocumentanchorfound,
          |   |_(#PCDATA)
          |
          |__sitewebdocumentfilenotfound,
          |   |_(#PCDATA)
          |
          |__sitewebdocumentanchornotfound,
          |   |_(#PCDATA)
          |
          |__sitefile+)
              |_(sitefilename,
              |   |_(#PCDATA)
              |
              |__sitefilemeta,
              |   |_(sitefilecontent,
              |   |   |_(#PCDATA)
              |   |
              |   |__sitefiledescription,
              |   |   |_(#PCDATA)
              |   |
              |   |__sitefilegenerator,
              |   |   |_(#PCDATA)
              |   |
              |   |__sitefilekeywords,
              |   |   |_(#PCDATA)
              |   |
              |   |__sitefiletitle,
              |   |   |_(#PCDATA)
              |   |
              |   |__sitefileauthor)
              |       |_(#PCDATA)
              |
              |
              |__sitefilestructure,
              |   |_(sitefileelements,
              |   |   |_(sitefileelementsnb,
              |   |   |   |_(#PCDATA)
              |   |   |
              |   |   |__sitefileelementdesc+)
              |   |       |_(#PCDATA)
              |   |
              |   |
              |   |__sitefiletxtelements,
              |   |   |_(#PCDATA)
              |   |
              |   |__sitefileimagenb,
              |   |   |_(#PCDATA)
              |   |
              |   |__sitefileimagedesc,
              |   |   |_(#PCDATA |
              |   |   |__extimage |
              |   |   |   |_(#PCDATA)
              |   |   |
              |   |   |__intimage)*
              |   |       |_(#PCDATA)
              |   |
              |   |
              |   |__sitefilelinks)
              |       |_(sitefilelinksnumber,
              |       |   |_(#PCDATA)
              |       |
              |       |__sitefileexternallinks,
              |       |   |_(#PCDATA)
              |       |
              |       |__sitefileinternallinks,
              |       |   |_(#PCDATA)
              |       |
              |       |__sitefilehtmlfilelinks,
              |       |   |_(#PCDATA)
              |       |
              |       |__sitefileexthypertextuallinks,
              |       |   |_(#PCDATA)
              |       |
              |       |__sitefileinthypertextuallinks,
              |       |   |_(#PCDATA)
```

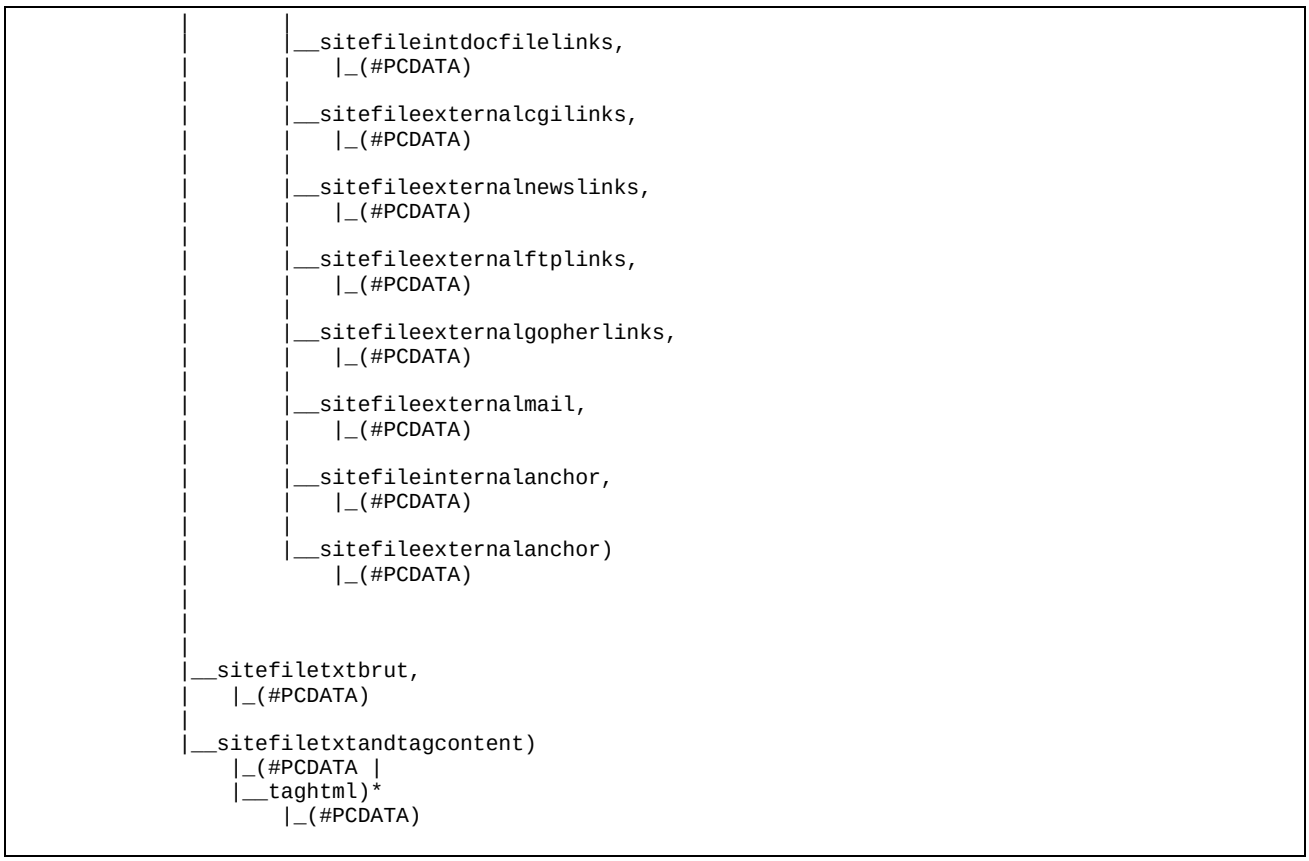


Figure 11 : arbre des éléments pour DTD corpus Typweb

Le schéma[4] général d'un site produit par cette chaîne est le suivant :

[4] Schéma produit par l'éditeur XML nommé XMLSpy (v. 3.5)

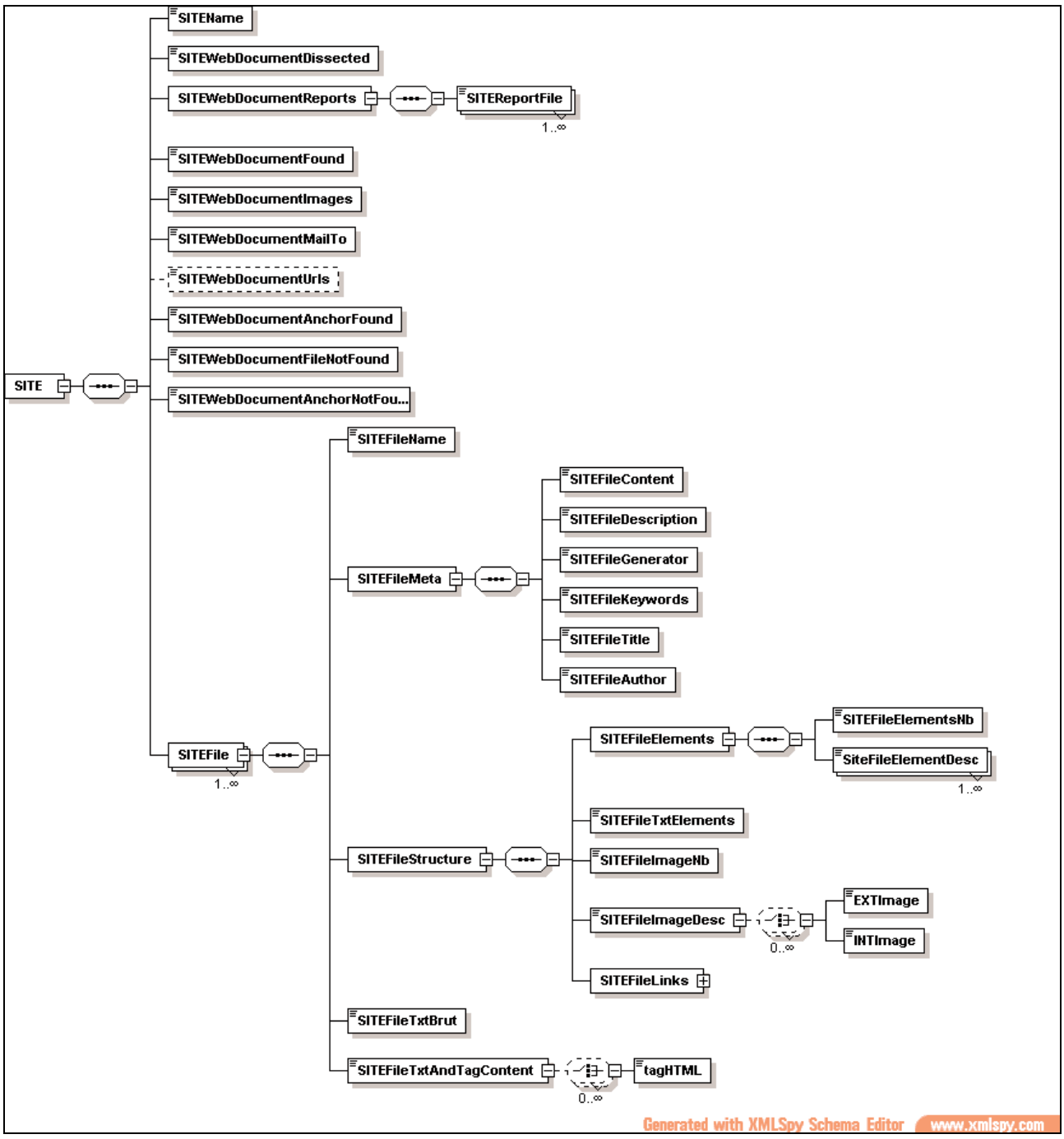


Figure 12 : schéma d'un corpus Typweb (un site)

Celui résultant de la concaténation de plusieurs sites normalisés est le suivant (un entête est intégré en début du fichier concaténé) :

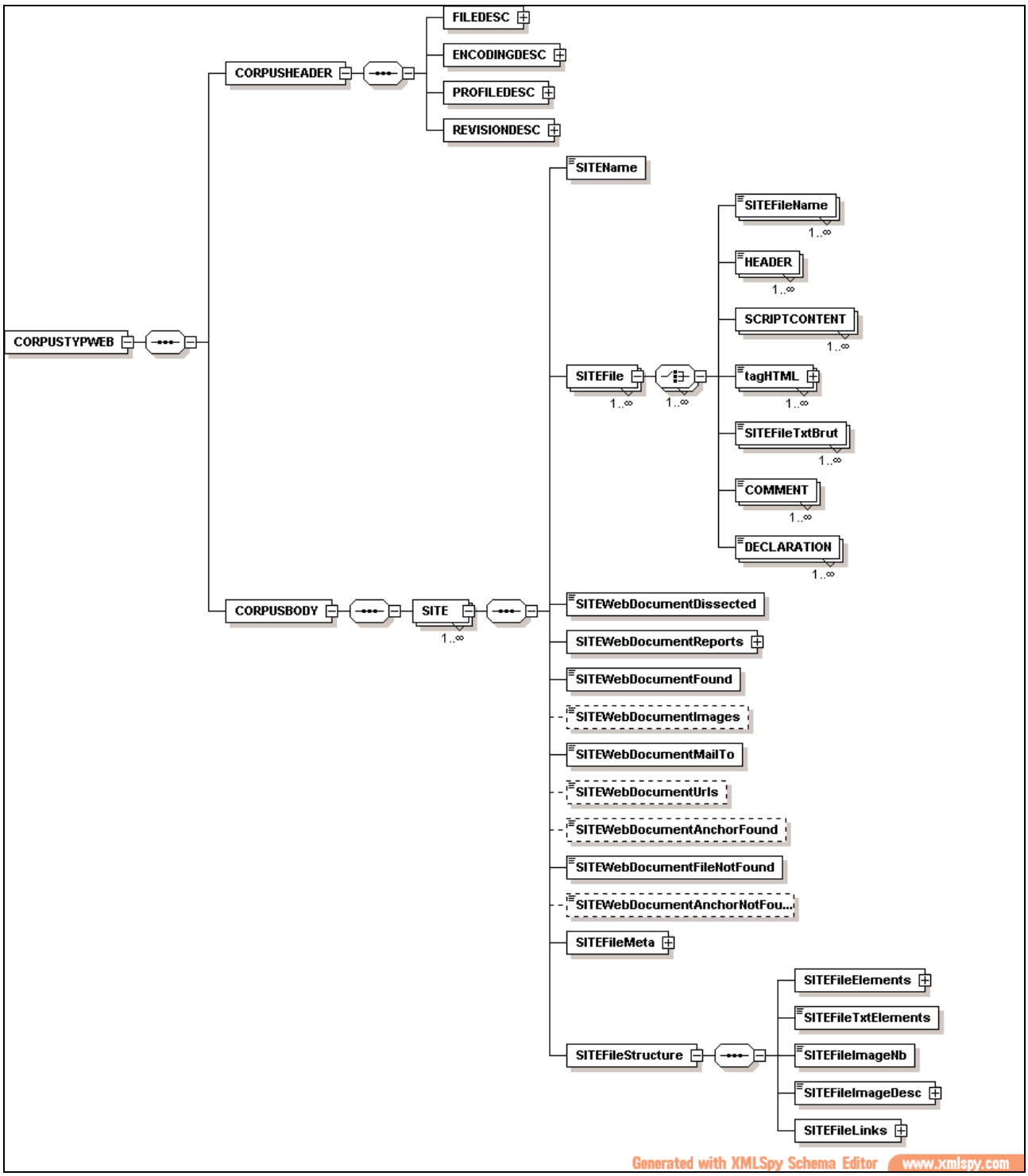


Figure 13 : schéma d'un corpus Typweb (un ensemble de sites)

5.2.1 Eléments textuels et structurels

Le corpus construit par `mktipo` contient dans des champs balisés associés à des informations représentant des données relatives aux éléments structurels et aux éléments textuels.

- Certains champs donnent des comptages sur les éléments HTML présents dans les pages traitées (liens, images…). De même le champ `sitefiletxtbrut` donne l'intégralité du texte contenu dans la page visée.
- Un champ supplémentaire vise à maintenir l'articulation entre les données textuelles de la page HTML initiale et les éléments structurels qui la composent. Le champ `sitefiletxtandtagcontent` donne en effet une présentation des portions de texte présents et des éléments HTML qui structurent la présentation de ces zones textuelles. Dans l'exemple ci-dessous, extrait d'un site construit par mktipo, on peut facilement isolé certaines zones

de texte placés entre certains marqueurs HTML du type "bold", "font", "a" (liens), ces zones sont soulignées à la main ici :

```
<SITEFileTxtAndTagContent>

<tagHTML type="HTML">begin_HTML</tagHTML>
<tagHTML type="HEAD">begin_HEAD</tagHTML>
<tagHTML type="TITLE">begin_TITLE</tagHTML>
<tagHTML type="META">begin_META</tagHTML>
<tagHTML type="META">begin_META</tagHTML>
<tagHTML type="META">begin_META</tagHTML>
<tagHTML type="SCRIPT">begin_SCRIPT</tagHTML>
<tagHTML type="script">end_script</tagHTML>
<tagHTML type="head">end_head</tagHTML>
<tagHTML type="BODY">begin_BODY</tagHTML>
<tagHTML type="CENTER">begin_CENTER</tagHTML>
<tagHTML type="A">begin_A</tagHTML>
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="TABLE">begin_TABLE</tagHTML>
<tagHTML type="TR">begin_TR</tagHTML>
<tagHTML type="TD">begin_TD</tagHTML>
<tagHTML type="P">begin_P</tagHTML>
<tagHTML type="FONT">begin_FONT</tagHTML>
<tagHTML type="IMG">begin_IMG</tagHTML>
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="td">end_td</tagHTML>
<tagHTML type="TD">begin_TD</tagHTML>
<tagHTML type="CENTER">begin_CENTER</tagHTML>
<tagHTML type="FONT">begin_FONT</tagHTML>
<tagHTML type="B">begin_B</tagHTML>
Bienvenue sur
<tagHTML type="b">end_b</tagHTML>
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="P">begin_P</tagHTML>
<tagHTML type="FONT">begin_FONT</tagHTML>
<tagHTML type="B">begin_B</tagHTML>

<tagHTML type="b">end_b</tagHTML>
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="A">begin_A</tagHTML>
<tagHTML type="FONT">begin_FONT</tagHTML>
<tagHTML type="B">begin_B</tagHTML>
<tagHTML type="IMG">begin_IMG</tagHTML>
<tagHTML type="b">end_b</tagHTML>
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="a">end_a</tagHTML>
<tagHTML type="P">begin_P</tagHTML>
<tagHTML type="FONT">begin_FONT</tagHTML>
<tagHTML type="B">begin_B</tagHTML>
le site des
 spéléologues jurassiens
<tagHTML type="b">end_b</tagHTML>
<tagHTML type="font">end_font</tagHTML>
<tagHTML type="P">begin_P</tagHTML>
<tagHTML type="FONT">begin_FONT</tagHTML>
<tagHTML type="I">begin_I</tagHTML>
Jura (France)
…
</SITEFileTxtAndTagContent>
```

Les corpus manipulés et construits tout au long de cette chaîne de traitements dans les phases d'expérimentation présentées *infra* ont conduit à ajouter une option de traitement au programme `mktipo`. On peut en effet choisir de produire un corpus qui ne contient que le zone regroupant les portions textuelles brutes ou la zone regroupant les zones textuelles accompagnées des marqueurs HTML.

5.3 ExtAndStatFrCorpTwp

Ce programme permet de générer des statistiques élémentaires sur les corpus issus de mktipo. On donne ci-dessous les résultats produits sur le site démo :

- On dispose tout d'abord d'informations sur l'intégralité du site visé :

```
----------
StatWord & element on corp-siteDemo.xml
----------
StatElement

Elt.                                    Freq
----                                    ----
meta                                      20
LinksNumber                                8
a                                          8
font                                       8
p                                          8
br                                         7
HtmlFileLinks                              5
head                                       4
html                                       4
body                                       4
title                                      4
ExternalLinks                              3
ImgNb(2)                                   2
img                                        2
ImgNb(1)                                   2
td                                         2
tr                                         1
ImgInterneNb                               1
table                                      1
ImgExterneNb                               1
ExternalGopherLinks                        0
IntHypertextualLinks                       0
InternalAnchor                             0
ExternalMail                               0
ExternalFtpLinks                           0
IntDocFileLinks                            0
ExtHypertextualLinks                       0
InternalLinks                              0
ExternalCgiLinks                           0
ExternalAnchor                             0
ExternalNewsLinks                          0

----------
Stat word

WORD                                   Freq.
----                                    ----
page                                       8
1                                          6
3                                          4
externe                                    4
welcome                                    4
revoir                                     3
la                                         3
lien                                       3
2                                          3
au                                         3
sur                                        3
image                                      2
surf                                       1
interne                                    1
bon                                        1
retour                                     1
vers                                       1
-----                                  -----
TOTAL                                  TOTAL
WORD                                  OCCURR
17                                        51
```

On dispose ensuite d'un description pour chaque page HTML du site :

```
----------
Stat Element for file : res1$siteDemo:index


meta                                        5
a                                           3
font                                        2
p                                           2
td                                          2
title                                       1
body                                        1
tr                                          1
head                                        1
table                                       1
br                                          1
html                                        1

-----
Stat Img for file : res1$siteDemo:index

ImgNb(1)                                    0
ImgNb(2)                                    0
ImgInterneNb                                0
ImgExterneNb                                0

-----
Stat Links for file : res1$siteDemo:index

LinksNumber                                 3
ExternalLinks                               0
InternalLinks                               0
HtmlFileLinks                               3
ExtHypertextualLinks                        0
IntHypertextualLinks                        0
IntDocFileLinks                             0
ExternalCgiLinks                            0
ExternalNewsLinks                           0
ExternalFtpLinks                            0
ExternalGopherLinks                         0
ExternalMail                                0
InternalAnchor                              0
ExternalAnchor                              0

----------
Stat Word for file : res1$siteDemo:index

WORD                                     Freq
----                                     ----
page                                        3
1                                           1
welcome                                     1
surf                                        1
bon                                         1
2                                           1
3                                           1
-----                                   -----
TOTAL                                   TOTAL
WORD                                   OCCURR
7                                           9

----------
Stat Element for file : res1$siteDemo:page2


meta                                        5
p                                           2
font                                        2
title                                       1
body                                        1
a                                           1
head                                        1
br                                          1
html                                        1

-----
Stat Img for file : res1$siteDemo:page2

ImgNb(1)                                    0
ImgNb(2)                                    0
ImgInterneNb                                0
ImgExterneNb                                0

-----
Stat Links for file : res1$siteDemo:page2

LinksNumber                                 1
ExternalLinks                               0
InternalLinks                               0
HtmlFileLinks                               1
ExtHypertextualLinks                        0
IntHypertextualLinks                        0
IntDocFileLinks                             0
ExternalCgiLinks                            0
ExternalNewsLinks                           0
ExternalFtpLinks                            0
ExternalGopherLinks                         0
ExternalMail                                0
InternalAnchor                              0
ExternalAnchor                              0

----------
Stat Word for file : res1$siteDemo:page2

WORD                                     Freq
----                                     ----
page                                        2
revoir                                      1
1                                           1
retour                                      1
2                                           1
welcome                                     1
au                                          1
sur                                         1
la                                          1
-----                                   -----
TOTAL                                   TOTAL
WORD                                   OCCURR
9                                          10

----------
Stat Element for file : res1$siteDemo:page1


br                                          5
meta                                        5
a                                           4
font                                        2
p                                           2
img                                         2
body                                        1
title                                       1
head                                        1
html                                        1

-----
Stat Img for file : res1$siteDemo:page1

ImgNb(1)                                    2
ImgNb(2)                                    2
ImgInterneNb                                1
ImgExterneNb                                1

-----
Stat Links for file : res1$siteDemo:page1

LinksNumber                                 4
ExternalLinks                               3
InternalLinks                               0
HtmlFileLinks                               1
ExtHypertextualLinks                        0
IntHypertextualLinks                        0
IntDocFileLinks                             0
ExternalCgiLinks                            0
ExternalNewsLinks                           0
ExternalFtpLinks                            0
```

```
ExternalGopherLinks                          0
ExternalMail                                 0
InternalAnchor                               0
ExternalAnchor                               0

----------
Stat Word for file : res1$siteDemo:page1

WORD                                      Freq
----                                      ----
externe                                      4
1                                            4
lien                                         3
image                                        2
page                                         2
3                                            2
revoir                                       1
interne                                      1
2                                            1
welcome                                      1
vers                                         1
au                                           1
la                                           1
sur                                          1
-----                                     -----
TOTAL                                     TOTAL
WORD                                     OCCURR
14                                           25

----------
Stat Element for file : res1$siteDemo:page3


meta                                         5
p                                            2
font                                         2
title                                        1
body                                         1
head                                         1
html                                         1

-----
```

```
Stat Img for file : res1$siteDemo:page3

ImgNb(1)                                     0
ImgNb(2)                                     0
ImgInterneNb                                 0
ImgExterneNb                                 0

-----
Stat Links for file : res1$siteDemo:page3

LinksNumber                                  0
ExternalLinks                                0
InternalLinks                                0
HtmlFileLinks                                0
ExtHypertextualLinks                         0
IntHypertextualLinks                         0
IntDocFileLinks                              0
ExternalCgiLinks                             0
ExternalNewsLinks                            0
ExternalFtpLinks                             0
ExternalGopherLinks                          0
ExternalMail                                 0
InternalAnchor                               0
ExternalAnchor                               0

----------
Stat Word for file : res1$siteDemo:page3

WORD                                      Freq
----                                      ----
revoir                                       1
page                                         1
3                                            1
welcome                                      1
au                                           1
la                                           1
sur                                          1
-----                                     -----
TOTAL                                     TOTAL
WORD                                     OCCURR
7                                             7
```

5.4 Webxref-038

Le programme `webxref-038` intégre les différents programmes présentés supra, il construit en sortie :

- un corpus XML (cf exemple infra)
- 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. On présente ci-dessous un exemple de ces sorties produites sur le site démo :

5.4.1 Corpus XML chaîne 038

```
<SITE>
<SITEName>siteDemo</SITEName>
<SITEFile>
<SITEFileName>C:/SFleury/Recherche/Typweb/siteDemo/index.htm</SITEFileName>
<SITETReportFileName>index.htm</SITEReportFileName>
<HEADER NUM="1">title: Bienvenue sur le site DEMOTYPWEB...</HEADER>
<HEADER NUM="2">meta http-equiv: Content-Type content: text/html; charset: iso-8859-1</HEADER>
<HEADER NUM="3">meta name: Author content: fleury@msh-paris.fr</HEADER>
<HEADER NUM="4">meta name: Description content: Un site de DEMO pour TypWeb : typologie de sites
web</HEADER>
<HEADER NUM="5">meta name: Keywords content: typologie, internet, trait, ...</HEADER>
<HEADER NUM="6">meta name: GENERATOR content: Mano</HEADER>
<tagHTML TAGType="html" NBATTR="0">BEGIN-html
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="head" NBATTR="0">BEGIN-head
</tagHTML>
```

```
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="3">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="charset" VALUE="iso-8859-1"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="content" VALUE="text/html; "/>
<tagHTMLAttr TAG="meta" NUM="3" ATTRType="http-equiv" VALUE="Content-Type"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="fleury@msh-paris.fr"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Author"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="Un site de DEMO pour TypWeb : typologie de
sites web"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Description"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="typologie, internet, trait, ..."/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Keywords"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="Mano"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="GENERATOR"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="title" NBATTR="0">BEGIN-title
</tagHTML>
<SITEFileTxtBrut>
Bienvenue sur le site DEMOTYPWEB...
</SITEFileTxtBrut>
<tagHTML TAGType="title">END-title</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="head">END-head</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body" NBATTR="2">BEGIN-body
<tagHTMLAttr TAG="body" NUM="1" ATTRType="bgcolor" VALUE="#FFFFFF"/>
<tagHTMLAttr TAG="body" NUM="2" ATTRType="bgproperties" VALUE="fixed"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="p" NBATTR="0">BEGIN-p
</tagHTML>
<tagHTML TAGType="font" NBATTR="1">BEGIN-font
<tagHTMLAttr TAG="font" NUM="1" ATTRType="size" VALUE="4"/>
</tagHTML>
<SITEFileTxtBrut>
Welcome...
</SITEFileTxtBrut>
<tagHTML TAGType="font">END-font</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="table" NBATTR="4">BEGIN-table
<tagHTMLAttr TAG="table" NUM="1" ATTRType="cellpadding" VALUE="0"/>
<tagHTMLAttr TAG="table" NUM="2" ATTRType="border" VALUE="0"/>
<tagHTMLAttr TAG="table" NUM="3" ATTRType="width" VALUE="100%"/>
<tagHTMLAttr TAG="table" NUM="4" ATTRType="cellspacing" VALUE="0"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="tr" NBATTR="0">BEGIN-tr
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="td" NBATTR="2">BEGIN-td
<tagHTMLAttr TAG="td" NUM="1" ATTRType="valign" VALUE="bottom"/>
<tagHTMLAttr TAG="td" NUM="2" ATTRType="rowspan" VALUE="2"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="a" NBATTR="1">BEGIN-a
<tagHTMLAttr TAG="a" NUM="1" ATTRType="href" VALUE="page1.htm"/>
</tagHTML>
<SITEFileTxtBrut>
Page 1
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="td">END-td</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="td" NBATTR="2">BEGIN-td
<tagHTMLAttr TAG="td" NUM="1" ATTRType="valign" VALUE="bottom"/>
<tagHTMLAttr TAG="td" NUM="2" ATTRType="nowrap" VALUE="nowrap"/>
```

```
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="a" NBATTR="1">BEGIN-a
<tagHTMLAttr TAG="a" NUM="1" ATTRType="href" VALUE="ss-dossier/page2.htm"/>
</tagHTML>
<SITEFileTxtBrut>
Page 2
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="td">END-td</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="tr">END-tr</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="table">END-table</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="a" NBATTR="1">BEGIN-a
<tagHTMLAttr TAG="a" NUM="1" ATTRType="href" VALUE="page3.htm"/>
</tagHTML>
<SITEFileTxtBrut>
Page 3
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="br" NBATTR="0">BEGIN-br
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="p" NBATTR="1">BEGIN-p
<tagHTMLAttr TAG="p" NUM="1" ATTRType="align" VALUE="right"/>
</tagHTML>
<tagHTML TAGType="font" NBATTR="1">BEGIN-font
<tagHTMLAttr TAG="font" NUM="1" ATTRType="size" VALUE="4"/>
</tagHTML>
<SITEFileTxtBrut>
...Bon surf !!
</SITEFileTxtBrut>
<tagHTML TAGType="font">END-font</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body">END-body</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="html">END-html</tagHTML>
<tagHTML  TAGType="LINK" NUM="1" TYPELink="INTERNAL_HTMLFILE" TAG="a"/>
<tagHTML  TAGType="LINK" NUM="2" TYPELink="INTERNAL_HTMLFILE" TAG="a"/>
<tagHTML  TAGType="LINK" NUM="3" TYPELink="INTERNAL_HTMLFILE" TAG="a"/>
</SITEFile>
<SITEFile>
<SITEFileName>C:/SFleury/Recherche/Typweb/siteDemo/ss-dossier/page2.htm</SITEFileName>
<SITETReportFileName>page2.htm</SITEReportFileName>
<HEADER NUM="1">title: Page 2...</HEADER>
<HEADER NUM="2">meta http-equiv: Content-Type content: text/html; charset: iso-8859-1</HEADER>
<HEADER NUM="3">meta name: Author content: fleury@msh-paris.fr</HEADER>
<HEADER NUM="4">meta name: Description content: Un site de DEMO pour TypWeb : typologie de sites
web</HEADER>
<HEADER NUM="5">meta name: Keywords content: typologie, internet, trait, ...</HEADER>
<HEADER NUM="6">meta name: GENERATOR content: Mano</HEADER>
<tagHTML TAGType="html" NBATTR="0">BEGIN-html
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="head" NBATTR="0">BEGIN-head
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="3">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="charset" VALUE="iso-8859-1"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="content" VALUE="text/html; "/>
<tagHTMLAttr TAG="meta" NUM="3" ATTRType="http-equiv" VALUE="Content-Type"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="fleury@msh-paris.fr"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Author"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="Un site de DEMO pour TypWeb : typologie de
sites web"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Description"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="typologie, internet, trait, ..."/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Keywords"/>
```

```
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="Mano"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="GENERATOR"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="title" NBATTR="0">BEGIN-title
</tagHTML>
<SITEFileTxtBrut>
Page 2...
</SITEFileTxtBrut>
<tagHTML TAGType="title">END-title</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="head">END-head</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body" NBATTR="2">BEGIN-body
<tagHTMLAttr TAG="body" NUM="1" ATTRType="bgcolor" VALUE="#FFFFFF"/>
<tagHTMLAttr TAG="body" NUM="2" ATTRType="bgproperties" VALUE="fixed"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="p" NBATTR="0">BEGIN-p
</tagHTML>
<tagHTML TAGType="font" NBATTR="1">BEGIN-font
<tagHTMLAttr TAG="font" NUM="1" ATTRType="size" VALUE="4"/>
</tagHTML>
<SITEFileTxtBrut>
Welcome sur la page 2...
</SITEFileTxtBrut>
<tagHTML TAGType="font">END-font</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="a" NBATTR="1">BEGIN-a
<tagHTMLAttr TAG="a" NUM="1" ATTRType="href" VALUE="../page1.htm"/>
</tagHTML>
<SITEFileTxtBrut>
retour page 1
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="br" NBATTR="0">BEGIN-br
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="p" NBATTR="1">BEGIN-p
<tagHTMLAttr TAG="p" NUM="1" ATTRType="align" VALUE="right"/>
</tagHTML>
<tagHTML TAGType="font" NBATTR="1">BEGIN-font
<tagHTMLAttr TAG="font" NUM="1" ATTRType="size" VALUE="4"/>
</tagHTML>
<SITEFileTxtBrut>
...Au revoir !!
</SITEFileTxtBrut>
<tagHTML TAGType="font">END-font</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body">END-body</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="html">END-html</tagHTML>
<tagHTML  TAGType="LINK" NUM="1" TYPELink="INTERNAL_HTMLFILE" TAG="a"/>
</SITEFile>
<SITEFile>
<SITEFileName>C:/SFleury/Recherche/Typweb/siteDemo/page1.htm</SITEFileName>
<SITETReportFileName>page1.htm</SITEReportFileName>
<HEADER NUM="1">title: Page 1...</HEADER>
<HEADER NUM="2">meta http-equiv: Content-Type content: text/html; charset: iso-8859-1</HEADER>
<HEADER NUM="3">meta name: Author content: fleury@msh-paris.fr</HEADER>
<HEADER NUM="4">meta name: Description content: Un site de DEMO pour TypWeb : typologie de sites
web</HEADER>
<HEADER NUM="5">meta name: Keywords content: typologie, internet, trait, ...</HEADER>
<HEADER NUM="6">meta name: GENERATOR content: Mano</HEADER>
<tagHTML TAGType="html" NBATTR="0">BEGIN-html
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="head" NBATTR="0">BEGIN-head
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="3">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="charset" VALUE="iso-8859-1"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="content" VALUE="text/html; "/>
<tagHTMLAttr TAG="meta" NUM="3" ATTRType="http-equiv" VALUE="Content-Type"/>
</tagHTML>
```

```
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="fleury@msh-paris.fr"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Author"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="Un site de DEMO pour TypWeb : typologie de
sites web"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Description"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="typologie, internet, trait, ..."/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Keywords"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="Mano"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="GENERATOR"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="title" NBATTR="0">BEGIN-title
</tagHTML>
<SITEFileTxtBrut>
Page 1...
</SITEFileTxtBrut>
<tagHTML TAGType="title">END-title</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="head">END-head</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body" NBATTR="2">BEGIN-body
<tagHTMLAttr TAG="body" NUM="1" ATTRType="bgcolor" VALUE="#FFFFFF"/>
<tagHTMLAttr TAG="body" NUM="2" ATTRType="bgproperties" VALUE="fixed"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="p" NBATTR="0">BEGIN-p
</tagHTML>
<tagHTML TAGType="font" NBATTR="1">BEGIN-font
<tagHTMLAttr TAG="font" NUM="1" ATTRType="size" VALUE="4"/>
</tagHTML>
<SITEFileTxtBrut>
Welcome sur la page 1...
</SITEFileTxtBrut>
<tagHTML TAGType="font">END-font</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="a" NBATTR="1">BEGIN-a
<tagHTMLAttr TAG="a" NUM="1" ATTRType="href" VALUE="http://www.netscape.fr/"/>
</tagHTML>
<SITEFileTxtBrut>
Lien externe 1
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="br" NBATTR="0">BEGIN-br
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="img" NBATTR="4">BEGIN-img
<tagHTMLAttr TAG="img" NUM="1" ATTRType="height" VALUE="31"/>
<tagHTMLAttr TAG="img" NUM="2" ATTRType="alt" VALUE="800*600"/>
<tagHTMLAttr TAG="img" NUM="3" ATTRType="src" VALUE="Images/800x600.gif"/>
<tagHTMLAttr TAG="img" NUM="4" ATTRType="width" VALUE="88"/>
</tagHTML>
<SITEFileTxtBrut>
Image interne 1
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="br" NBATTR="0">BEGIN-br
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="a" NBATTR="1">BEGIN-a
<tagHTMLAttr TAG="a" NUM="1" ATTRType="href" VALUE="http://www.microsoif.com/"/>
</tagHTML>
<tagHTML TAGType="img" NBATTR="4">BEGIN-img
<tagHTMLAttr TAG="img" NUM="1" ATTRType="height" VALUE="31"/>
<tagHTMLAttr TAG="img" NUM="2" ATTRType="alt" VALUE="FrontPage"/>
<tagHTMLAttr TAG="img" NUM="3" ATTRType="src" VALUE="http://www.microsoif.com/Images/fpcreate.gif"/>
<tagHTMLAttr TAG="img" NUM="4" ATTRType="width" VALUE="88"/>
</tagHTML>
<SITEFileTxtBrut>
```

```
Lien externe 2 + Image externe 1
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="br" NBATTR="0">BEGIN-br
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="a" NBATTR="1">BEGIN-a
<tagHTMLAttr TAG="a" NUM="1" ATTRType="href" VALUE="http://www.microsoft.com/france/"/>
</tagHTML>
<SITEFileTxtBrut>
Lien externe 3
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="br" NBATTR="0">BEGIN-br
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="a" NBATTR="1">BEGIN-a
<tagHTMLAttr TAG="a" NUM="1" ATTRType="href" VALUE="page3.htm"/>
</tagHTML>
<SITEFileTxtBrut>
vers page 3
</SITEFileTxtBrut>
<tagHTML TAGType="a">END-a</tagHTML>
<tagHTML TAGType="br" NBATTR="0">BEGIN-br
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="p" NBATTR="1">BEGIN-p
<tagHTMLAttr TAG="p" NUM="1" ATTRType="align" VALUE="right"/>
</tagHTML>
<tagHTML TAGType="font" NBATTR="1">BEGIN-font
<tagHTMLAttr TAG="font" NUM="1" ATTRType="size" VALUE="4"/>
</tagHTML>
<SITEFileTxtBrut>
...Au revoir !!
</SITEFileTxtBrut>
<tagHTML TAGType="font">END-font</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body">END-body</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="html">END-html</tagHTML>
<tagHTML  TAGType="LINK" NUM="1" TYPELink="EXTERNAL_HTTP" TAG="a"/>
<tagHTML  TAGType="LINK" NUM="2" TYPELink="EXTERNAL_HTTP" TAG="a"/>
<tagHTML  TAGType="LINK" NUM="3" TYPELink="INTERNAL_IMAGE" TAG="img"/>
<tagHTML  TAGType="LINK" NUM="4" TYPELink="EXTERNAL_HTTP" TAG="a"/>
<tagHTML  TAGType="LINK" NUM="5" TYPELink="INTERNAL_HTMLFILE" TAG="a"/>
<tagHTML  TAGType="LINK" NUM="6" TYPELink="EXTERNAL_IMAGE" TAG="img"/>
</SITEFile>
<SITEFile>
<SITEFileName>C:/SFleury/Recherche/Typweb/siteDemo/page3.htm</SITEFileName>
<SITETReportFileName>page3.htm</SITEReportFileName>
<HEADER NUM="1">title: Page 3...</HEADER>
<HEADER NUM="2">meta http-equiv: Content-Type content: text/html; charset: iso-8859-1</HEADER>
<HEADER NUM="3">meta name: Author content: fleury@msh-paris.fr</HEADER>
<HEADER NUM="4">meta name: Description content: Un site de DEMO pour TypWeb : typologie de sites
web</HEADER>
<HEADER NUM="5">meta name: Keywords content: typologie, internet, trait, ...</HEADER>
<HEADER NUM="6">meta name: GENERATOR content: Mano</HEADER>
<tagHTML TAGType="html" NBATTR="0">BEGIN-html
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="head" NBATTR="0">BEGIN-head
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="3">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="charset" VALUE="iso-8859-1"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="content" VALUE="text/html; "/>
<tagHTMLAttr TAG="meta" NUM="3" ATTRType="http-equiv" VALUE="Content-Type"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="fleury@msh-paris.fr"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Author"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="Un site de DEMO pour TypWeb : typologie de
sites web"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Description"/>
</tagHTML>
```

```
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="typologie, internet, trait, ..."/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="Keywords"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="meta" NBATTR="2">BEGIN-meta
<tagHTMLAttr TAG="meta" NUM="1" ATTRType="content" VALUE="Mano"/>
<tagHTMLAttr TAG="meta" NUM="2" ATTRType="name" VALUE="GENERATOR"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="title" NBATTR="0">BEGIN-title
</tagHTML>
<SITEFileTxtBrut>
Page 3...
</SITEFileTxtBrut>
<tagHTML TAGType="title">END-title</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="head">END-head</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body" NBATTR="2">BEGIN-body
<tagHTMLAttr TAG="body" NUM="1" ATTRType="bgcolor" VALUE="#FFFFFF"/>
<tagHTMLAttr TAG="body" NUM="2" ATTRType="bgproperties" VALUE="fixed"/>
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="p" NBATTR="0">BEGIN-p
</tagHTML>
<tagHTML TAGType="font" NBATTR="1">BEGIN-font
<tagHTMLAttr TAG="font" NUM="1" ATTRType="size" VALUE="4"/>
</tagHTML>
<SITEFileTxtBrut>
Welcome sur la page 3...
</SITEFileTxtBrut>
<tagHTML TAGType="font">END-font</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="p" NBATTR="1">BEGIN-p
<tagHTMLAttr TAG="p" NUM="1" ATTRType="align" VALUE="right"/>
</tagHTML>
<tagHTML TAGType="font" NBATTR="1">BEGIN-font
<tagHTMLAttr TAG="font" NUM="1" ATTRType="size" VALUE="4"/>
</tagHTML>
<SITEFileTxtBrut>
...Au revoir !!
</SITEFileTxtBrut>
<tagHTML TAGType="font">END-font</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body">END-body</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="html">END-html</tagHTML>
</SITEFile>
<SITEWebDocumentDissected> 4</SITEWebDocumentDissected>
<SITEWebDocumentReports>
<SITEReportFile NUM="1">index.html</SITEReportFile>
<SITEReportFile NUM="2">page2.html</SITEReportFile>
<SITEReportFile NUM="3">page1.html</SITEReportFile>
<SITEReportFile NUM="4">page3.html</SITEReportFile>
</SITEWebDocumentReports>
<SITEWebDocumentFound> 4</SITEWebDocumentFound>
<SITEWebDocumentUrls> 4</SITEWebDocumentUrls>
<SITEWebDocumentFileNotFound> 1</SITEWebDocumentFileNotFound>
<SITEFileMeta>
<SITEFileContent> text/html; charset: iso-8859-1</SITEFileContent>
<SITEFileDescription> Un site de DEMO pour TypWeb : typologie de sites web</SITEFileDescription>
<SITEFileGenerator> Mano</SITEFileGenerator>
<SITEFileKeywords> typologie, internet, trait, ...</SITEFileKeywords>
<SITEFileTitle> Page 3...</SITEFileTitle>
<SITEFileAuthor> fleury@msh-paris.fr</SITEFileAuthor>
</SITEFileMeta>
<SITEFileStructure>
<SITEFileElements>
<SITEFileElementsNb>73</SITEFileElementsNb>
</SITEFileElements>
<SITEFileTxtElements>88</SITEFileTxtElements>
<SITEFileImageNb>2</SITEFileImageNb>
<SITEFileImageDesc>2<EXTImage>1</EXTImage><INTImage>1</INTImage></SITEFileImageDesc>
<SITEFileLinks>
<SITEFileLinksNumber>8</SITEFileLinksNumber>
<SITEFileExternalLinks>3</SITEFileExternalLinks>
```

```
<SITEFileInternalLinks>0</SITEFileInternalLinks>
<SITEFileHtmlFileLinks>5</SITEFileHtmlFileLinks>
<SITEFileExtHypertextualLinks>0</SITEFileExtHypertextualLinks>
<SITEFileIntHypertextualLinks>0</SITEFileIntHypertextualLinks>
<SITEFileIntDocFileLinks>0</SITEFileIntDocFileLinks>
<SITEFileExternalCgiLinks>0</SITEFileExternalCgiLinks>
<SITEFileExternalNewsLinks>0</SITEFileExternalNewsLinks>
<SITEFileExternalFtpLinks>0</SITEFileExternalFtpLinks>
<SITEFileExternalGopherLinks>0</SITEFileExternalGopherLinks>
<SITEFileExternalMail>0</SITEFileExternalMail>
<SITEFileInternalAnchor>0</SITEFileInternalAnchor>
<SITEFileExternalAnchor>0</SITEFileExternalAnchor>
</SITEFileLinks>
</SITEFileStructure>
</SITE>
```

5.4.2 Corpus TXT chaîne 038

```
<SITEName>siteDemo</SITEName>
<SITEFile>
<SITEFileName>C:/SFleury/Recherche/Typweb/siteDemo/index.htm</SITEFileName>
       Bienvenue sur le site DEMOTYPWEB...   Welcome...    Page 1  Page 2   Page 3 ...Bon surf !!
</SITEFile>
<SITEFile>
<SITEFileName>C:/SFleury/Recherche/Typweb/siteDemo/ss-dossier/page2.htm</SITEFileName>
       Page 2...   Welcome sur la page 2... retour page 1 ...Au revoir !!
</SITEFile>
<SITEFile>
<SITEFileName>C:/SFleury/Recherche/Typweb/siteDemo/page1.htm</SITEFileName>
       Page 1...   Welcome sur la page 1... Lien externe 1 Image interne 1 Lien externe 2 + Image
externe 1 Lien externe 3 vers page 3 ...Au revoir !!
</SITEFile>
<SITEFile>
<SITEFileName>C:/SFleury/Recherche/Typweb/siteDemo/page3.htm</SITEFileName>
       Page 3...   Welcome sur la page 3... ...Au revoir !!
</SITEFile>
</SITE>
```

5.4.3 Etat statistique par page chaîne 038

```
<TAGS>
<SITE>siteDemo</SITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/index.htm</PAGE>
<ELEMENTS>
<ITEM>META</ITEM><FRQ>5</FRQ>
<ITEM>A</ITEM><FRQ>3</FRQ>
<ITEM>INTERNAL (HTMLFILE)</ITEM><FRQ>3</FRQ>
<ITEM>TD</ITEM><FRQ>2</FRQ>
<ITEM>P</ITEM><FRQ>2</FRQ>
<ITEM>FONT</ITEM><FRQ>2</FRQ>
<ITEM>BR</ITEM><FRQ>1</FRQ>
<ITEM>HTML</ITEM><FRQ>1</FRQ>
<ITEM>HEAD</ITEM><FRQ>1</FRQ>
<ITEM>TITLE</ITEM><FRQ>1</FRQ>
<ITEM>TABLE</ITEM><FRQ>1</FRQ>
<ITEM>BODY</ITEM><FRQ>1</FRQ>
<ITEM>TR</ITEM><FRQ>1</FRQ>
</ELEMENTS>
<ELEMENTS_ATTR>
<ITEM>META(CONTENT)</ITEM><FRQ>5</FRQ>
<ITEM>META(NAME)</ITEM><FRQ>4</FRQ>
<ITEM>A(HREF)</ITEM><FRQ>3</FRQ>
<ITEM>FONT(SIZE)</ITEM><FRQ>2</FRQ>
<ITEM>TD(VALIGN)</ITEM><FRQ>2</FRQ>
<ITEM>BODY(BGCOLOR)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(CELLSPACING)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(BORDER)</ITEM><FRQ>1</FRQ>
<ITEM>P(ALIGN)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(CELLPADDING)</ITEM><FRQ>1</FRQ>
<ITEM>META(HTTP-EQUIV)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(WIDTH)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGPROPERTIES)</ITEM><FRQ>1</FRQ>
<ITEM>META(CHARSET)</ITEM><FRQ>1</FRQ>
<ITEM>TD(ROWSPAN)</ITEM><FRQ>1</FRQ>
```

```
<ITEM>TD(NOWRAP)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTR>
<ELEMENTS_ATTRVALUE>
<ITEM>TD(VALIGN=BOTTOM)</ITEM><FRQ>2</FRQ>
<ITEM>FONT(SIZE=4)</ITEM><FRQ>2</FRQ>
<ITEM>TD(ROWSPAN=2)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=AUTHOR)</ITEM><FRQ>1</FRQ>
<ITEM>META(HTTP-EQUIV=CONTENT-TYPE)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(WIDTH=100%)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(CELLSPACING=0)</ITEM><FRQ>1</FRQ>
<ITEM>P(ALIGN=RIGHT)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=SS-DOSSIER/PAGE2.HTM)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=TYPOLOGIE, INTERNET, TRAIT, ...)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGPROPERTIES=FIXED)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=PAGE3.HTM)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=UN SITE DE DEMO POUR TYPWEB : TYPOLOGIE DE SITES WEB)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(BORDER=0)</ITEM><FRQ>1</FRQ>
<ITEM>META(CHARSET=ISO-8859-1)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=GENERATOR)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=KEYWORDS)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=MANO)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=DESCRIPTION)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=TEXT/HTML; )</ITEM><FRQ>1</FRQ>
<ITEM>TD(NOWRAP=NOWRAP)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(CELLPADDING=0)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=PAGE1.HTM)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=FLEURY@MSH-PARIS.FR)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGCOLOR=#FFFFFF)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTRVALUE>
</TAGS>
<WORDS>
<SITE>siteDemo</SITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/index.htm</PAGE>
<ITEM>page</ITEM><FRQ>3</FRQ>
<ITEM>site</ITEM><FRQ>1</FRQ>
<ITEM>1</ITEM><FRQ>1</FRQ>
<ITEM>surf</ITEM><FRQ>1</FRQ>
<ITEM>2</ITEM><FRQ>1</FRQ>
<ITEM>le</ITEM><FRQ>1</FRQ>
<ITEM>3</ITEM><FRQ>1</FRQ>
<ITEM>bienvenue</ITEM><FRQ>1</FRQ>
<ITEM>welcome</ITEM><FRQ>1</FRQ>
<ITEM>bon</ITEM><FRQ>1</FRQ>
<ITEM>demotypweb</ITEM><FRQ>1</FRQ>
<ITEM>sur</ITEM><FRQ>1</FRQ>
<TOTALFORM>12</TOTALFORM>
<TOTALOCCUR>14</TOTALOCCUR>
</WORDS>
<TAGS>
<SITE>siteDemo</SITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/ss-dossier/page2.htm</PAGE>
<ELEMENTS>
<ITEM>META</ITEM><FRQ>5</FRQ>
<ITEM>P</ITEM><FRQ>2</FRQ>
<ITEM>FONT</ITEM><FRQ>2</FRQ>
<ITEM>INTERNAL (HTMLFILE)</ITEM><FRQ>1</FRQ>
<ITEM>BR</ITEM><FRQ>1</FRQ>
<ITEM>HTML</ITEM><FRQ>1</FRQ>
<ITEM>TITLE</ITEM><FRQ>1</FRQ>
<ITEM>HEAD</ITEM><FRQ>1</FRQ>
<ITEM>A</ITEM><FRQ>1</FRQ>
<ITEM>BODY</ITEM><FRQ>1</FRQ>
</ELEMENTS>
<ELEMENTS_ATTR>
<ITEM>META(CONTENT)</ITEM><FRQ>5</FRQ>
<ITEM>META(NAME)</ITEM><FRQ>4</FRQ>
<ITEM>FONT(SIZE)</ITEM><FRQ>2</FRQ>
<ITEM>BODY(BGCOLOR)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF)</ITEM><FRQ>1</FRQ>
<ITEM>P(ALIGN)</ITEM><FRQ>1</FRQ>
<ITEM>META(HTTP-EQUIV)</ITEM><FRQ>1</FRQ>
<ITEM>META(CHARSET)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGPROPERTIES)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTR>
<ELEMENTS_ATTRVALUE>
<ITEM>FONT(SIZE=4)</ITEM><FRQ>2</FRQ>
<ITEM>META(NAME=AUTHOR)</ITEM><FRQ>1</FRQ>
<ITEM>META(HTTP-EQUIV=CONTENT-TYPE)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=../PAGE1.HTM)</ITEM><FRQ>1</FRQ>
<ITEM>P(ALIGN=RIGHT)</ITEM><FRQ>1</FRQ>
```

```
<ITEM>META(CONTENT=TYPOLOGIE, INTERNET, TRAIT, ...)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGPROPERTIES=FIXED)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=UN SITE DE DEMO POUR TYPWEB : TYPOLOGIE DE SITES WEB)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=GENERATOR)</ITEM><FRQ>1</FRQ>
<ITEM>META(CHARSET=ISO-8859-1)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=MANO)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=KEYWORDS)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=DESCRIPTION)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=TEXT/HTML; )</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=FLEURY@MSH-PARIS.FR)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGCOLOR=#FFFFFF)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTRVALUE>
</TAGS>
<WORDS>
<SITE>siteDemo</SITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/ss-dossier/page2.htm</PAGE>
<ITEM>page</ITEM><FRQ>3</FRQ>
<ITEM>2</ITEM><FRQ>2</FRQ>
<ITEM>revoir</ITEM><FRQ>1</FRQ>
<ITEM>1</ITEM><FRQ>1</FRQ>
<ITEM>retour</ITEM><FRQ>1</FRQ>
<ITEM>welcome</ITEM><FRQ>1</FRQ>
<ITEM>au</ITEM><FRQ>1</FRQ>
<ITEM>la</ITEM><FRQ>1</FRQ>
<ITEM>sur</ITEM><FRQ>1</FRQ>
<TOTALFORM>9</TOTALFORM>
<TOTALOCCUR>12</TOTALOCCUR>
</WORDS>
<TAGS>
<SITE>siteDemo</SITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/page1.htm</PAGE>
<ELEMENTS>
<ITEM>META</ITEM><FRQ>5</FRQ>
<ITEM>BR</ITEM><FRQ>5</FRQ>
<ITEM>A</ITEM><FRQ>4</FRQ>
<ITEM>EXTERNAL (HTTP)</ITEM><FRQ>3</FRQ>
<ITEM>IMG</ITEM><FRQ>2</FRQ>
<ITEM>P</ITEM><FRQ>2</FRQ>
<ITEM>FONT</ITEM><FRQ>2</FRQ>
<ITEM>HTML</ITEM><FRQ>1</FRQ>
<ITEM>TITLE</ITEM><FRQ>1</FRQ>
<ITEM>HEAD</ITEM><FRQ>1</FRQ>
<ITEM>INTERNAL (HTMLFILE)</ITEM><FRQ>1</FRQ>
<ITEM>INTERNAL (IMAGE)</ITEM><FRQ>1</FRQ>
<ITEM>BODY</ITEM><FRQ>1</FRQ>
</ELEMENTS>
<ELEMENTS_ATTR>
<ITEM>META(CONTENT)</ITEM><FRQ>5</FRQ>
<ITEM>META(NAME)</ITEM><FRQ>4</FRQ>
<ITEM>A(HREF)</ITEM><FRQ>4</FRQ>
<ITEM>IMG(WIDTH)</ITEM><FRQ>2</FRQ>
<ITEM>FONT(SIZE)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(ALT)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(HEIGHT)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(SRC)</ITEM><FRQ>2</FRQ>
<ITEM>BODY(BGPROPERTIES)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGCOLOR)</ITEM><FRQ>1</FRQ>
<ITEM>META(CHARSET)</ITEM><FRQ>1</FRQ>
<ITEM>META(HTTP-EQUIV)</ITEM><FRQ>1</FRQ>
<ITEM>P(ALIGN)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTR>
<ELEMENTS_ATTRVALUE>
<ITEM>IMG(WIDTH=88)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(HEIGHT=31)</ITEM><FRQ>2</FRQ>
<ITEM>FONT(SIZE=4)</ITEM><FRQ>2</FRQ>
<ITEM>A(HREF=HTTP://WWW.MICROSOFT.COM/FRANCE/)</ITEM><FRQ>1</FRQ>
<ITEM>P(ALIGN=RIGHT)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=TYPOLOGIE, INTERNET, TRAIT, ...)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGPROPERTIES=FIXED)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=PAGE3.HTM)</ITEM><FRQ>1</FRQ>
<ITEM>IMG(ALT=800*600)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=UN SITE DE DEMO POUR TYPWEB : TYPOLOGIE DE SITES WEB)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=GENERATOR)</ITEM><FRQ>1</FRQ>
<ITEM>META(CHARSET=ISO-8859-1)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=MANO)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=KEYWORDS)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=FLEURY@MSH-PARIS.FR)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=AUTHOR)</ITEM><FRQ>1</FRQ>
<ITEM>IMG(ALT=FRONTPAGE)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=DESCRIPTION)</ITEM><FRQ>1</FRQ>
```

```
<ITEM>IMG(SRC=IMAGES/800X600.GIF)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=TEXT/HTML; )</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=HTTP://WWW.NETSCAPE.FR/)</ITEM><FRQ>1</FRQ>
<ITEM>IMG(SRC=HTTP://WWW.MICROSOIF.COM/IMAGES/FPCREATE.GIF)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=HTTP://WWW.MICROSOIF.COM/)</ITEM><FRQ>1</FRQ>
<ITEM>META(HTTP-EQUIV=CONTENT-TYPE)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGCOLOR=#FFFFFF)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTRVALUE>
</TAGS>
<WORDS>
<SITE>siteDemo</SITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/page1.htm</PAGE>
<ITEM>1</ITEM><FRQ>5</FRQ>
<ITEM>externe</ITEM><FRQ>4</FRQ>
<ITEM>lien</ITEM><FRQ>3</FRQ>
<ITEM>page</ITEM><FRQ>3</FRQ>
<ITEM>image</ITEM><FRQ>2</FRQ>
<ITEM>3</ITEM><FRQ>2</FRQ>
<ITEM>revoir</ITEM><FRQ>1</FRQ>
<ITEM>interne</ITEM><FRQ>1</FRQ>
<ITEM>2</ITEM><FRQ>1</FRQ>
<ITEM>welcome</ITEM><FRQ>1</FRQ>
<ITEM>vers</ITEM><FRQ>1</FRQ>
<ITEM>au</ITEM><FRQ>1</FRQ>
<ITEM>la</ITEM><FRQ>1</FRQ>
<ITEM>sur</ITEM><FRQ>1</FRQ>
<TOTALFORM>14</TOTALFORM>
<TOTALOCCUR>27</TOTALOCCUR>
</WORDS>
<TAGS>
<SITE>siteDemo</SITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/page3.htm</PAGE>
<ELEMENTS>
<ITEM>META</ITEM><FRQ>5</FRQ>
<ITEM>P</ITEM><FRQ>2</FRQ>
<ITEM>FONT</ITEM><FRQ>2</FRQ>
<ITEM>TITLE</ITEM><FRQ>1</FRQ>
<ITEM>HEAD</ITEM><FRQ>1</FRQ>
<ITEM>HTML</ITEM><FRQ>1</FRQ>
<ITEM>BODY</ITEM><FRQ>1</FRQ>
</ELEMENTS>
<ELEMENTS_ATTR>
<ITEM>META(CONTENT)</ITEM><FRQ>5</FRQ>
<ITEM>META(NAME)</ITEM><FRQ>4</FRQ>
<ITEM>FONT(SIZE)</ITEM><FRQ>2</FRQ>
<ITEM>BODY(BGPROPERTIES)</ITEM><FRQ>1</FRQ>
<ITEM>P(ALIGN)</ITEM><FRQ>1</FRQ>
<ITEM>META(HTTP-EQUIV)</ITEM><FRQ>1</FRQ>
<ITEM>META(CHARSET)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGCOLOR)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTR>
<ELEMENTS_ATTRVALUE>
<ITEM>FONT(SIZE=4)</ITEM><FRQ>2</FRQ>
<ITEM>META(NAME=AUTHOR)</ITEM><FRQ>1</FRQ>
<ITEM>META(HTTP-EQUIV=CONTENT-TYPE)</ITEM><FRQ>1</FRQ>
<ITEM>P(ALIGN=RIGHT)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=TYPOLOGIE, INTERNET, TRAIT, ...)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGPROPERTIES=FIXED)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=UN SITE DE DEMO POUR TYPWEB : TYPOLOGIE DE SITES WEB)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=GENERATOR)</ITEM><FRQ>1</FRQ>
<ITEM>META(CHARSET=ISO-8859-1)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=MANO)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=KEYWORDS)</ITEM><FRQ>1</FRQ>
<ITEM>META(NAME=DESCRIPTION)</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=TEXT/HTML; )</ITEM><FRQ>1</FRQ>
<ITEM>META(CONTENT=FLEURY@MSH-PARIS.FR)</ITEM><FRQ>1</FRQ>
<ITEM>BODY(BGCOLOR=#FFFFFF)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTRVALUE>
</TAGS>
<WORDS>
<SITE>siteDemo</SITE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/page3.htm</PAGE>
<ITEM>3</ITEM><FRQ>2</FRQ>
<ITEM>page</ITEM><FRQ>2</FRQ>
<ITEM>revoir</ITEM><FRQ>1</FRQ>
<ITEM>welcome</ITEM><FRQ>1</FRQ>
<ITEM>au</ITEM><FRQ>1</FRQ>
<ITEM>la</ITEM><FRQ>1</FRQ>
<ITEM>sur</ITEM><FRQ>1</FRQ>
<TOTALFORM>7</TOTALFORM>
```

```
<TOTALOCCUR>9</TOTALOCCUR>
</WORDS>
```

5.4.4 Etat statistique global chaîne 038

```
<StatWordElement site="siteDemo">
<TAGS>
<ELEMENTS>
<ITEM>META</ITEM><FRQ>20</FRQ>
<ITEM>FONT</ITEM><FRQ>8</FRQ>
<ITEM>P</ITEM><FRQ>8</FRQ>
<ITEM>A</ITEM><FRQ>8</FRQ>
<ITEM>BR</ITEM><FRQ>7</FRQ>
<ITEM>INTERNAL (HTMLFILE)</ITEM><FRQ>5</FRQ>
<ITEM>HTML</ITEM><FRQ>4</FRQ>
<ITEM>HEAD</ITEM><FRQ>4</FRQ>
<ITEM>TITLE</ITEM><FRQ>4</FRQ>
<ITEM>BODY</ITEM><FRQ>4</FRQ>
<ITEM>EXTERNAL (HTTP)</ITEM><FRQ>3</FRQ>
<ITEM>TD</ITEM><FRQ>2</FRQ>
<ITEM>IMG</ITEM><FRQ>2</FRQ>
<ITEM>TABLE</ITEM><FRQ>1</FRQ>
<ITEM>INTERNAL (IMAGE)</ITEM><FRQ>1</FRQ>
<ITEM>EXTERNAL (IMAGE)</ITEM><FRQ>1</FRQ>
<ITEM>TR</ITEM><FRQ>1</FRQ>
</ELEMENTS>
<ELEMENTS_ATTR>
<ITEM>META(CONTENT)</ITEM><FRQ>20</FRQ>
<ITEM>META(NAME)</ITEM><FRQ>16</FRQ>
<ITEM>A(HREF)</ITEM><FRQ>8</FRQ>
<ITEM>FONT(SIZE)</ITEM><FRQ>8</FRQ>
<ITEM>META(CHARSET)</ITEM><FRQ>4</FRQ>
<ITEM>P(ALIGN)</ITEM><FRQ>4</FRQ>
<ITEM>META(HTTP-EQUIV)</ITEM><FRQ>4</FRQ>
<ITEM>BODY(BGPROPERTIES)</ITEM><FRQ>4</FRQ>
<ITEM>BODY(BGCOLOR)</ITEM><FRQ>4</FRQ>
<ITEM>IMG(WIDTH)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(SRC)</ITEM><FRQ>2</FRQ>
<ITEM>TD(VALIGN)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(ALT)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(HEIGHT)</ITEM><FRQ>2</FRQ>
<ITEM>TD(NOWRAP)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(CELLPADDING)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(CELLSPACING)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(WIDTH)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(BORDER)</ITEM><FRQ>1</FRQ>
<ITEM>TD(ROWSPAN)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTR>
<ELEMENTS_ATTRVALUE>
<ITEM>FONT(SIZE=4)</ITEM><FRQ>8</FRQ>
<ITEM>META(CONTENT=FLEURY@MSH-PARIS.FR)</ITEM><FRQ>4</FRQ>
<ITEM>META(HTTP-EQUIV=CONTENT-TYPE)</ITEM><FRQ>4</FRQ>
<ITEM>META(CONTENT=TEXT/HTML; )</ITEM><FRQ>4</FRQ>
<ITEM>BODY(BGCOLOR=#FFFFFF)</ITEM><FRQ>4</FRQ>
<ITEM>BODY(BGPROPERTIES=FIXED)</ITEM><FRQ>4</FRQ>
<ITEM>META(NAME=AUTHOR)</ITEM><FRQ>4</FRQ>
<ITEM>META(NAME=DESCRIPTION)</ITEM><FRQ>4</FRQ>
<ITEM>META(NAME=KEYWORDS)</ITEM><FRQ>4</FRQ>
<ITEM>META(CONTENT=TYPOLOGIE, INTERNET, TRAIT, ...)</ITEM><FRQ>4</FRQ>
<ITEM>P(ALIGN=RIGHT)</ITEM><FRQ>4</FRQ>
<ITEM>META(CONTENT=UN SITE DE DEMO POUR TYPWEB : TYPOLOGIE DE SITES WEB)</ITEM><FRQ>4</FRQ>
<ITEM>META(NAME=GENERATOR)</ITEM><FRQ>4</FRQ>
<ITEM>META(CHARSET=ISO-8859-1)</ITEM><FRQ>4</FRQ>
<ITEM>META(CONTENT=MANO)</ITEM><FRQ>4</FRQ>
<ITEM>A(HREF=PAGE3.HTM)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(WIDTH=88)</ITEM><FRQ>2</FRQ>
<ITEM>IMG(HEIGHT=31)</ITEM><FRQ>2</FRQ>
<ITEM>TD(VALIGN=BOTTOM)</ITEM><FRQ>2</FRQ>
<ITEM>A(HREF=HTTP://WWW.MICROSOFT.COM/FRANCE/)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(WIDTH=100%)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(CELLSPACING=0)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=../PAGE1.HTM)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=SS-DOSSIER/PAGE2.HTM)</ITEM><FRQ>1</FRQ>
<ITEM>IMG(ALT=800*600)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(BORDER=0)</ITEM><FRQ>1</FRQ>
<ITEM>IMG(ALT=FRONTPAGE)</ITEM><FRQ>1</FRQ>
```

```
<ITEM>IMG(SRC=IMAGES/800X600.GIF)</ITEM><FRQ>1</FRQ>
<ITEM>TD(NOWRAP=NOWRAP)</ITEM><FRQ>1</FRQ>
<ITEM>IMG(SRC=HTTP://WWW.MICROSOIF.COM/IMAGES/FPCREATE.GIF)</ITEM><FRQ>1</FRQ>
<ITEM>TABLE(CELLPADDING=0)</ITEM><FRQ>1</FRQ>
<ITEM>TD(ROWSPAN=2)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=HTTP://WWW.NETSCAPE.FR/)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=PAGE1.HTM)</ITEM><FRQ>1</FRQ>
<ITEM>A(HREF=HTTP://WWW.MICROSOIF.COM/)</ITEM><FRQ>1</FRQ>
</ELEMENTS_ATTRVALUE>
</TAGS>
<WORDS>
<ITEM>page</ITEM><FRQ>11</FRQ>
<ITEM>1</ITEM><FRQ>7</FRQ>
<ITEM>3</ITEM><FRQ>5</FRQ>
<ITEM>welcome</ITEM><FRQ>4</FRQ>
<ITEM>2</ITEM><FRQ>4</FRQ>
<ITEM>sur</ITEM><FRQ>4</FRQ>
<ITEM>externe</ITEM><FRQ>4</FRQ>
<ITEM>au</ITEM><FRQ>3</FRQ>
<ITEM>revoir</ITEM><FRQ>3</FRQ>
<ITEM>la</ITEM><FRQ>3</FRQ>
<ITEM>lien</ITEM><FRQ>3</FRQ>
<ITEM>image</ITEM><FRQ>2</FRQ>
<ITEM>site</ITEM><FRQ>1</FRQ>
<ITEM>surf</ITEM><FRQ>1</FRQ>
<ITEM>retour</ITEM><FRQ>1</FRQ>
<ITEM>interne</ITEM><FRQ>1</FRQ>
<ITEM>bienvenue</ITEM><FRQ>1</FRQ>
<ITEM>bon</ITEM><FRQ>1</FRQ>
<ITEM>demotypweb</ITEM><FRQ>1</FRQ>
<ITEM>le</ITEM><FRQ>1</FRQ>
<ITEM>vers</ITEM><FRQ>1</FRQ>
<TOTALFORM>21</TOTALFORM>
<TOTALOCCUR>62</TOTALOCCUR>
</WORDS>
</StatWordElement>
```

Les éléments pris en compte dans ces états statistiques ont été balisés pour permettre ,en aval, de filtrer facilement ces informations.

5.4.5 Schéma du corpus XML 038

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

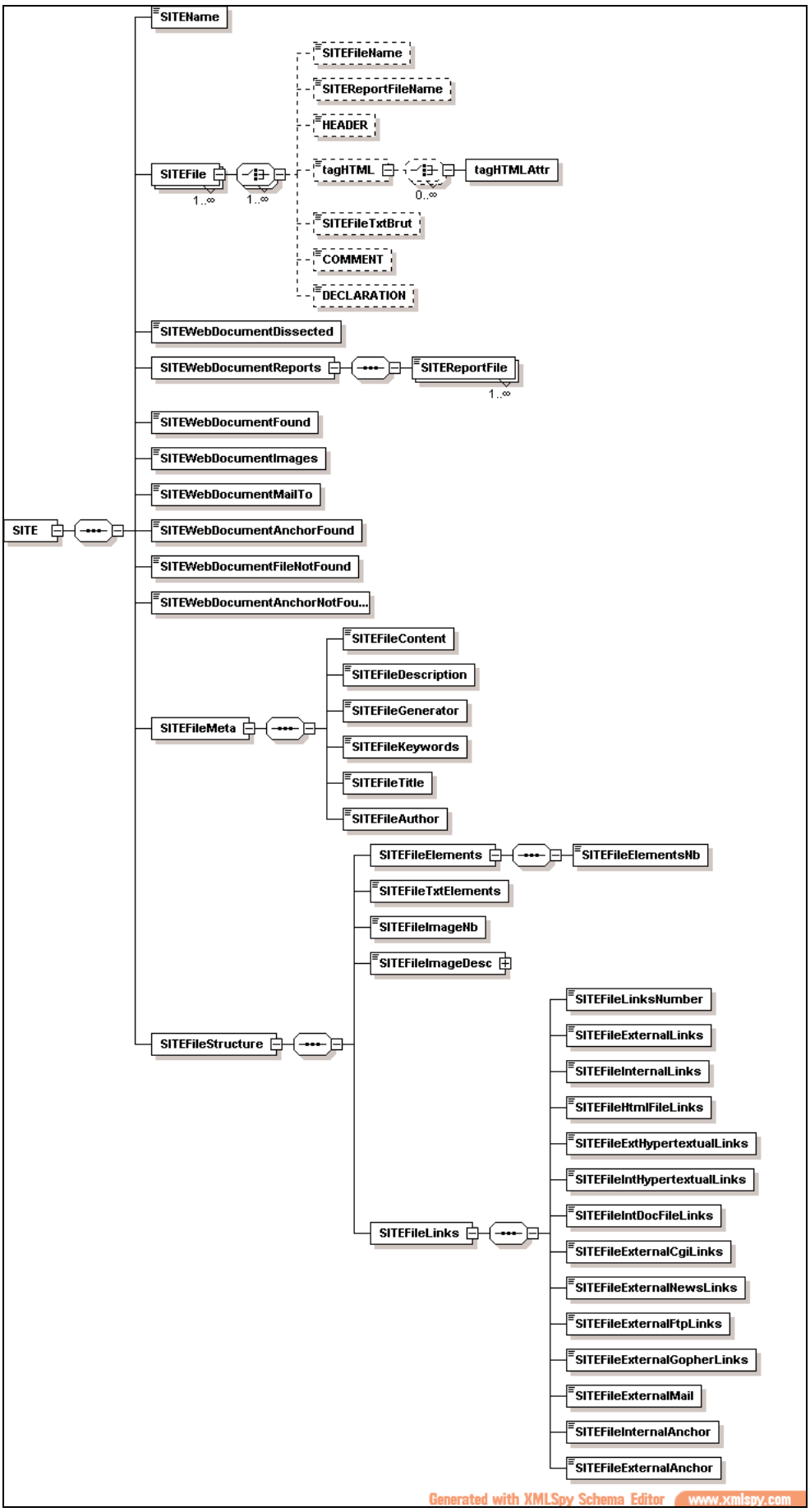


Figure 14 : schéma d'un corpus Typweb (un site)

5.4.6 Développements en cours

Le programme `webxref` présenté ci-dessus a été modifié pour permettre 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 "opérationnelle" de webxref : 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/i
ndex.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/site
Demo/page1.htm
    2.
file://localhost/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/site
Demo/ss-dossier/page2.htm
    3.
file://localhost/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/site
Demo/page3.htm

</DUMPTEXT>
</FILE>

<FILE>
<FILENAME>/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/siteDemo/s
s-dossier/page2.htm</FILENAME>
<DUMPTEXT>

    Welcome sur la page 2...
    [1]retour page 1
```

```
                                                                ...Au revoir !!

 References

    1.
file://localhost/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/site
Demo/page1.htm

</DUMPTEXT>
</FILE>

<FILE>
<FILENAME>/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/siteDemo/p
age1.htm</FILENAME>
<DUMPTEXT>

    Welcome sur la page 1...
    [1]Lien externe 1
    800*600 Image interne 1
    [2]FrontPage Lien externe 2 + Image externe 1
    [3]Lien externe 3
    [4]vers page 3

                                                                ...Au revoir !!

 References

    1. http://www.netscape.fr/
    2. http://www.microsoif.com/
    3. http://www.microsoft.com/france/
    4.
file://localhost/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/site
Demo/page3.htm

</DUMPTEXT>
</FILE>

<FILE>
<FILENAME>/windows/C/SFleury/Programmes/perl/perlTk/MkCorpus/sitesWeb/siteDemo/p
age3.htm</FILENAME>
<DUMPTEXT>

    Welcome sur la page 3...

                                                                ...Au revoir !!

</DUMPTEXT>
</FILE>

</DUMPLYNX>
```

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

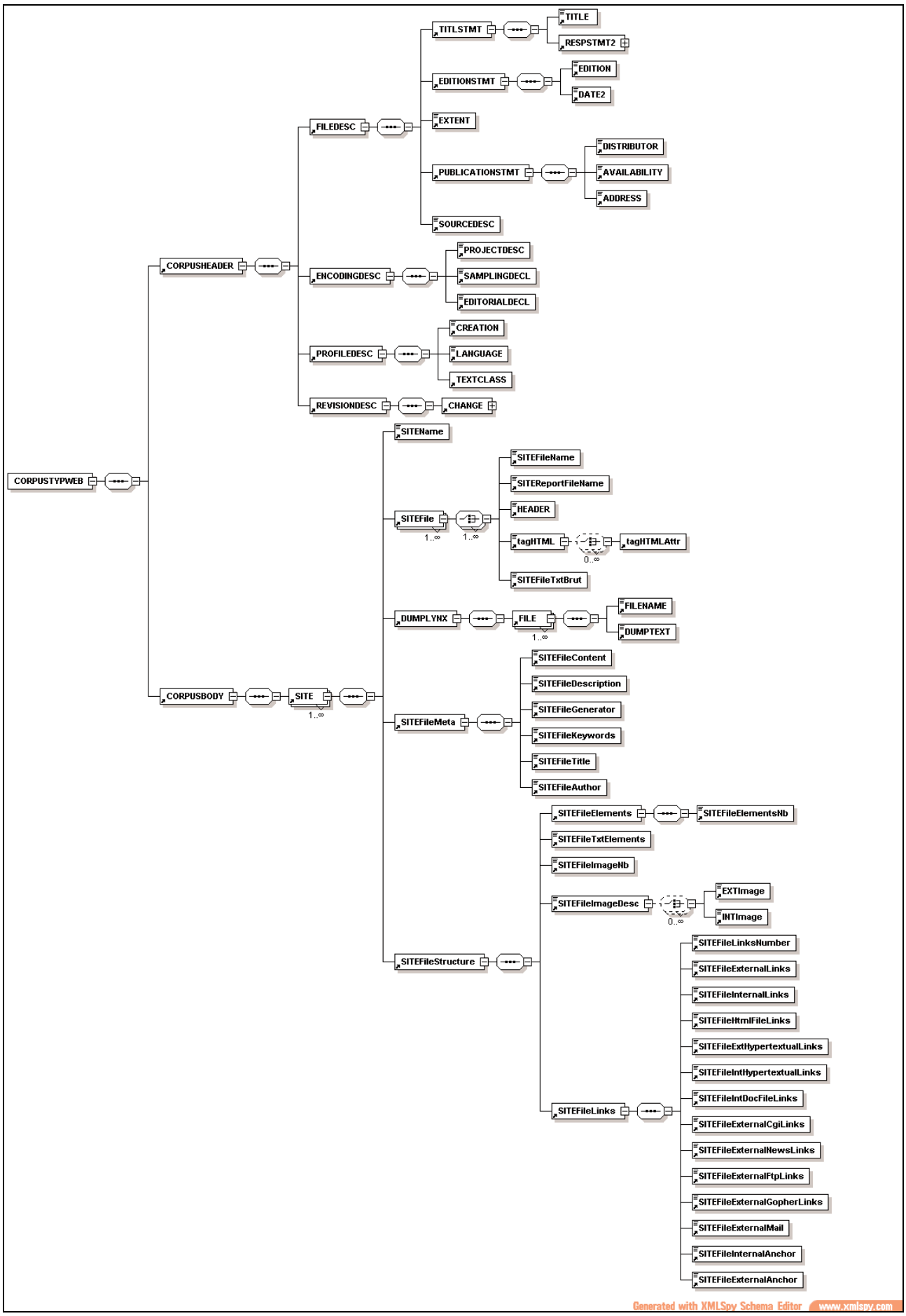


Figure 14bis : schéma d'un corpus Typweb avec lynx

5.5 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…).
- Les outils développés dans le cadre du projet Typweb ont été intégrés à cet interface : les deux chaînes 036 et 038 sont disponibles dans un menu de cet outil.

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. On trouve à l'adresse suivante une documentation provisoire sur cet outil : http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/mkcorpusProject.htm

6 Expérimentations des outils

6.1 Expérience n°1

Une première expérimentation a été menée par Serge Fleury au mois de juillet 2000 sur 445 sites correspondant à des sites personnels d'abonnés à Wanadoo. Cette expérimentation avait pour but de mettre à l'épreuve la démarche et les outils construits. Elle a aussi permet d'affiner la mise au point des outils utilisés.

La chaîne de traitements présentée supra a été reproduite intégralement et a produit le corpus complet normalisé et les différents états présentés.

Nombre de sites	445
Nombres de pages	10 659
Nombre moyen de pages par site	23,95
Nombre de mots/occurrences	146 580 / 3 798 841
Nombre moyen de mots/occurrences par site	329 / 8 536
Nombre moyen de mots/occurrences par page	13 / 356
Nombre de liens	66 471
Nombre de liens internes (fichiers)	40 993
Nombre de liens externes	11 259

6.2 Expérience n°2

Une deuxième expérimentation a été menée au mois de juillet et août 2000 par Aude Maisondieu et Andréa Kuncova sur l'ensemble des sites. Cette expérimentation avait là encore pour but de mettre à l'épreuve la démarche et les outils construits. Le rapport établi par les deux stagiaires donne une présentation complète de cette expérimentation. On donne ci-dessous les résultats quantitatifs sur les traitements réalisés.

Nombre de sites	534
Nombres de pages	11 129
Nombre moyen de pages par site	20,84
Nombre de mots/occurrences	150 316 / 3 862 198
Nombre moyen de mots/occurrences par site	281 / 7 232
Nombre moyen de mots/occurrences par page	13 / 347
Nombre de liens	69 932
Nombre de liens internes (fichiers)	43 046
Nombre de liens externes	11 433

6.2.1 Aspiration de site et constitution du corpus

- *pages persos :*

"Un premier corpus a été constitué à partir des pages personnelles hébergées par Wanadoo. Les aspirations ont été lancées à partir du site leader (jura.speleo), puis complétées par des aspirations aléatoires.
Au cours de notre stage, nous avons constitué un second corpus de pages personnelles. La stratégie d'aspiration retenue est différente puisque l'on tient compte des parcours des utilisateurs. En effet, nous avons aspirés tous les sites personnels hébergés par le fournisseur d'accès Wanadoo visités au cours du mois de mars 2000. Cela représente environ 580 sites personnels".

- *sites marchands :*

"Les sites aspirés correspondent aux entretiens menés dans le laboratoire (entretiens auprès des plus importants sites marchands français et de leurs prestataires de technologies, pour le commerce des biens tangibles). Nous avons poursuivi les aspirations des sites marchands déjà en cours en tenant compte des entretiens ayant déjà eu lieu ou programmés pour les mois à venir. Nous avons ainsi aspiré 40 sites marchands".

6.2.2 Test des outils

Aucun problème n'est survenu avec Webxref sur les 584 sites personnels aspirés.
Mktipo a quant à lui eu des problèmes sur les sites suivants :

sites	Rapport webxref en entrée	Descriptif du problème
http://perso.wanadoo.fr/antoine.flahault	//rep1<index>.html	page à redirection
http://perso.wanadoo.fr/bourse.graphic	//rep1<index>.html	page à redirection
http://perso.wanadoo.fr/dancy.sound	//rep164<here-i-go-again>.html	la source de la page est complète mais la page ne s'affiche pas
http://perso.wanadoo.fr/gisèle.bartaland	//rep18<prevention.educnat>.html	page à redirection
http://perso.wanadoo.fr/guillaume.blanc	//rep82<sommaire>.html	
http://perso.wanadoo.fr/j-françois.thiers	//rep10<alien3>.html	
http://perso.wanadoo.fr/jeux.lulu	//rep1<index>.html	
http://perso.wanadoo.fr/kiss07	//rep1<index>.html	
http://perso.wanadoo.fr/midinet	//rep4<accueil>.html	
http://perso.wanadoo.fr/nguyen595	//rep1<index>.html	page à redirection
http://perso.wanadoo.fr/olivier.lalorette	//rep15<menu>.html	
http://perso.wanadoo.fr/sahira	//rep1<index>.html	
http://perso.wanadoo.fr/ultimage3d	//rep5<basdepage>.html	
http://perso.wanadoo.fr/velonin	//rep1<index>.html	

6.3 Expérience n°3

Expériences de Marie Pasquier.

7 Etats d'avancement du projet Typweb LOT1

7.1 Petite synthèse sur les travaux actuels : du corpus aux matrices

7.1.1 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 (appelée aussi *supra* `webxref_VersionTypweb`) : 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.

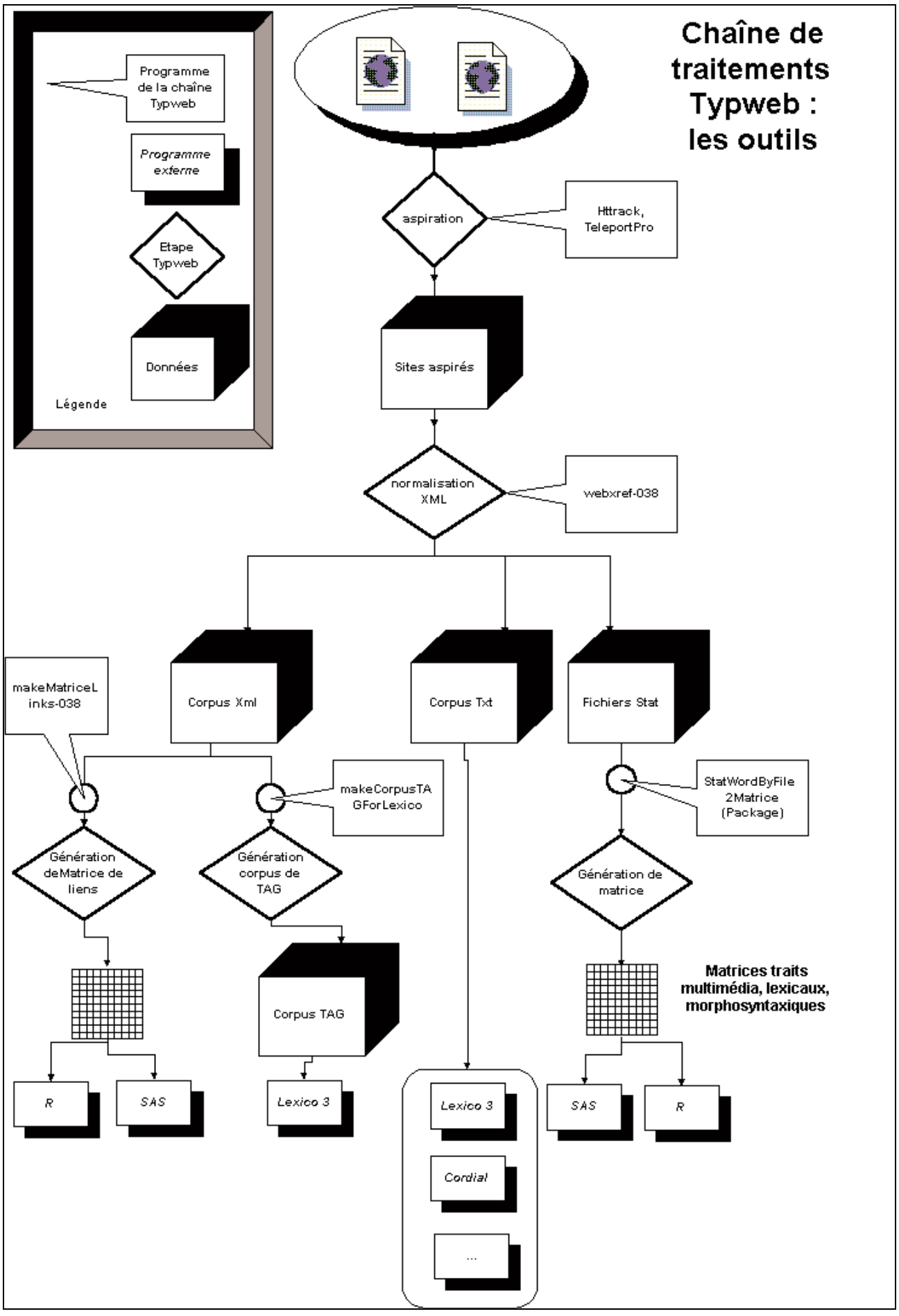


Figure 15 : Chaîne des outils Typweb

7.2 Constitution des corpus

Ce travail est en cours via les outils présentés *supra* : environ 500 sites de pages personnelles et une centaine de sites commerciaux.

7.3 Constitution des matrices

7.3.1 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)

7.3.2 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.

7.3.3 Construction des matrices

EtiquetageTyPWeb : Etiquetage du corpus TyPWeb

7.3.3.1 Documentation

7.3.3.1.1 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*
```

7.3.3.1.2 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.

7.3.3.1.3 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

7.3.3.2 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/>).

7.3.3.3 . 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").

7.3.3.4 Filtrage a posteriori d'une matrice

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

7.3.3.4.1 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.

7.3.3.4.2 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]

7.3.3.5 . 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.

7.3.3.6 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 0 -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      O
alain.bosmans_new       3318    14513   1.8     1125.3  alain.bosmans_new       O
alain.cf_new    139     285     0.0     29.5    alain.cf_new    O
alain.dubus_new 250     697     0.1     77.4    alain.dubus_new O

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      O
alain.bosmans_new       3318    14513   1.8     1125.3  alain.bosmans_new       O
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 O
```

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     O
A(HREF) 4       205     51.2    105.3   _aaaaqb_        417     O
```

```
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
```

7.3.3.7 A voir

- Refaire les essais ci-dessus à partir de SitesAlain.StatWordByFile.txt (fourni dans l'archive) pour vérifier si toutes les étapes aboutissent aux mêmes résultats.
- A priori, TOUS les fichiers sont correctement engendrés par StatWordByFile2Matrice.pl. Vérifier si c'est le cas sous Windows. Si ce n'est pas le cas, deux solutions, au moins : a) problème de place (hypothèse déjà faite) ; b) des ouvertures de fichiers intermédiaires qui ne s'effectuent pas correctement.
- Les appels à rm et mv dans StatWordByFile2Matrice.pl ont été remplacés par des commandes Perl. Les appels à system ont aussi été modifiés en fonction de tes précédentes indications. Voir si cela fonctionne correctement.

7.3.3.7.1 A faire

- Ecrire les documentations en utilisant les outils Perl pour associer la documentation au code.
- Tests
- Lire tous les paramètres communs dans un seul fichier pour éviter les risques de divergences.
- Outil de regroupement de traits, par exemple pour ramener :

```
<CHOIX/>A(NAME)   _aaaavf_
<CHOIX/>A(NAME= BANDES DESSINéES)  _aaaavg_
<CHOIX/>A(NAME=1) ETYMOLOGIE.)     _aaaavh_
<CHOIX/>A(NAME=1) LA RELIGION.)    _aaaavi_
<CHOIX/>A(NAME=1) RECENSEMENT SUCCINCT DES RéC)_aaaavj_
<CHOIX/>A(NAME=2) L'HISTOIRE.)     _aaaavk_
<CHOIX/>A(NAME=2) LES COMPOSANTS DU MYTHE.)    _aaaavl_
<CHOIX/>A(NAME=2) LES DéVELOPPEMENTS HISTORIQU)_aaaavm_
<CHOIX/>A(NAME=3) CLASSEMENT.)     _aaaavn_
<CHOIX/>A(NAME=3) L'ARCHéTYPE.)    _aaaavo_
<CHOIX/>A(NAME=3) MYTHES ET CROYANCES ASSOCIéS)_aaaavp_
<CHOIX/>A(NAME=4) LA LITTéRATURE.) _aaaavq_
<CHOIX/>A(NAME=A) LES INVARIANTS.) _aaaavr_
<CHOIX/>A(NAME=ARTS PLASTIQUES)    _aaaavs_
<CHOIX/>A(NAME=B) LES MYTHèMES.)   _aaaavt_
<CHOIX/>A(NAME=BIBLIOGRAPHIE)_aaaavu_
<CHOIX/>A(NAME=C) LES ATTRIBUTS.)  _aaaavv_
<CHOIX/>A(NAME=CINéMA)  _aaaavw_
<CHOIX/>A(NAME=CONCLUSION)    _aaaavx_
<CHOIX/>A(NAME=II - LES LEGENDES DES AMAZONES) _aaaavy_
<CHOIX/>A(NAME=III - VERS UNE TYPOLOGIE LITTER)_aaaavz_
<CHOIX/>A(NAME=INTRODUCTION )_aaaawa_
<CHOIX/>A(NAME=MYTHE1)  _aaaawb_
<CHOIX/>A(NAME=OPéRA)   _aaaawc_
```

```
<CHOIX/>A(NAME=REMERCIEMENTS) _aaaawd_
<CHOIX/>A(NAME=TABLE)   _aaaawe_
```

à

<CHOIX/>A(NAME=)

ce qui va réduire la dispersion des traits (y compris le non-respect de la syntaxe même d'HTML, signalé par SF).

Outil de vérification de la cohérence entre les données de départ et les occurrences des traits dans les matrices.

Lien entre les traits et les documents dans lesquels ils figurent et leur localisation exacte. D'après Helka Folch, ce n'est pas une bonne idée de regénérer les pages .html avec des ancres devant chaque élément et un index vers ces ancres.

Transformer en un trait le nom de domaine qui permet de connaître l'hébergeur.

7.3.4 Remarques, bugs et problèmes

1. Traitement du nom des pages

Ce format générique est produit par les 2 versions avec les variantes suivantes :

Dans la version 036, le programme ExtAndStat produit un index des fichiers traités avec indexation des noms de fichier du type siteX$fileY associé à un fichier précis. C'est cette référence unique qui est utilisée dans la description précédente notée "pagename" et ce pour éviter des pbs de conflit de noms de page.

Dans la version 038, c'est le nom physique complet qui se retrouve dans le champ <PAGE>.

2. Traitements des attributs

Seule la version 038 intègre des fréquences pour les attributs, le fichier de stat de la version 036 ne contiendra donc que des champs vides du type :

```
<ELEMENTS_ATTR>
</ELEMENTS_ATTR>
```

3. Installation : vérifier que le chemin de Perl localement est bien celui qui est indiqué dans la première ligne de ces programmes, sinon changer cette première ligne (quand on veut des scripts directement exécutables).
4. Vérifier si les programmes de profilage et d'émondage de profil ne saturent pas la mémoire allouée à/par Perl en cas de très grosse matrice de départ
5. XMLiser toutes les étapes
6. Faire une vérification de la cohérence de la matrice par rapport aux données de départ et de la cohérence d'une matrice après émondage (par sondages).
7. Questions à Serge Fleury : Les noms de site/page ne sont-ils pas trop longs et alors aussi à changer comme les noms de traits ? Que faire pour faciliter le passage à Windows, en termes d'entrées/sorties principalement ? Une bonne partie des noms créés est paramétrée. Les noms longs (très) que j'emploie posent-ils problèmes ? Il faudra une petite moulinette in fine d'ajout d'un \r aux matrices une fois émondées pour permettre le travail ultérieur sous Windows.

7.3.5 Programmes annexes

7.3.5.1 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>

7.3.5.2 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    Fréq    Lien    Fréq    Lien    Fréq    Lien    Fréq
                Interne         Externe         Mail            Ftp
```

Des traitements statistiques sur ces données sont en cours.

Exemple de sorties produites sur le site Démo :

```
<SITE>
<NAME>siteDemo</NAME>
<LINKSBYPAGE>
<PAGE>C:/SFleury/Recherche/Typweb/siteDemo/index.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/page1.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/page3.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>
```

7.3.5.3 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>
```

7.3.5.4 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[5] 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…).

7.3.6 Points à traiter

- Utilisation de Cordial pour les marquages morpho-syntaxiques et autres
- Marquages sémantiques : quels traits retenir ?
- Marquages multimédia : lesquels ?

Cette étape correspond à l'adaptation de l'architecture à construire s'inspire et adapte celle mise en œuvre pour TyPText. On dispose au départ d'une base de sites. Chaque site comprend un « cartouche » documentaire donnant des indications sur les sites visés. Cette étape se décompose en plusieurs phases :

- Création de sous-corpus de sites sur la base de critères à déterminer.
- Etiquetage morpho-syntaxique générique des sites traités : intervention possible de DSM/GRI avec TLT.
- Définition des traits spécifiques à prendre en compte pour l'analyse des sites : ces traits peuvent décrire les contenus formels ou textuels des sites à analyser.
- Marquage des sites sur la base des indicateurs dont on veut étudier la distribution.
- Sélection des traits à analyser et constitution d'une matrice dans laquelle chaque site est représenté par un vecteur de traits : la matrice sert tant à la recherche optimale de traits pertinents à une opposition, qu'à la classification inductive ou supervisée.

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

8 Traitements statistiques sur les matrices

Ce travail est en cours de réalisation sur la base des programmes présentés supra.

- Traitements statistiques sur les matrices produites par TyPWeb : indicateurs de la structure formelle des sites (liens internes, externes, profondeur, articulation texte-image, sons, vidéos…)
- Constitution d'une typologie des contenus : utilisation d'index thématique prédéfini ou bien constitution de catégories de contenu (démarche inductive propre à l'architecture).
- Croisement de la structure formelle et des typologies de contenus

8.1 Données traitées

Des matrices ont été produites pour 4 corpus.

- corpus Page Perso Eté 99 (noté PPEte99)

Nombre de sites	
Nombres de pages	
Nombre moyen de pages par site	
Nombre de mots/occurrences	
Nombre moyen de mots/occurrences par site	
Nombre moyen de mots/occurrences par page	
Nombre de liens	
Nombre de liens internes (fichiers)	
Nombre de liens externes	

- corpus Page Perso Mars 2000 (noté PPMars2000)

Nombre de sites	
Nombres de pages	
Nombre moyen de pages par site	
Nombre de mots/occurrences	
Nombre moyen de mots/occurrences par site	
Nombre moyen de mots/occurrences par page	
Nombre de liens	
Nombre de liens internes (fichiers)	
Nombre de liens externes	

- corpus Sites Marchand n°1 (Noté SM1)

Nombre de sites	
Nombres de pages	
Nombre moyen de pages par site	
Nombre de mots/occurrences	
Nombre moyen de mots/occurrences par site	
Nombre moyen de mots/occurrences par page	
Nombre de liens	
Nombre de liens internes (fichiers)	
Nombre de liens externes	

- corpus Sites Marchand n°2 (Noté SM2)

Nombre de sites	
Nombres de pages	
Nombre moyen de pages par site	

Nombre de mots/occurrences	
Nombre moyen de mots/occurrences par site	
Nombre moyen de mots/occurrences par page	
Nombre de liens	
Nombre de liens internes (fichiers)	
Nombre de liens externes	

8.2 Matrices textuelles

8.3 Matrices de TAGs HTML

8.4 Matrice de liens

9 Références bibliographiques

Habert Benoît, Salem André, Nazarenko Adeline (1998) : *Les linguistiques de corpus*, Armand Colin, Paris.

Habert, B., Fabre, C., Issac, F. (1998) : *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*, Masson, Paris.

Typologie de sites Web

Valérie Beaudouin**, Serge Fleury*, Julia Velkovska**, « Analyse des espaces de communication sur internet et intranet, de l'illusion de la transparence », JADT'2000, Lausanne.
* UMR8503 CNRS/ENS Fontenay Saint-Cloud, France
** CNET/DIH/UCE, Issy les Moulineaux, France

Calin Mosut (2000) : "WEBXREF version TypWeb, mise à plat des éléments html des pages WEB", rapport technique.

Aude Maisondieu*, Andréa Kuncova* (2000), "Constitution d'un corpus Web dans le cadre du Projet Typweb", rapport de stage TYPWEB.
* Université Paris 3, ILPGA

Typologie des textes

B. Habert *, G. Illouz **, H. Folch* , S. Fleury *, S. Heiden *, P. Lafon*, « Maîtriser les déluges de données hétérogènes », Actes de *TALN'99*, Atelier Corpus et traitement automatique des langues~: pour une réflexion méthodologique, p.37-46, Anne Condamines, Cécile Fabre, Marie-Paule Péry-Woodley éditeurs, Cargèse, 12-17 juillet 1999.
* UMR8503 CNRS/ENS Fontenay Saint-Cloud, France
** LIMSI CNRS Orsay, France

B. Habert *, G. Illouz **, H. Folch* , S. Fleury *, S. Heiden *, P. Lafon *, S. Prévost *, « Profilage de textes : cadre de travail et expérience », JADT'2000, Lausanne.
* UMR8503 CNRS/ENS Fontenay Saint-Cloud, France
** LIMSI CNRS Orsay, France

10 Annexes

10.1 Webxref

10.1.1 Présentation générale du programme

10.1.1.1 Sources *Web d'origine*

Pour créer l'outil de désossage, on a repris le logiciel *Webxref035* distribué gratuitement sur le site *http://www.perl/com*. C' est un programme *Perl* conçu pour vérifier rapidement un ensemble local de pages *HTML* et mettre à jour les dysfonctionnements possibles : structure défectueuse du site, liens pointant vers des URL, des fichiers ou des ancres manquant, etc. *Webxref* comporte des options permettant la recherche de motifs dans les fichiers référencés et l'édition de documents *HTML* à partir de structures récupérées dans ces fichiers. Partant de cet outil existant, on l'a modifié et enrichi afin d'en faire un outil de désossage de sites adapté aux objectifs visés.

10.1.1.2 Création d'un outil de désossage de sites

L'outil d'origine *Webxref* produit des tableaux contenant les références des documents trouvés, URLs, ancres, images et fichiers afférents. La fonction *DissectFile* incluse dans la nouvelle version *Webxref036* prend appui sur les algorithmes utilisés dans le script *dissectsite.hts* disponible à l' adresse *http://worldwidemart.com/scripts/*. Ce script est utilisable dans un navigateur. Il prend en argument une URL (en ligne) et génère des variables stockant l' information sur les en-têtes et les éléments envoyés par le serveur (« text objects » ou « tag objects »). Les résultats produits sont présentés au format *.html* dans une nouvelle fenêtre du navigateur.
Notre outil (*Webxref036*) reprend ce mode d' adressage des éléments dans chaque rapport de « dissection » avec cependant quelques modifications. Il comporte des options qui contrôlent le format de sortie (*.html* ou *.txt*), permettant en outre la récupération des en-têtes *MIME* et des liens contenus dans chaque document *HTML* avant son éclatement en « text objects » et « tag objects ».

10.1.1.3 Perspectives de développement

Une fonction permettant la visualisation des indications de structure des sites est en cours de réalisation. Pour mener à bien ce travail, nous avons choisi d' adapter le script *sitemapper.pl* distribué également par *http://www.perl.com*. Les objectifs que nous poursuivons sont la révision du code en vue de la greffe sur *Webxref* et la redéfinition des formats de sortie en accord avec les résultats produits actuellement par *Webxref*.

10.1.1.4 Bugs constatés

La version d'origine de *Webxref* distribuée par le site *www.perl.com* comporte, selon nos tests, quelques problèmes de flux interne qui se manifestent au niveau de la fonction *Check_External_URLs*. Nous avons proposé une solution qui semble pratique pour la vérification des références externes sans avoir pu expliquer les causes initiales de dysfonctionnement du programme (voir le code commenté de la nouvelle version de l'outil *036)*.

La version initiale ne collecte pas et ne vérifie pas les références des liens présents dans les balises *body*, *base*, *link*, *script*, *input*, *applet*, *embed*. Nous avons enrichi la liste des liens à vérifier et nous avons corrigé, avec la version *036*, une erreur de code qui ne permettait pas la vérification des liens du type *<a href='file.html#anchor'>*.

Après l'ajout de la routine *DissectFile*, les fonctions annexes reprises à la version *035* et partiellement modifiées n'ont pas été testées systématiquement. Notre corpus-test comporte seulement des arborescences aspirées via *HTTrack*. Nous ne savons pas encore si *Webxref036* dépiste correctement la présence des scripts *.cgi* et des liens du type *<a href='/image.gif'>* dans un site structuré différemment. Nous vous remercions de nous adresser, à ce propos, vos remarques et suggestions.

10.1.2 Description technique de l'outil *Webxref036*

10.1.2.1 Objectifs initiaux

Cet outil conçu pour des traitements en série répond à la nécessité de gérer un nombre important de sites dans le cadre de l' étude en cours. Les documents *Web* doivent être repérés dans l' arborescence de chaque site et soumis à des traitements sélectifs ; *Webxref036* peut parcourir les sites et appliquer à chaque document *HTM(L)* trouvé la routine de « désossage » *DissectFile*. Les algorithmes que nous proposons essaient d' améliorer l' allocation de la mémoire durant les procédures, le format des sorties, et l' utilisation du programme sur plusieurs plates-formes (sous *Windows*, *Linux* et *Unix*). En effet, la version d'origine *Webxref035* ne fonctionne pas correctement sous *Windows* avec le shell *Bash* (*Cygwin Beta 19*) et pas du tout si le shell utilisé est *DOS*. La version adaptée *036* résout ce problème : elle récupère du code Perl (bibliothèques standard *HTML::HeadParser.pm*, *HTTP::Headers.pm*, *HTML::Parser.pm*,

HTML::Elements.pm, *HTML::TreeBuilder.pm*, *HTML::Entities.pm*), *HTML* (*dissectsite.hts*, disponible à l'adresse : *http://worldwidemart.com/scripts/*) et, de ce fait, accepte les plates-formes de travail *Windows*, *Unix* et *Linux*.

10.1.2.2 Fonctionnalités du programme

Par rapport à la version *035*, les adaptations suivantes ont été opérées :

* ajout de la nouvelle fonction *DissectFile* remplaçant la fonction *GetLinks* (version *035*) ;
* modification du code pour permettre des traitements en série ;
* ajout de structures de protection supplémentaires ;
* gestion de la durée des traitements et de la mémoire ;
* gestion des chemins complets pour les fichiers d'entrée (« input files » ;
* création de répertoires distincts pour les résultats produits en fin de traitement ;
* modification en vue d'un usage sur des plates-formes *Windows* ;
* modification de l'option *-long* : dans la version d'origine, l'option *-long* implique *-nohttp* c'est-à-dire que les références des URLs externes sont effacées durant l'impression des listes et par conséquent ne peuvent être vérifiées par la suite.

10.1.2.3 La routine **DissectFile**

Si l'option *-rep* est activée, *Webxref036* appliquera cette fonction aux documents *HTM(L)* qui sont référencés dans un site stocké sur les disques locaux. Avec l'option *-spell*, le flux de caractères qui entre dans le parser sera filtré en vue de la correction des erreurs de syntaxe *HTML* : balises défectueuses, couples attribut-valeur tronqués, etc.
La fonction *DissectFile* récupère, avant l'éclatement de chaque page, les en-têtes du type *MIME* du document pour les imprimer dans le fichier de rapport selon la grammaire :

nombre courant
champ (type)
valeur

Avant le passage du parser, les scripts sont extraits des documents *HTM(L)* afin d'être directement imprimés dans les rapports au format *.txt*. En effet, les scripts ne sauraient être « désossés » et leur affichage en format *.html* pourrait provoquer des accidents d'interaction avec le navigateur. Les liens sont recueillis et stockés dans une liste afin d'être typés ultérieurement.
Par la suite, l'élément *<html>* de chaque document et ses descendants subissent un traitement sélectif, selon qu'ils constituent des « déclarations » (habituellement la déclaration du type de document), des « tag objects » (balises ouvrantes ou fermantes), des « text objects » (séquences textuelles) ou des « comments » (des balises de commentaire).
La sous-fonction *elements* appelée à ce point fait appel à une version adaptée du parser distribué avec *Perl5* pour *Windows* (*HTML::Parser.pm*) pour typer ces éléments et les imprimer après encodage-décodage des entités HTML. Les attributs *SGML* des « tag objects » et les séquences textuelles seront affichés dans le format demandé (*.html* ou *.txt*) selon la grammaire du script *dissectsite.hts* (cf. références *supra*).
Les attributs des liens sont soumis à un test pour le typage. La typologie est reprise en partie à l'auteur de *Webxref035.pl*.

Algorithmes:

*** mailto: <mailto:>**
*** cgi-bin: <cgi-bin>**
*** gopher: <gopher:>**
*** news: <news:>**
*** lien ftp : <ftp:>**
*** image : <img src>**
*** lien http : <http:>**
*** lien hypertextuel: <link rel><link rev>**
*** lien interne vers fichier : <file:>**
*** lien vers ancre : <#>**

Les liens typés sont imprimés dans le fichier de rapport selon la grammaire :

numéro courant
classe
type
attributs

10.1.2.4 Contrôle des formats de sortie

Si, par défaut, la fonction *DissectFile* est chargée au lancement du programme, celle-ci peut être inhibée en utilisant l'option *-norep.*

commande : **./Webxref036.pl -norep site/**

L'index des résultats, plus complexe que celui de la version *035* du programme, présente les résultats produits dans des tableaux (format *.html*) pour plus de clarté. Les options *-nohttp -long* et *-txt* étant activées par défaut, pour obtenir des rapports en format *.html*, qui demandent plus de mémoire pour le stockage, if faut exécuter l'option *-html* :

commande : **./Webxref036.pl -html site/**

Les listes imprimées sont constituées de :

*** fichiers.html**
*** répertoires**
*** images**
*** ancres**
*** mailto's**
*** news**
*** ftp**
*** telnet**
*** gopher**
*** URLs externes**
*** scripts cgi-bin**
* fichiers non trouvés
* fichiers non reconnus par Word
* répertoires non trouvés
* ancres non trouvées

Les listes suivantes sont disponibles (si les options correspondantes sont activées) :

*** fichiers et répertoires non référencés**
*** adresses externes vérifiées « ok » et adresses externes manquantes :** option *-http*
*** fichiers dans lesquels un motif de recherche a été trouvé** : option *-find*
*** fichiers dans lesquels un motif a été remplacé** : option *-replace*
*** fichiers anciens ou récents par rapport à une certaine date :** options *-after*, *-before*

10.1.2.5 Les traitements en série

Si on exécute *Webxref036* avec un répertoire site en argument, la fonction *DissectFile* crée un rapport pour chaque document *HTML* dont le nom est référencé à l'intérieur du site. On peut décider d'aligner plusieurs arguments sur la ligne de commande afin d'abréger les temps de traitement. Le programme cherchera tour à tour les sites (ou les fichiers), descendra les arborescences, trouvera le répertoire *root*, la page d'accueil, extraira les liens et mettra à plat l'information sur chaque page référencée. Les résultats seront stockés dans des répertoires distincts. Durant le traitement d'un site, *Webxref* imprime les éventuels messages d'erreur dans le fichier *trace.txt*.
Notons que la version *036* a été prévue pour gérer les chemins complets, lire et écrire sur plusieurs disques à la fois. En effet, le programme peut prendre en entrée des répertoires distincts, portant le nom du site et contenant un set de documents *HTML* inter-référencés, et crée des répertoires distincts contenant les rapports sur chaque site et sur chaque page.

Exemple :

répertoire de travail : **c:**
entrée : **c:/site1/ a:/site2/ g:/site3/**
commande : **./Webxref036.pl site1 a:/site2/ g:/site3/**
sortie : **c:/res1(site1) c:/res2(site2) c:/res3(site3)**

Monitoring du temps de traitement :

En fonction de la vitesse du processeur, des options activées et du nombre d'arguments, les temps de traitement peuvent être longs. Pour suivre visuellement l'avancement des traitements, le programme envoie des messages sur le *STDOUT*, énonçant les étapes de traitement. *Webxref036* imprime un « + » pour chaque référence trouvée et un « - » pour chaque référence absente. Cette information, reprise à la version antérieure, est incluse dans les fichiers de résultats.

Exemple :

répertoire de travail : **c:/site/**
entrée : **c:/site/test.html**
commande : **./Webxref.pl c:/site/test.html**
sortie STDOUT :

Getting parameters...

Checking c:/site/test.html...

Dissecting files...
file 1
Printing lists...
All done.
See C:/site/res1(site)/analysis_results

Structures de protection :

Après chaque traitement d'un site, *Webxref036* fait une mise à jour de la table d'arguments. La fonction *UpdateARGV* vérifie si les fichiers ou les répertoires spécifiés dans la ligne de commande ont été déjà référencés pour éviter les itérations supplémentaires. Si tel est le cas ou si la table d'arguments contient des doubles (accident possible lors des traitements en série) ceux-ci sont réduits. Les arguments commençant par « - » venant après un nom de fichier/répertoire sont ignorés.

Exemple :

répertoire de travail : **c:/site/**
entrée : **c:/site/file1.html c:/site/file2.html ; lien file1->file2**
commande : **./Webxref036.pl c:/site/file1.html -html file1.html file2.html**
sortie : **c:/site/res1(site)**

Remarque : le code a fait une seule itération, créant en sortie un seul répertoire de résultats. L'option *-html* n'ayant pas été passée au programme en premier argument, le format de sortie est, par défaut, *.txt*. Le second argument a été réduit étant similaire au premier ; le troisième, étant déjà référencé, a été réduit également. *res1(site)* contient *analysis_results.txt*, *rep1(file1).txt* et *rep2(file2).txt*.

10.1.2.6 Gestion des résultats

La fonction *WriteResFile* ouvre un répertoire de résultats distinct pour chaque traitement, sans risque d'effacement durant les traitements en série ou après plusieurs exécutions du programme dans le même répertoire. La valeur d'incrément d'un compteur (qui dénombre chaque appel de la fonction) est incluse dans le nom du répertoire des résultats. Pour les attributions ultérieures, ce répertoire porte le nom du site traité (le nom du répertoire dans lequel se trouve le fichier passé en argument ou bien le nom du répertoire passé en argument). Si l'on ne demande pas un endroit particulier pour le stockage des résultats, ils seront mis dans le répertoire depuis lequel *Webxref036* a été lancé.

* Exemple :

répertoire de travail : **c:/**
entrée : **c:/site1/file1.html a:/site2/file2.html**
commande : **./Webxref036.pl site1 a:/site2**
sortie : **c:/res1(site1) c:/res2(site2)**

Si l'on souhaite que les rapports soient stockés dans un répertoire précis, il y a deux possibilités : soit on indique un chemin complet vers ce répertoire, soit on demande un répertoire descendant, sans arborescence. Pour cela, il faut utiliser l'option *-at* suivie du chemin demandé.

* Exemple :

répertoire de travail : **c:/**
entrée : **c:/site/**
commande : **./Webxref036.pl -at c:/documents/site/**

sortie : c:/Documents/res1(site)
commande : ./Webxref036.pl -at Documents/site/
sortie : c:/Documents/res1(site)

Structures de protection :

Le chemin spécifié après *-at* doit mener à un répertoire valide. Aucun répertoire (excepté les « res ») n'est créé à la demande, le programme sortant de son itération avec un message d'erreur sur le *STDOUT*. Le traitement est bloqué si le répertoire des résultats ne peut être ouvert.

10.1.2.7 Le travail sur les plates-formes Windows

L'outil original *Webxref035* a été conçu pour le système d'exploitation *UNIX*. Le code du programme contient des appels du *Shell*, ce qui fait qu'il ne peut pas reconstituer l'arborescence des fichiers sous le shell *DOS* ou bien remonter au point d'insertion sous *Windows* même si le shell utilisé est *Bash*. Aussi, lorsque le format de sortie est *.html*, la navigation entre les fichiers devient-elle impossible. Avec la nouvelle version *Webxref036*, on peut naviguer entre résultats et fichiers de départ depuis n'importe quel répertoire, même si l'option *-fullpath* n'est pas activée. *Webxref036* collabore bien avec le shell *DOS*.

Structure de protection :

Si le slash inverse apparaît après l'options *-at*, ou au niveau des noms de fichiers avec arborescence passés en arguments, celui-ci est « traduit » pour l'uniformisation des paths, selon le modèle *UNIX*.

Les fonctions du programme

Toutes les fonctions de *Webxref035*, à l'exception de *-root* et *-files* (« assumées » par la nouvelle structure du programme) ont été conservées pour la souplesse des traitements. Quelques-unes ont subi de légères adaptations. Ainsi avons-nous récupéré sous le nom *-http* l'option qui commande la recherche des références externes. Dans la version d'origine, les fonctions *-http* et *-long* sont incompatibles, car une fois les listes imprimées, les variables stockant les URLs externes sont déchargées. L'utilisation par défaut de certaines options a été changée également, mais les correspondances avec la version initiale sont évidentes.
Voici le comportement par défaut du programme : *Webxref036* dresse la liste de tous les fichiers trouvés dans les répertoires, rapporte les problèmes, dissèque les fichiers *.htm(l)* et imprime les résultats format *.txt* dans le répertoire *res(site)*.

Liste des options :

./Webxref -html fichier.htm(l)
Donne des rapports en format *.html*

./Webxref -norep file.html (voir aussi les possibilités de configuration)
Dresse la liste des fichiers trouvés et rapporte les problèmes

./Webxref -at path fichier.htm(l)
Offre la possibilité de choix d'un répertoire pour les résultats

./Webxref -http fichier.html
Cherche le fichier.html et les URLs externes

./Webxref -brief fichier.html
Génère un rapport bref (liste les problèmes)

./Webxref -fullpath fichier.html
Imprime les noms des fichiers avec les arborescences).

10.1.2.8 Analyse des algorithmes

La structure du répertoire *main* de la nouvelle version de *Webxref* a subi plusieurs modifications par rapport à la version d'origine *035*. Les adaptations du code d'origine se trouvent dans les fonctions suivantes :

10.1.2.9 Fonctions reprises à la version initiale

GetParams

Régit les options et lance les traitements en série .Une boucle parcourt la table des arguments et la variable *$InFile* est assignée à chaque argument.

GetCWD

Dans le code original de *Webxref*, le répertoire courant était obtenu par un appel au *Shell* ; ceci fonctionne de manière imparfaite avec le *Bash*, lorsque le système d'exploitation est *Windows*, et bloque le programme sous le shell *DOS*. Nous avons proposé la solution d'un appel à la bibliothèque Perl standard *Cwd.pm*.

Check_External_URLs

La liste des URLs externes est récupérée par la modification de *PrintLists*. C'est au niveau de cette sous-fonction qu'elle est déclarée vide pour préparer une nouvelle itération du code.

SplitURL

Une arborescence *Windows* sera reconnue à ce niveau avec la version *036*. C'est ici également que les fichiers locaux sont pourvus de l'en-tête « file:/ » pour que la navigation soit possible à partir de n'importe quel répertoire

PrintFile

Ajoute l'en-tête « file:/ » aux fichiers

AddedToList

Les éléments de la liste sont pourvus de l'en-tête « file:/ » pour constituer une adresse valide pour la navigation sous *Windows*

PrintList

La fonction a été modifiée pour des raisons d'uniformisation des critères de présentation des résultats

PrintLists

A ce niveau, la table des URLs externes est récupérée. Un traitement différentiel est appliqué aux listes : nous en avons modifié la présentation par l'ajout des tables *HTML*.

10.1.2.10 Nouvelles fonctions

Les nouvelles procédures de la version *036* ont comme support les fonctions et les sous-fonctions suivantes :

shell_path

Transforme les paths *DOS* en paths *UNIX*.

WriteResFile

Génère les fichiers de résultats en accord avec l'option *-at*, faisant appel à la sous-fonction *get_path_to results*.

get_path_to_results

Cette fonction choisit le répertoire d'ouverture des résultats en accord avec l'option *-at*. L'argument venant après est considéré comme « path ». Par défaut les résultats sont mis dans le répertoire depuis lequel *Webxref* a été lancé. Des boucles de protection vérifient la présence du slash final pour les répertoires. Un compteur dénombre les itérations du programme ; l'incrément donnera, avec le nom du répertoire passé en argument (ou, si l'argument est un fichier, avec le répertoire contenant le fichier), le nom du répertoire des résultats - un par site.

DissectFile

Cette routine collecte l'information sur les liens présents dans chaque document HTM(L) et appelle la fonction *elements* (qui dispense l'information sur chaque élément *HTML*) si l'option *-rep* est activée. Le code continue son itération jusqu'à ce que la fin de la table d'arguments soit atteinte, même si un problème survient (les sites peuvent ne pas être correctement aspirés). *DissectFile* fait appel à la routine *AddedToList* pour créer la liste des fichiers disséqués.

spell

Si l'option *-spell* est activée, cette sous-fonction filtre le flux de caractères envoyé vers le parser, proposant des corrections pour les erreurs de syntaxe *HTML*. Il y a trois types d'action possibles : modification de la syntaxe sans message d'avertissement, lorsque l'erreur dépistée peut être réduite sans risque (exemple : complétion d'une balise fermante défectueuse), modification de la syntaxe avec message d'avertissement envoyé vers le fichier *trace.txt* (comme la fermeture d'une balise défectueuse suivie d'une séquence textuelle après le dernier couple attribut-valeur reconnu) et message d'avertissement sur la présence d'une balise trouvée non conforme et qui ne peut être modifiée faute de repères solides. Le message « Not conforming tag or simple mask found at 'path/file'. May have caused parse errors ! » n'est pas nécessairement un message d'erreur. Si la balise défectueuse trouvée n'est pas une balise *HTML* mais un simple masque contenant du texte ou une balises marquée comme valeur dans un couple attribut-valeur, elle sera correctement interprétée par le parser. Vu la présence de ce genre de balises dans notre corpus, nous avons préféré ne pas modifier les séquences du type <*tag text... <another_tag*>. Généralement, le parser considérera les balises défectueuses qui n'auront pas été corrigées comme des séquences textuelles (« text objects »). Le traitement pourra éventuellement être refait après correction manuelle des erreurs signalées.

entities

Il s'agit d'une version adaptée de la fonction *decode_entities* (routine définie dans le corps de la bibliothèque Perl *HTML::Entities.pm*). Traduit les entités *HTML* pour le format de sortie *.txt*.

elements

Cette sous-fonction est une version adaptée de la bibliothèque standard Perl *HTML::Parser.pm*. Nous avons développé les fonctions prédéfinies : *parse*, *start*, *end*, *declaration*, *comment* et nous avons créé trois nouvelles routines, *headers*, *scripts* et *links*, qui dispensent l'information sur les en-têtes *MIME* et les liens présents dans les documents *HTML*.

headers

Lors de la dispense des éléments contenus dans la section <*head*>... <*/head*> les entités sont décodées seulement pour le format *.txt*.

scripts

Pour la protection de l'affichage, la routine n'imprimera pas les scripts dans les fichiers de résultats format *.html*. Les scripts peuvent être récupérés avec l'option *-txt* (activée par défaut).

parse

Parse chaque fichier *.html* comme une chaîne de caractères et en catégorise les éléments. Les fonctions internes du parser, *start*, *end*, *comment*, *declaration*, appliquent un traitement différentiel à ces catégories en vue de l'impression dans le fichier de rapport.

links

Applique la routine de typage selon que les attributs des liens sont reconnus. La typologie a été reprise en partie à l'auteur de *Webxref*. Si le format de sortie est *.html*, les liens sont référencés.

display

Imprime périodiquement la variable *$tmp* stockant les informations des collectes.

UpdateARGV

Cette fonction a été conçue pour protéger le programme durant les traitements en série. *UpdateARGV* est appelée à chaque itération. Deux arguments identiques comme nom et emplacement peuvent être donnés au programme et l'on désire éviter dans ce cas un traitement supplémentaire et l'encombrement du disque avec les répertoires de résultats. Si plusieurs fichiers co-référencés sont passés en arguments, on souhaiterait ne pas ouvrir des répertoires distincts pour les résultats, mais lancer un seul traitement. Dans le premier cas, la fonction vérifie la présence du nom du fichier dans les tables (aussi le traitement se répète-t-il à chaque itération de la boucle principale), le réduisant si la recherche est positive. Pour le deuxième cas, le test consiste en la comparaison entre l' « input file » et chaque élément de la table d'arguments, dont les arborescences sont éventuellement reconstituées.

10.1.2.11 Contenu du programme

10.1.2.11.1 Répertoire main de Webxref036

```
&GetParams;
# vérifie les options passées en arguments
die "No input file(s).\n(try Webxref -help)\n" unless @ARGV;
$CWD=&GetCWD;
# répertoire courant de travail
foreach (@ARGV) {
  # traitements en série
  next if ($_ eq""||$_=~/^-/);
  # structures de protection
  $_=&shell_path($_);
  # path UNIX
  $InFile=$_;
  if (! -e $InFile) {
    &AddedToList(*LostFileList,$InFile,$WebxrefReferer);
    print "Cannot find file $InFile\n";
    next;
  }
  $SiteRoot=$CWD;
  # par défaut, le répertoire depuis lequel Webxref est lancé est considéré comme répertoire root du
site
  if ($_ =~ m#^((?:\w:)?/(?:.+/)*)[^/]+/?$#) {
    # chemin complet
    $SiteRoot=$1;
    s/$1//;
  }
  print <<EOM && next unless ( -e $SiteRoot);
  The site directory $SiteRoot does not exist!
EOM
  print <<EOM && next unless ( -d $SiteRoot);
  "\"$SiteRoot\" is not a directory!
EOM
  print <<EOM && next unless ( -r $SiteRoot);
  Cannot access directory \"$SiteRoot\"!
EOM
  &WriteResFiles;
  # ouvre les fichiers de résultats
  # Si le programme est interrompu à ce moment
  # NOTE (n.a.) : This is unreliable if Webxref was interrupted
  # asynchronously. The C-library is not re-entrant, so
  # if printing was in progress printing may well fail
  # due to malloc running into trouble. Oh well. It does
  # work sometimes.
  $SIG{INT} = 'InterruptHandler' if (! $NoInterrupt);
  $WebxrefReferer = '--Webxref--';
  # Pendant l'impression, les paths seront optionnellement abrégés
  $SiteRootExpr = $SiteRoot;
  if ($SiteRootExpr !~ m#/$#) { $SiteRootExpr .= '/'; }
  $SiteRootExpr =~ s/(\W)/\\$1/g;  #  déspécifie les caractères
  ($d,$f,$a,$RootDepth) = &SplitFile($SiteRoot);
  # print "\nSiteRoot=$SiteRoot, \nd=$d \nf=$f \na=$a \ndep=$RootDepth\n";
  $MaxDepth = $MaxDepth + $RootDepth;
  # print "Maxdepth=$MaxDepth\n";
  print RES "<pre>\n" if ($HTMLReport);
  &GetFluffFiles($SiteRoot) if ($Fluff);
  # fichiers non-référencés présents dans les sites
  $InFile = $SiteRoot . $_;
  print "\nChecking $InFile...\n\n";
  print "Dissecting files...\n" if ($ReportFiles);
  &GetReferences($InFile,"--Webxref--");
  # cherche les références
  &UpdateARGV;
  # la table d'arguments est mise à jour
  &PickFluff if ($Fluff);
  print RES "</pre>\n" if ($HTMLReport);
  &PrintLists;
  # rapport sur les références vérifiées
  $DotCount = 0;
  if ($Do_External_URLs) {
  print "External HTTP checking starts...\n" if ($Dots);
  print RES "\n\n";
  print RES "<p>" if ($HTMLReport);

  # Check external URLs
  if (!$Silent) {
  print <<"E_O_T";

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Going to really check external URLs via the network.
This may take some time. Simply abort Webxref if you
are out of patience.
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
E_O_T
```

```
  }
  &InitStatusMessages;
  &Check_External_URLs;
  # cherche les références externes
  }
  &PrintHTTPLists;
  # rapport sur les références externes
  print RES "</body></html>\n" if ($HTMLReport);
  close RES;
  close TRA;
  print "\n" if ($HTTP);
  print "All done.\nSee $path_to_results","analysis_results\n";
}
```

10.1.2.11.2 Sous-fonction get_path_to_results

```
$loop++;
  # compte les itérations
  $path_to_results=$CWD;
  # par défaut les résultats seront mis dans le répertoire courant de travail
  my $site;
  if ($AskedPath) {
    # un endroit sur le disque est spécifié
    s/\\/\//g;# chemin Shell
    $AskedPath .= '/' unless ($AskedPath =~ m#/$#);
    # s'assure de la présence du slash final
    if ($AskedPath =~ m#^(?:\w:)?/#) {
       # un chemin complet est demandé
      $path_to_results=$AskedPath;
    }
    else {
       # un répertoire en descendance a été demandé
      $path_to_results .= "$AskedPath";
       # ajoute le chemin au répertoire de travail
    }
  }
  die (<<EOM) unless (-d $path_to_results);
"$path_to_results" is not a valid path. $!\n
EOM
   # structure de protection
  if ( -d $InFile) {
    if ($InFile =~ m#([^/]+)/?$#) {
       $site=$1;
       # le nom du répertoire est récupéré
    }
  }
  else {
    if ($SiteRoot =~ m#([^/]+)/$#) {
       $site=$1;
       # le nom du répertoire passé en argument est récupéré
    }
  }
  $site=~s/://g;
  # protection contre les caractères spéciaux

  mkdir "$path_to_results/res$loop($site)",'ugo+rwx' ||die (<<EOM);
  Cannot create results directory at $path_to_results!
EOM
  $path_to_results .= "res$loop($site)/";
}
```

10.1.2.11.3 Fonction *DissectFile*

```
# lit le fichier .html et extrait tous les éléments; retourne @links, @Newlist
  local($filename) = @_;
  local(@Tags,   @headers,@scripts,   %links,   @links,   %Newlist,   @Newlist,   %LocalAnchorsFound,
%LocalAnchorsWanted);
  local($elements,$tag,$Link);
  unless (open(HTML, $filename)) {
    print TRA "sub DissectFile warn:\n   Could not open file '$filename'\n\n";
    return;
  }
  # print "Opening $filename\n";
  @Tags = <HTML>;
  # le fichier est lu comme une chaîne de caractères
  close(HTML);
  $elements = join('',@Tags);
  &DoFind($filename,$elements) if (defined($FindExpr));
  # trouve les fichiers qui contiennent l'expression recherchée avec l'option -find
  if ($ReportFiles) {
    if ($elements =~ m#<(TITLE)>(.+)</TITLE>#si) {
      # récupère le titre
      push @headers,"$1=$2";
    }
    while ($elements =~ m#<SCRIPT.+?>(.+?)</SCRIPT>#sig) {
      push @scripts, "$1";
    # récupère les scripts
    }
    $elements =~ s#(<SCRIPT.+?>).+?(</SCRIPT>)#$1 $2#sig;
    $elements = &spell($elements);
  }
```

```
  # enlève les scripts avant le passage du parser
  @Tags = split/</, $elements;
  for(@Tags) {
    s/\n/ /g;
    s/>.*//;
    # print "tag: $_\n";
    if ($ReportFiles) {
      if (m#^(?:META|ISINDEX|STILE|BASE|LINK)#i) {
        # elements de la section <head>...</head>
        push @headers,"$_";
        # extrait les en-têtes
      }
      if                          (m#^(A|IMG|FRAME|AREA|FORM|BODY|BASE|LINK|SCRIPT|INPUT|APPLET|EMBED
).+(HREF|SRC|ACTION|BACKGROUND|CODEBASE|URL)\s*=\s*"?([^"\s]*)"?#i){
        # éléments contenant des attributs liens
        $links{"$1 $2 = $3"} = 1;
        # extrait les liens
      }
      @links = keys (%links);
    }
    # <a href/name
    if (/^A\s+/i) {
      #print "-anchor: $_\n";
      # -- a href
      if (m#HREF\s*=\s*"?([^"\s]*)"?#i) {
        $Link = $1;
        #print "  href: -$Link-\n";
        # Lien vers une ancre dans le document même? (<a href=#anchor>)
        if ($Link =~ m/^#/) {
          #print "  -$filename$Link- wanted\n";
          # Situation spéciale : ne donne pas d'erreur avec "href=file.html#"
          #print "  -$Link- wanted\n";
          if ($Intermediair) {
            next if ($Link eq "#");
          }
          $LocalAnchorsWanted{"$filename$Link"} = 1;
        }
        # Lien vers un autre fichier?  a href=file.html#anchor
        elsif ($Link =~ m/#/) {
          $Link =~ m/(.+)#(.+)/;
          # print "LINK: $Link  $1 $file - equal?\n";
          if ($1 eq $file) {  # Current file after all
            $LocalAnchorsWanted{("$filename" . '#' . "$2")} = 1;
          }
          else {
            $Link =~ s/#.+$//;
            $Newlist{$Link} = 1;
          }
        }
        else {  # Just a file ref
          $Newlist{$Link} = 1;
        }
      }
      elsif (m#NAME\s*=\s*"?([^"\s]*)"?#i) {
        # -- a name=...
        $Link = $1;
        #print "  name: $Link\n";
        #print "  -$filename$Link- found\n";
        $LocalAnchorsFound{"$filename#$Link"} = 1;
      }
    }
    # <img src=...
    elsif (/^IMG/i) {
      if (m#SRC\s*=\s*"?([^"\s]*)"?#i) {
        $Link = $1;
        #print "  img: $Link\n";
        # Ajoute le fichier à la liste
        $Newlist{$Link} = 1;
      }
      else { print TRA "sub DissectFile warn:\n   Image parse error in '$filename':\n  $_\n\n"; }
    # le seul message d'erreur retenu de la fonction GetLinks antérieure
    }
    elsif (/^(?:FRAME|AREA|FORM|BODY|BASE|LINK|SCRIPT|INPUT|APPLET|EMBED)\s+/i) {
      if (m#(?:SRC|HREF|ACTION|BACKGROUND|CODEBASE)\s*=\s*"?([^"\s]*)"?#i) {
        # quelques attributs ont été rajoutés par rapport à la version 035
        $Link = $1;
        # Ajoute le nom du fichier à la liste
        $Newlist{$Link} = 1;
      }
    }
  }
  &elements;
  # récupère tous les éléments du document HTML
  #print "LocalAnchorsFound: \n",join("\n",keys(%LocalAnchorsFound)),"\n";
  #print "\nLocalAnchorsWanted: \n",join("\n",keys(%LocalAnchorsWanted)),"\n";
  # Ajoute les " local anchors " trouvées à la liste globale
  foreach $Anchor (keys(%LocalAnchorsFound)) {
    &AddedToList(*AnchorList,$Anchor,$filename);
  }
  # Voit si les ancres référencés localement y sont et les efface de la liste %Newlist si cela en
est ainsi
```

```
  foreach $Anchor (keys(%LocalAnchorsWanted)) {
    if (!defined($LocalAnchorsFound{$Anchor})) {
      &AddedToList(*LostAnchorList,$Anchor,$filename);
      print "\n" if ($Errors && $Dots);
      &PrintDot('-') if ($Dots);
      print "Anchor",&PrintFile($Anchor),"is NOT present in file",
            &PrintFile($filename),"\n" if (!$Silent) || ($Errors);
    }
    else {
      $Newlist{$Anchor} = 0; # effacé du hachage
    }
  }
  foreach (keys(%Newlist)) {
    if ($Newlist{$_}) { push(@Newlist, $_);}
  }
  # print "\nnewlist: \n",join("\n",@Newlist),"\n";
  return @Newlist;
```

10.1.2.11.4 Sous-fonction *spell*

```
if ($Spell) {
    # corrige les erreurs de syntaxe HTML
    local ($input) = @_;
    $input =~ s#\240#\040#g;
    $input =~ s#(<[!/a-zA-Z][a-zA-Z0-9\.\-]*)#\240$1#g;
    $input =~ s#\s([a-zA-Z][a-zA-Z0-9\.\-_]*\s*=["']?)# >_$1#g;
    # marque toutes les balises et les couples attribut-valeur
    $input =~ s#(_[^\s]+?\s*=\s*"[^"<]*?)>#$1">#sg;
    $input =~ s#(_[^\s]+?\s*=\s*'[^'<]*?)>#$1'>#sg;
    # corrige les valeurs tronquées à droite
    $input =~ s#(_[a-zA-Z][a-zA-Z0-9\.\-_]*\s*=)\s*([^\s"'][^"'<>]*?)(["'])#$1$3$2$3#sg;
    # corrige les valeurs tronquées à gauche
    $input =~ s#>_##g;
    # efface les marques additionnelles
    $input =~ s#(<![^-][^<>]*?(['"])[^<>]*?\2)\s*\240#$1>#sg;
    # ajoute un ">" aux balises de déclaration défectueuses
    $input              =~              s#(<[a-zA-Z][a-zA-Z0-9\.\-]*(?:\s*[a-zA-Z][a-zA-Z0-9\.\-
_]*\s*=\s*(?:(["'])[^<>]*?\2|[^\s<>]*))*)\s*\240#$1>#sg;
    # ajoute un ">" après des couples possibles attribut-valeur
    $input =~ s#(</[a-zA-Z][a-zA-Z0-9\.\-]*)[\s\240]+>?#$1>#sg;
    # corrige le ">" manquant aux balises fermantes
    $input =~ s#(<!--[^>]*?--)\s*\240#$1>#sg;
    # ajoute un ">" à la fin des commentaires
    while ($input =~ m#<[^_][^>]+?\240#sg) {
      # les modifications doivent être rapportées
      print TRA "sub spell warn:\n";
      if  ($input  =~  s#(<![^-][^<>]*?(['"])[^<>]*?\2)([^<>]+)\240#$1>$3#s){print  TRA  "      Not
conforming declaration tag: '$1$3' found at '$filename'\n\n     Interpreted as '$1> $3'\n\n";}
      elsif       ($input       =~       s#(<[a-zA-Z][a-zA-Z0-9\.\-]*[^<>]*[a-zA-Z][a-zA-Z0-9\.\-
_]\s*=\s*(["'])[[^<>]*?\2)([^<=>]+)\240#$1>$3#s){print  TRA  "     Not  conforming  start  tag:  '$1$3'
found at '$filename'\n\n    Interpreted as '$1> $3'\n\n";}
      elsif ($input =~ s#(<!--[^>]*?--)([^>]*?)\240#$1>$2#s){print TRA "    Not conforming comment
tag: '$1$2' found at '$filename'\n\n    Interpreted as '$1> $2'\n\n";}
      elsif ($input =~ s#(<[^_][^>]+?)\240#$1#s) {print TRA "    Not conforming tag or simple mask:
'$1' found at '$filename'\n\n     May have caused parse errors\n\n";}
    # limite du correcteur
    }
    $input =~ s#\240##g;
    # efface les marques additionnelles
    # print "$input\n";
    return $input;
  }
```

10.1.2.11.5 Fonction *UpdateARGV*

```
for (@ARGV) {
    s/\\/\//g;# chemin UNIX
    $argv=$CWD.$_;
    # par défaut, on considère que le nom du fichier/site ne rappelle pas l'arborescence
    if ($_ =~ m#^(?:\w:)?/(?:.+/)*[^/]+/?$#) {
    # chemin complet en argument
      $argv=$_;
    }
    undef $_ if ($argv eq $InFile)||
    ($^O=~/MSWin/)&&(lc$argv eq lc$InFile);
    # "sort unique": on vérifie si les chemins reconstitués se superposent pour éliminer
    # les doubles
    $argv="file:/$argv";
    for $key (keys%FileList){
      undef($_) && last if ($key eq $argv)||
        (lc$key eq lc$argv)&&($^O=~/MSWin/);
    }
    # si des fichiers déjà référencés sont trouvés dans la table d'arguments
    # le shell DOS ne distingue pas les majuscules des minuscules
    chdir ($CWD);
    # retour au répertoire de travail initial
  }
}
```