Analyses Textuelles sur le Web

Une introduction
Présentation du projet TyPWeb

ILPGA, 2 mai 2002

Marie Pasquier, FT R&D
Serge Fleury, SYLED/CLA²T

En guise d’introduction…

Le web : un réservoir « infini » de ressources

Existence de pages très différenciées

Qu’est-ce qu’un texte sur le web ?

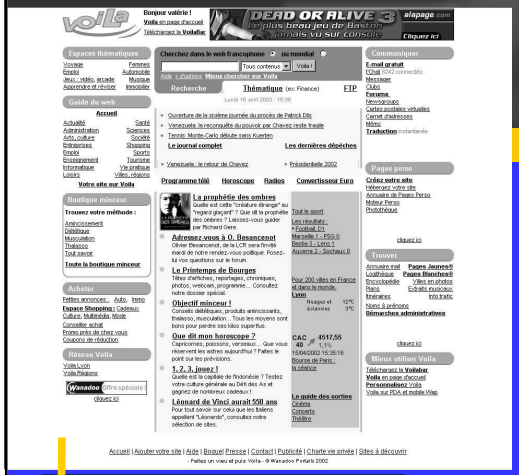
Le Web : réservoir « infini » de ressources

- En 1997 : la taille de l ’Internet = 2 terabytes
  - cf. Michel Lesk « How much information is there in the world » `http://www.lesk.com/mlesk/ksg97/ksg.html`
  - En comparaison : taille totale de l ’information 12 000 petabytes (6 millions de fois plus)
- Données disponibles
  - Du texte, des images, du son…
  - Des données hyper-connectées et articulées
  - Des données résultant d’usages de différentes natures
- Lieu d’articulation entre « outils » de communication et d’information

Des pages très différenciées

Exemples en image

- *Page d'accueil de portail*
- *Page d'accueil de site « perso marchand »*
- *Page d'accueil de site perso*
- *Page de contenu de site perso*
- *Page de réponses d'un moteur de recherche*
- *Page d'un formulaire d'inscription*
- *Page de chat*
- *Page de liens de site perso*



Page d'accueil portail

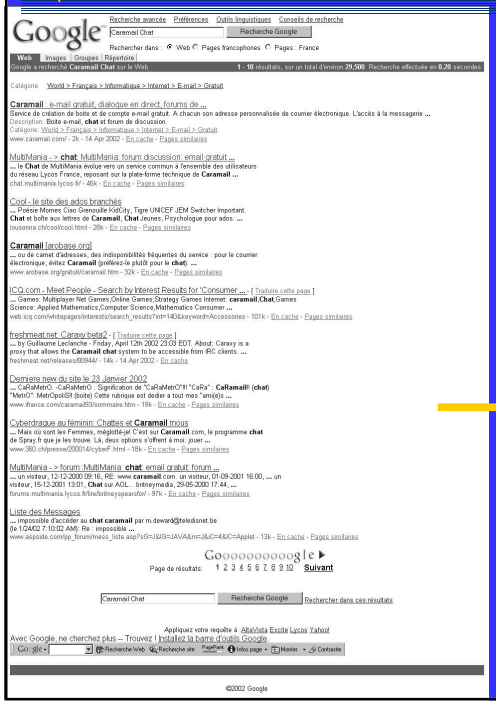
Page d'accueil site perso marchand

Page d'accueil site perso

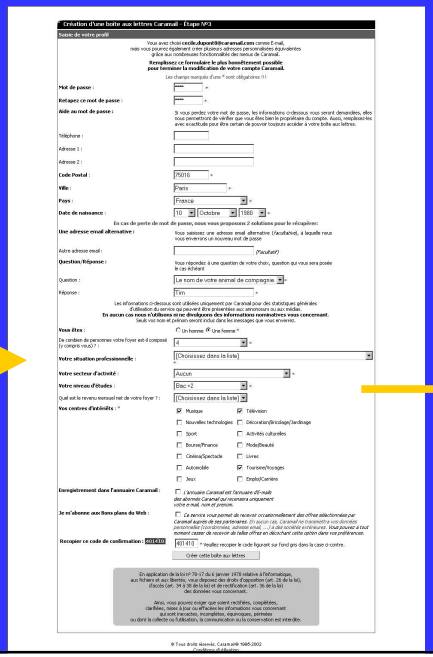
Page de contenu site perso

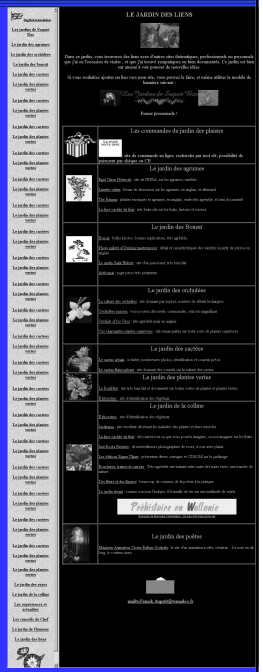
Requête : Chat+caramail

Page de réponses d'un moteur de recherche



Page formulaire d'inscription

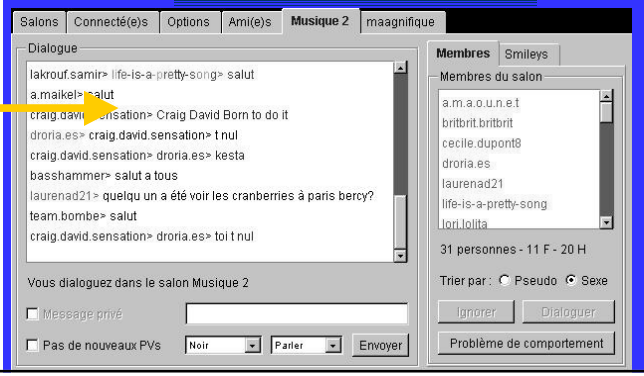




Page de liens site perso

De la diversité des documents du Web

Page de chat



Particularités des données disponibles sur le Web

- L'analyse de textes collectés sur Internet doit en prendre en compte la richesse linguistique des marquages structurels associés aux zones textuelles
  - les données présentes sur le web sont de différentes natures et organisées par des éléments « structurants »
  - les données textuelles sont réparties en « zones de langue »
- Modification des conditions générales de la textualité : le texte comme flux de données
  - les processus de productions textuelles s'entrecroisent via des relations-connexions, des données, des personnes hétérogènes
  - l'hypertexte permet de mettre en œuvre des mécanismes textuels fondés sur des enchaînements, des parcours de lecture, de réappropriation...
    - le texte n'est plus forcément un objet clos et homogène
    - les rapports « auteur/lecteur » se trouvent modifiées

Avantages des données sur le Web

- En raison de son libre accès
  - moins de contraintes éditoriales que les publications traditionnelles
- Grande variété d'usages et de niveaux de langue
  - terrain d'exploitation privilégié pour explorer les variations linguistiques
- Les données « brutes » sur Internet sont plus structurées que les données constituant la plupart des corpus « classiques »
  - hyperdocuments liés en réseau, constitués de zones fonctionnelles (via les marqueurs HTML ou XML) et très souvent enrichis de méta-descripteurs
- Ces données sont disponibles dans différentes langues
  - corpus alignables ou comparables

Exploitation « difficile » des données

- **Textualité électronique**
  - *« L'écriture électronique n'est pas linéaire. Elle n'est pas livresque. Elle n'est pas textuelle. Elle n'est pas cinématographique. Elle n'est pas photographique. Elle n'est pas graphique. Elle n'est pas stable. Elle n'est pas réflexive. Elle n'est pas picturale. Elle n'est pas sculpturale. Elle n'est pas sonore… »*
    - Hervé Fischer (Québec) : « Paradoxes de la textualité électronique », colloque « Textualités & nouvelles technologies » Musée d'Art Contemporain - Montréal (23-24-25 octobre 2001)
  - Emboîtés/Articulés les uns dans les autres, les composants du discours sur le Web modifient/amplifient les puissances du langage
- **Les moteurs de recherche ne permettent pas encore d'exploiter toute la richesse des données sur le Web**
  - **mode de requête disponible peu adaptée à la complexité des données disponibles**

Applications possibles

- [Jacquemin 2001] présente 3 exemples d ’applications fondées sur les données accessibles en ligne :
  - Constitution de corpus multilingues et alignables
    - Système STRAND : http://umiacs.umd.edu/~resnik/strand/ pour l ’acquisition de paires de documents (traduction l ’un de l ’autre) sur le Web
  - Traduction basée sur des exemples
  - Recherche et acquisition d ’entités nommées
- Pour ce type d’application :
  - Phase 1 : accès aux données du Web
  - Phase 2 : exploitation des données pour en extraire les « connaissances » recherchées

Christian Jacquemin, « L’internet, réservoir de langue infini », éc/art S:#2 [00_01], 2001 www.ecarts.org

Constituer des corpus à partir du web

Le web : un « réservoir » à corpus

Éléments de méthode

Qu'est ce qu'un corpus? (1)

- **Une définition**
  - **« Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage. » (in Habert *et al.* 1997)**

- **Types de corpus**
  - **Corpus spécialisé**
    - **« Il se restreint à une situation de communication, un domaine, une langue de spécialité, c'est-à-dire un langage très spécifique, très contraint du point de vue lexical, syntaxique, voire textuel, que l'on trouve dans les domaines scientifiques et techniques. » (*ibidem)***

      → ex. corpus Ménélas, INRA
  - **Corpus de référence**
    - **« Il est conçu pour fournir une information en profondeur sur une langue. Il vise à être suffisamment grand pour représenter toutes les variétés pertinentes de cette langue et son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres usuels fiables. » (*ibidem)***

      → ex. Brown Corpus, BNC (*British National Corpus*), base textuelle Frantext

Qu’est ce qu’un corpus? (2)

- **Caractéristiques des corpus**
  - **Corpus fermés vs corpus « ouverts »**
  - **Corpus achevés vs corpus de suivi (*monitoring corpus*)**
  - **Corpus éphémères vs corpus « persistants »**

L'essor du web reconfigure la notion de corpus (1)

- Le web entraîne une évolution des corpus et une mutation de la notion de corpus
- Une constante
  - Utilité de recourir aux corpus pour concevoir et valider des outils de TAL
- Le web : une immense source de documents
  - Avantages
    - facilité d'accès à des documents de toute nature (texte, image, vidéo…)
    - facilité d'accès à du texte numérique (vs. longue et fastidieuse numérisation de textes par saisie, par OCR, etc.)
    - mise à disposition d'outils de recherche, de fouille, d'aspiration, …
    - apparente facilité d'exploitation des textes par des outils de TAL

L'essor du web reconfigure la notion de corpus (2)

- **Inconvénients**
  - hétérogénéité des données, des contenus et des formats
  - volume, « mouvance » et personnalisation (relative) des données
  - durée de vie très variable des documents sur le web
  - documents « peu renseignés » : auteur, date de création, date de dernière modification, sources des informations présentées, etc. sont des indications rarement mentionnées
  - particularités du texte web
    - interférence de la mise en forme → articulation texte / non-texte
    - texte fortement bruité : usages propres au web (ex. smileys), changements de langue, scories, etc.
    - registre de langue spécifique (ni oral, ni écrit)
    - rôle primordial de la structure au détriment du contenu textuel

Éléments de méthode

- Avant la constitution, une question essentielle :

«un corpus web pour quoi faire? »

  - utilité de recourir à des corpus web : quelques exemples
    - visée applicative : création, validation et/ou évaluation d'outils et de ressources
      - ex. outil de résumés automatiques de sites web (Amitay & Paris 2001)
      - ex. projet Saphir : constitution d'une terminologie spécialisée (Grabar & Berland 2001)
    - visée descriptive : études de phénomènes linguistiques, typographiques, etc.
      - ex. nombreux travaux sur les spécificités des écritures en ligne (Anis 2001, Beaudouin & Velkovska 1999…)
  - → nécessité de cerner très distinctement son objectif

- L'objectif final gouverne l'ensemble du processus de constitution
  - contraintes internes
    - identification, sélection et rapatriement des données → déf. stratégie d'aspiration
    - traitements et exploitation des données → déf. ensemble des post-traitements
  - contraintes externes
    - faisabilité humaine, technique, juridique et temporelle

Identification et sélection des données

- **Origine des données**
  - site web, forum, chat, etc.
- **Nature des données**
  - texte, image, son, vidéo, etc.
- **Unité minimale des données**
  - site, page, extrait de pages, phrase, mots, balises, etc.

Rapatriement des données (1)

- **Utilisation d'aspirateurs web existant (en anglais *browser off line*) ou création d'un logiciel *ad-hoc* selon les besoins**
- **Principe général**
  - **logiciel qui se charge de recopier une partie ou l'intégralité d'un site web sur un disque dur → création d'un miroir local**
- **Intérêt**
  - **accès à tout ou partie d'un site en local lorsque l'ordinateur est déconnecté**
  - **sauvegarde des traces de tout ou partie d'un site à une date donnée**
- **Importance du choix de l'aspirateur**
  - **théoriquement le site aspiré doit être une copie conforme de l'original, mais dans la réalité, il est rare d'obtenir une copie conforme de l'original qui se trouve en ligne (images manquantes, liens non opérationnels, etc.)**
    → la qualité des aspirations dépend de l'outil utilisé

Rapatriement des données (2)

- comparaison de trois aspirateurs (Maisondieu & Kuncova 2000)

	MemoWeb3	HTTrack 3.1	Teleport Pro 1.29
vitesse d'aspiration	rapide	rapide	très rapide
environnement	Windows	Windows/Linux	Windows
mise à jour du logiciel	oui	oui (très fréquente)	oui
disponibilité	payant	gratuit	payant
chargement liste URL	**oui**	**oui**	non
plannificateur de capture	non	**oui**	non
mise à jour des sites	-	oui	-

- sélection d'un aspirateur : WinHTTrack (http://www.httrack.com)
  - \+ mises à jour très fréquentes du robot
  - \+ gratuité
  - \+ nombreux paramétrages du robot (filtres sur les éléments, index « sémantique », etc.)
  - \+ capture quasiment tous les sites (requêtes tolérantes auprès des serveurs)
  - \+ possibilité de planifier les captures
  - \+ chargement de fichiers textes contenant une liste d'URL
  - \- inapproprié pour faire de la veille (supprime et remplace les captures antérieures)

Rapatriement des données (3)

- **Problèmes techniques rencontrés**
  - **stockage des données**
  - **difficultés techniques liées aux architectures mises en place pour construire les sites (scripts, frames, accès par mot de passe, etc.)**
  - **temps d'aspiration très variables**

	Nb de pages	Durée aspiration	% de récupération	Nb de page/minute	Taille des fichiers (en Mo)
home-nordnet	**122**	**40 sec.**	**99**	**183**	
free	**3 687**	**24 min.**	**86**	**131**	**13,3**
multimania	**2856**	**54 min.**	**98**	**52**	**21,7**
www.fnac.com (5 niveaux)	**41 302**	**3j 6h 10 min.**	**78**	**0,9**	**42,8**

Exemple de temps d'aspiration de pages et sites web avec WinHTTrack

Traitements et exploitation des données

- **Rendre le corpus exploitable → normalisation du corpus**
  - principe : il s'agit de convertir tous les documents aspirés dans un même format
  - choix et/ou création d'un modèle de document
    - ex. extraction des parties textuelles → format texte
    - ex. représentation arborescente des divers éléments d'une page→ format XML
    - ....
- **Rendre le corpus réutilisable → documentation du corpus**
  - principe : il s'agit d'adjoindre au corpus et aux documents qui le compose un descriptif
  - possibilité d'insérer ces descriptifs sous forme d'un « cartouche descriptif »
    - sur la totalité du corpus
    - sur chaque document composant le corpus
  - Informations à consigner dans les descriptifs au cours de la constitution
    - informations portant sur le document : titre, auteur, taille, langue, etc.
    - informations portant sur la collecte : aspirateur utilisé, date d'aspiration, etc.
    - informations portant sur les modifications issues de la normalisation
    - informations portant sur les éventuelles annotations du corpus

TypWeb

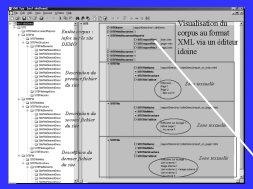
Traits textuels, structurels et présentationnels pour typer les sites web personnels et marchands

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

Équipe TyPWeb (FT R&D/DIH/UCE, Paris III, LIMSI & Paris X)

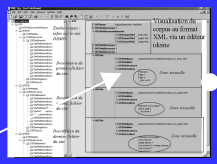
Valérie Beaudouin, Serge Fleury, Benoît Habert, Gabriel Illouz, Christian Licoppe, Marie Pasquier

TyPWeb



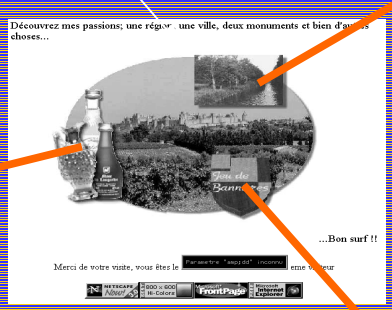
Page d'accueil

Thème 1
La ville de Carcassonne

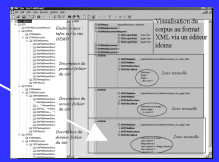


Visualisation XML
des indicateurs
- structurels
- morpho-syntaxiques
- sémantiques

Thème 2
La collection de mignonnettes



Thème 3
Liens vers d'autres sites (amis, centres d 'intérêts, …)



Liens externes

- **Analyse de la structure formelle et sémantique des sites**
- **Discours des concepteurs de sites web**
- **Analyse des parcours**

Sommaire

- Présentation du projet TyPWeb
  - Contexte et hypothèses
- Méthodologie générale et chaîne de traitements
- Résultats : présentation d ’ un exemple d’analyse de sites
  - Exploration de corpus de sites personnels et marchands
    - Objectif : distinguer deux grandes catégories de sites

Projet TyPWeb (LOT 1)

Objectifs

- Analyser finement la structure formelle et sémantique de sites web afin de mieux comprendre le projet de conception qui le sous-tend
  - Projet de vente électronique dans les sites marchands
  - Projet de présentation et mise en scène de soi, de ses centres d'intérêts et de ses passions dans les sites personnels
- Mettre au point une méthodologie de description de sites à l'aide de traits structurels, présentationnels et textuels
- Typer les sites afin de décrire et catégoriser des parcours sur le web

Comment ?

- Transposer et adapter à des corpus web l'architecture de typologie des textes TyPTex (Folch *et al.*, 2000)
- Décrire la structure et le contenu des sites web
  - Identifier des types de pages et de sites sur la base de traits

Projet TyPWeb (LOT 1)

Comment analyser le contenu textuel des sites ?

plusieurs approches d'analyse possibles

- **Possibilité de recourir au site lui-même ou à des textes périphériques qui le décrivent**
  - **Textes périphériques au site**
    - **Utilisation des descriptions d'annuaires (Assadi & Beauvisage 2002)**
    - **Utilisation des descriptions de moteurs de recherche (Amitay & Paris 2001)**
    - **Utilisation des pages de liens qui pointent vers la page ou le site analysé (*ibidem*)**
  - **Éléments internes aux pages et/ou aux sites → approche TyPWeb**
    - **Nom de fichier HTML**
    - **Balises META**
    - **Contenu textuel**

Identification de traits multimédia

traits structurels

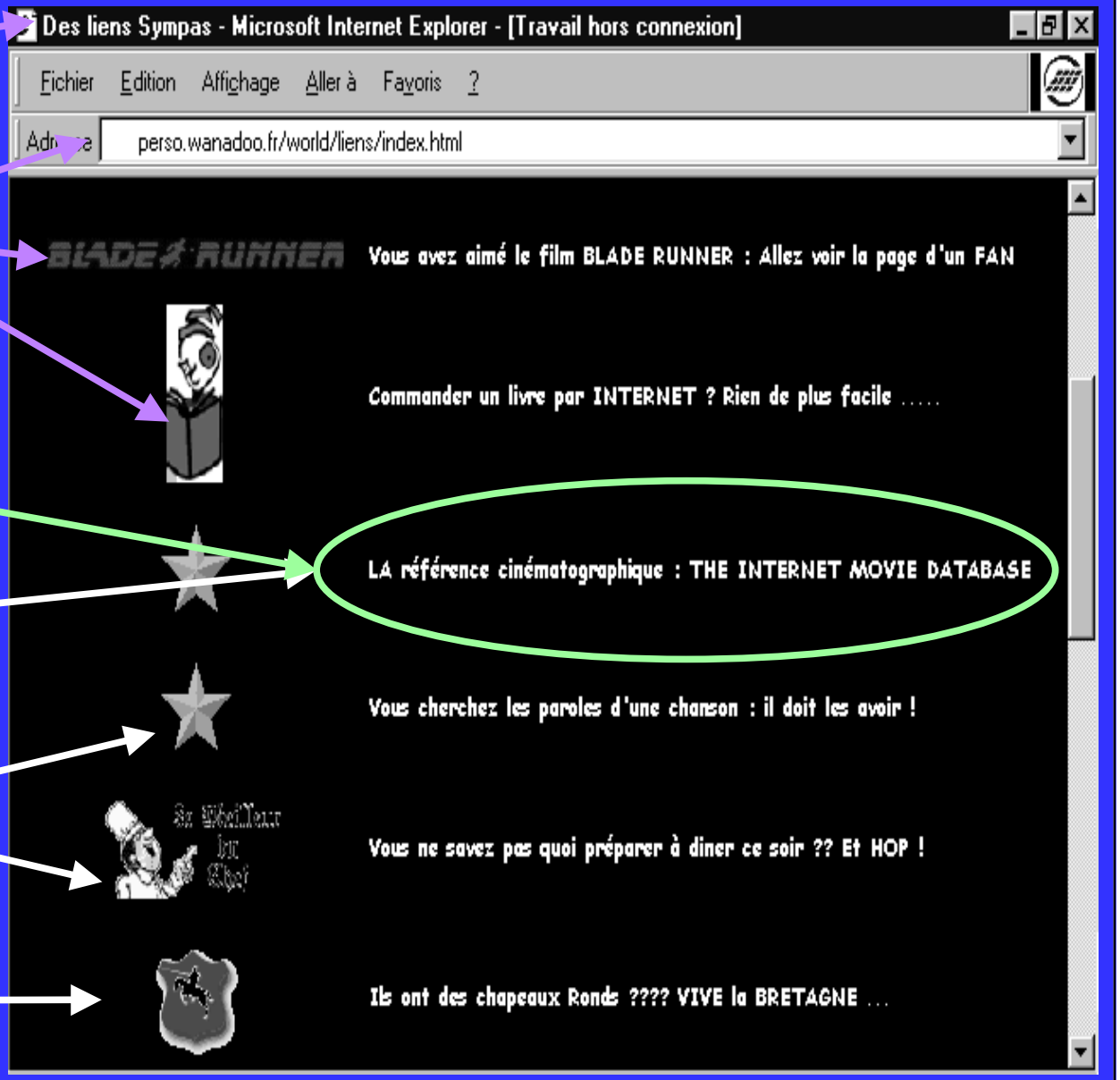
- titre
- liens
  interne et externe

traits textuels

- zone textuelle
  - ➢ contenu :
    LA, référence…
  - ➢ aspect :
    police, taille, couleur

traits présentationnels

- images
  taille, format
- fond d'écran
  couleur, motif,...

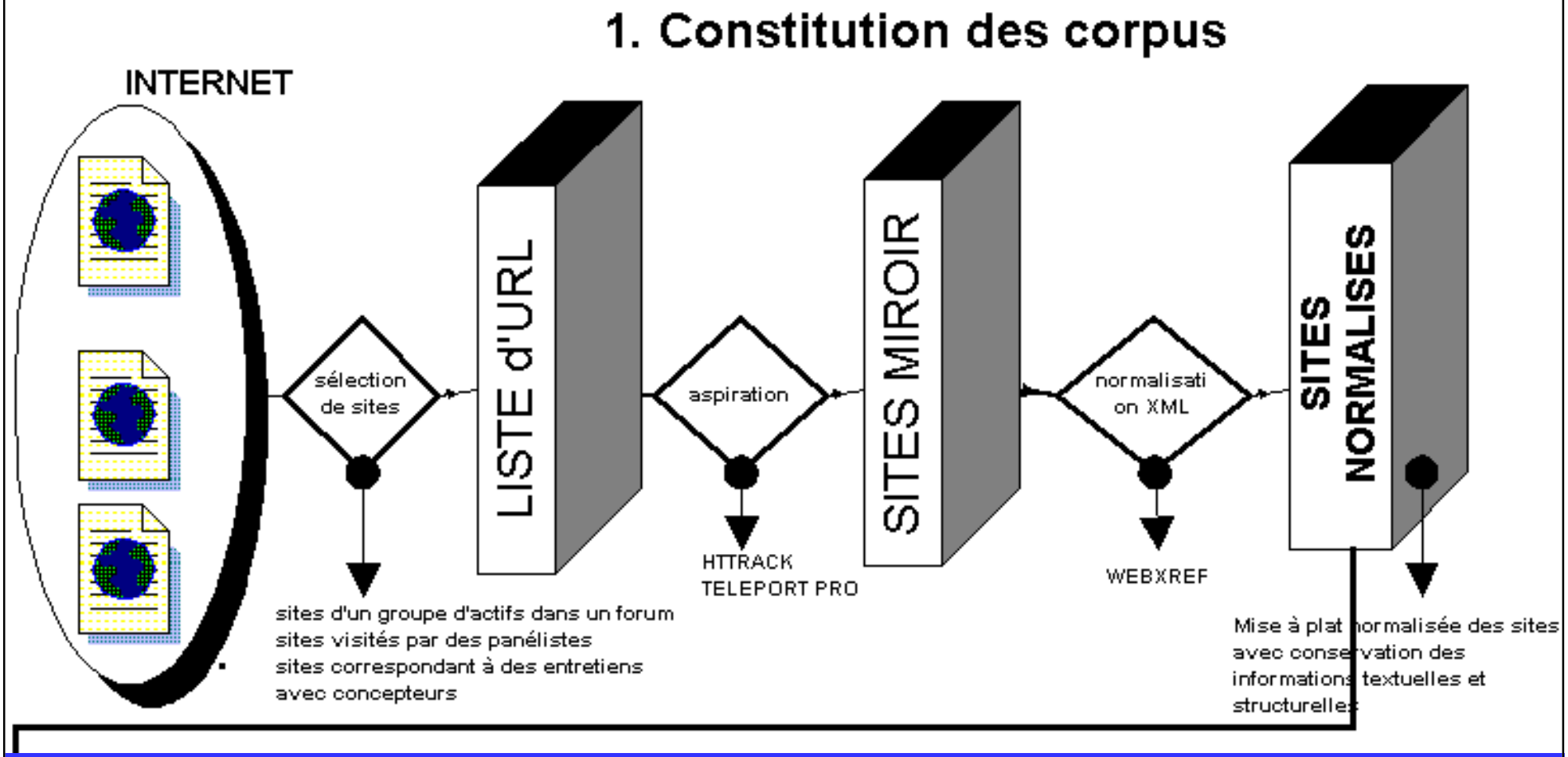


Méthodologie et chaîne de traitements

- **Méthodologie retenue pour l'analyse de sites web**
  - **Prise en compte la structure hypertextuelle des pages et des sites → identification de traits structurels, présentationnels et textuels permettant de décrire les pages et les sites**
  - **Typer les pages et les sites avec les traits retenus afin d'obtenir des répartitions de pages et de sites pertinents**

- **Chaîne de traitements**
  - **Définition de critères de constitution de corpus web**
  - **Aspiration de pages ou de sites**
  - **Mise en forme et normalisation des corpus avec outil webXref → corpus XML**
  - **Deux types d'approches pour identifier des traits pertinents**
    - **Approche hypothético-déductive → identification des spécificités de catégories données *a priori***
      - ***Sites marchands versus sites personnels,***
      - ***Secteurs économiques des sites marchands,***
      - ***Services d'hébergement des pages perso, …***
    - **Approche inductive → identification de types de pages ou de sites sur la base de leur forme et de leur contenu**
  - **Traitements avec des outils de statistiques générales (R, SAS) et de statistiques textuelles (Lexico 3, Alceste, Tropes)**

Phases d’aspiration et de normalisation



Normalisation des sites

- Un corpus XML de référence
  - réorganisation normalisée des éléments (structurels et textuels) qui composent les pages originales
- Des états statistiques sur la composition des pages de ce corpus (comptage de mots, d'éléments HTML : TAG, attributs, valeurs d'attributs…) et des états statistiques sur les sites
  - très grand nombre de traits obtenus par corpus (de l'ordre de 200000) ➔ difficulté de traitements par les outils de statistiques disponibles
  - sélection pour chaque étude particulière de sous-ensemble de traits dans la liste des traits (ex. pronoms personnels)
- Des corpus construits à partir de telles sélections :
  - Par ex.: corpus « textuel », corpus de TAG HTML, etc.
- Les représentations de ces corpus sont ensuite soumises, après reformatage, à des outils d'analyse tels que R, SAS, Lexico3, Alceste, Tropes, Cordial, …
- Recherche d'informations sur l'arbre XML avec XPath ou filtrage via programmes spécifiques (scripts Perl)

Corpus XML après normalisation (1)

```
<SITE>
<SITEName> jura.speleo1 </SITEName>
<SITEFile>
<SITEFileName>/windows/C/SFleury/Recherche/Typweb/Typweb2sql/data/sites/jura.spele\
o1/index.html</SITEFileName>
<SITEReportFileName>index.html</SITEReportFileName>
<HEADER NUM="1">TITLE: Jura Speleo</HEADER>
<HEADER NUM="2">META Name: keywords Content: speleo, grotte, gouffre, karst, Jura,\
 Franche-Comte</HEADER>
<HEADER NUM="3">META Name: author Content: Frachon Jean-Claude</HEADER>
<HEADER NUM="4">META Name: description Content: La speleologie dans le departement\
 du Jura (France): grottes, gouffres, karst</HEADER>
<SCRIPTCONTENT NUM="1" VALUE="INDISPONIBLE"/>
<tagHTML TAGType="HTML" NBATTR="0">BEGIN-HTML
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<COMMENT>Mirrored from perso.wanadoo.fr by HTTrack/1.x [RX/PY'99] </COMMENT>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="HEAD" NBATTR="0">BEGIN-HEAD
</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="TITLE" NBATTR="0">BEGIN-TITLE
</tagHTML>
<SITEFileTxtBrut>
Jura Speleo
</SITEFileTxtBrut>
<tagHTML TAGType="title">END-title</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="META" NBATTR="2">BEGIN-META
<tagHTMLAttr TAG="META" NUM="2" ATTRType="Name" VALUE="keywords"/>
</tagHTML>
<tagHTML TAGType="META" NBATTR="2">BEGIN-META
<tagHTMLAttr TAG="META" NUM="1" ATTRType="Content" VALUE="Frachon Jean-Claude"/>
<tagHTMLAttr TAG="META" NUM="2" ATTRType="Name" VALUE="author"/>
</tagHTML>
<tagHTML TAGType="META" NBATTR="2">BEGIN-META
<tagHTMLAttr TAG="META" NUM="1" ATTRType="Content" VALUE="La speleologie dans le d\
epartement du Jura (France): grottes, gouffres, karst"/>
<tagHTMLAttr TAG="META" NUM="2" ATTRType="Name" VALUE="description"/>
</tagHTML>
<tagHTML TAGType="SCRIPT" NBATTR="1">BEGIN-SCRIPT
<tagHTMLAttr TAG="SCRIPT" NUM="1" ATTRType="LANGUAGE" VALUE="JavaScript"/>
-(Unix)--  CORP-J~1.XML   5:17PM   (XML [SITE])--L??-- 1%--------------------------
```

Nom du site

Méta-informations sur fichier (balises META)

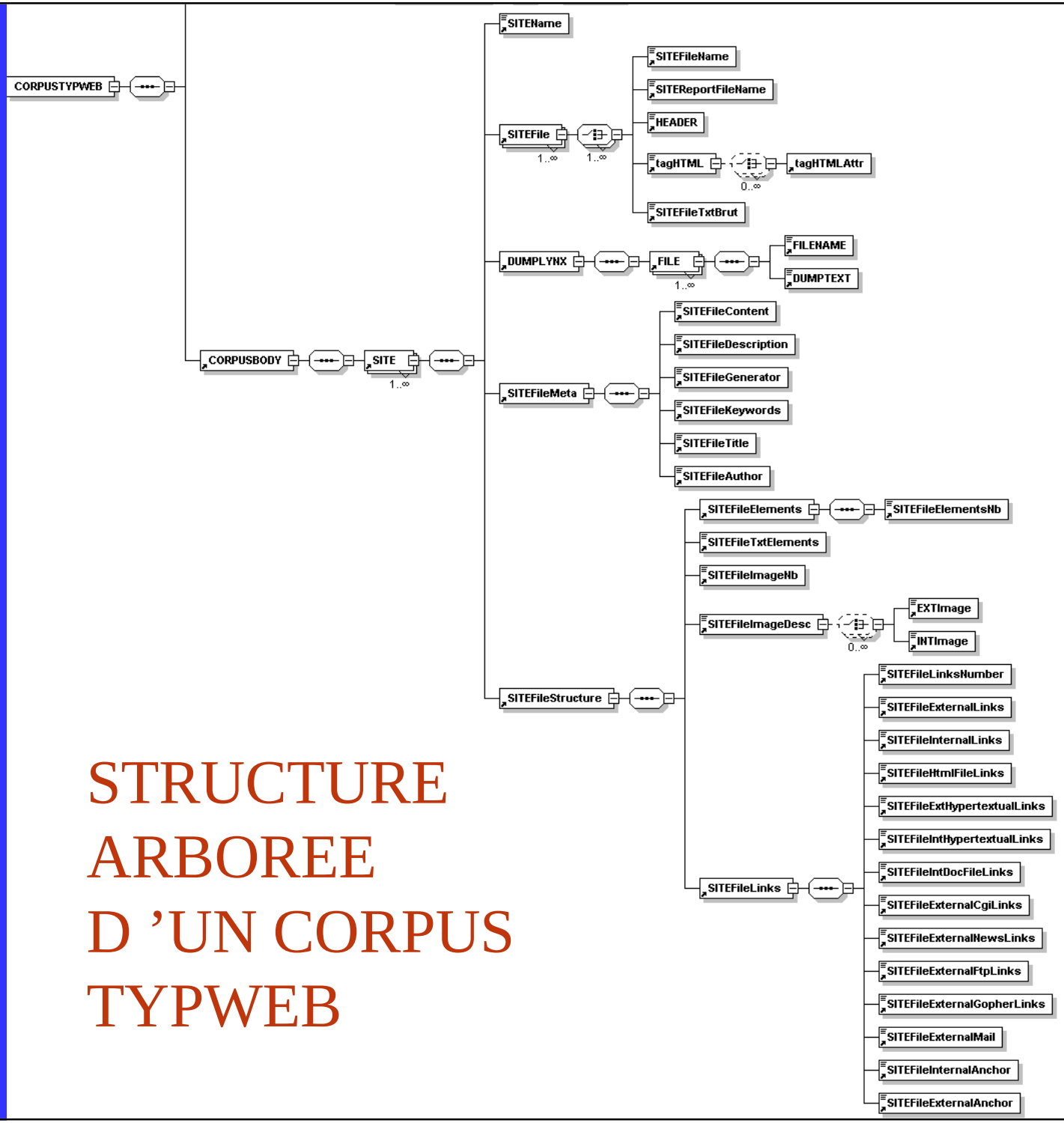
Description séquentielle de tous les éléments composant le fichier original (zones textuelles, informations structurelles, liens…)

Description de chaque fichier du site

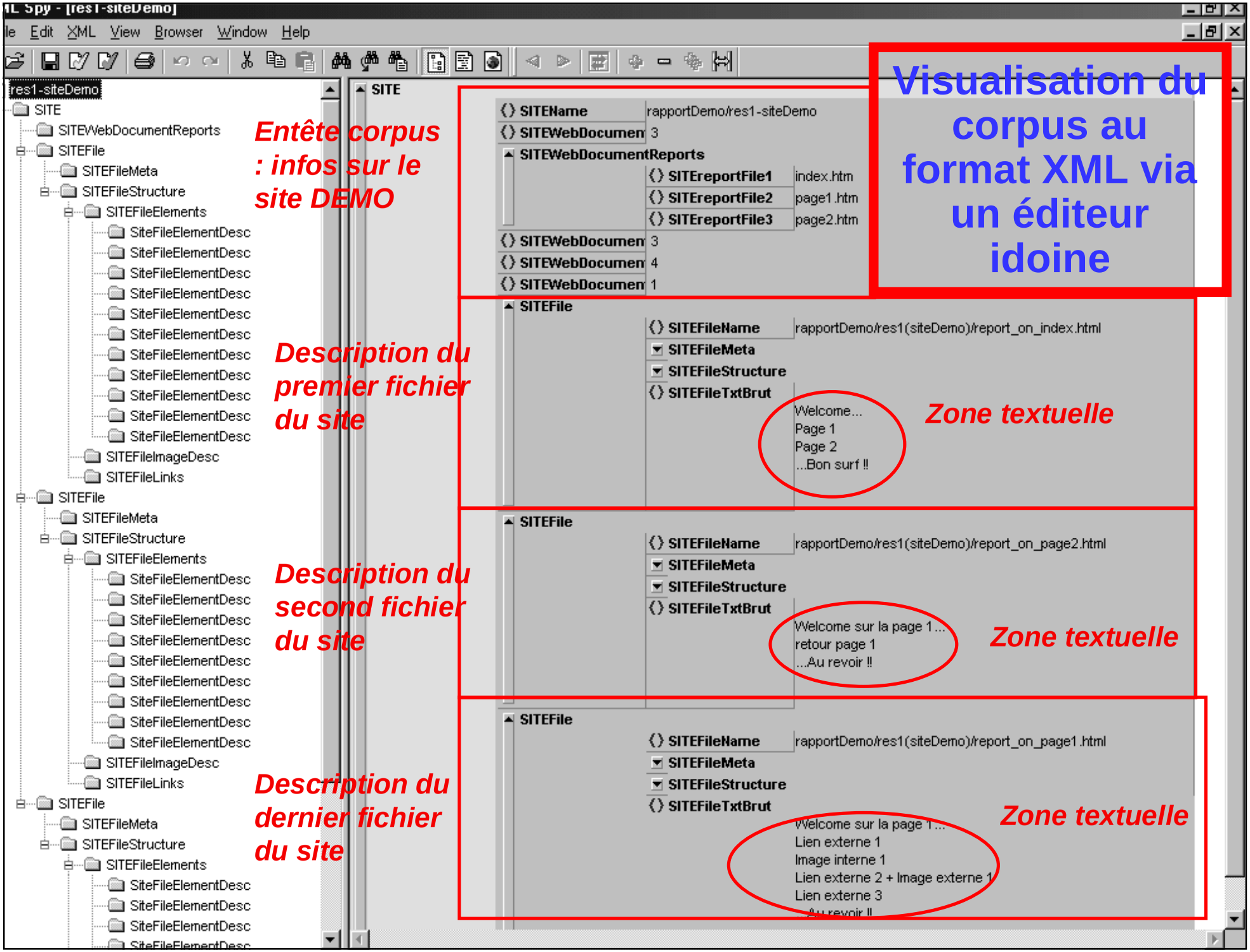
Corpus XML après normalisation (2)

```
<tagHTMLAttr TAG="IMG" NUM="7" ATTRType="SRC" VALUE="../images/barrenav.gif"/>
<tagHTMLAttr TAG="IMG" NUM="8" ATTRType="USEMAP" VALUE="#map1"/>
<tagHTMLAttr TAG="IMG" NUM="9" ATTRType="WIDTH" VALUE="340"/>
</tagHTML>
<tagHTML TAGType="center">END-center</tagHTML>
<tagHTML TAGType="p">END-p</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<tagHTML TAGType="body">END-body</tagHTML>
<SITEFileTxtBrut TYPE="BLANKSPACE"> </SITEFileTxtBrut>
<COMMENT>Mirrored from perso.wanadoo.fr by HTTrack/1.x [RX/PY'99] </COMMENT>
<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_ANCHOR" TAG="A"/>
<tagHTML  TAGType="LINK" NUM="4" TYPELink="INTERNAL_IMAGE" TAG="IMG"/>
</SITEFile>
<DUMPLYNX>
<FILE>
<FILENAME>/windows/C/SFleury/Recherche/Typweb/Typweb2sql/data/sites/jura.speleo1/index.html</FILENAM\
E>
<DUMPTEXT>

    Porche

                                Bienvenue sur

                               [1] Jura Speleo

                      le site des sp\351l\351ologues jurassiens

                                Jura (France)
```

« Isolation » des zones textuelles pour analyses textuelles traditionnelles



STRUCTURE ARBOREE D ’UN CORPUS TYPWEB


Visualisation du corpus au format XML via un éditeur idoine
Entête corpus : infos sur le site DEMO
Description du premier fichier du site
Description du second fichier du site
Description du dernier fichier du site
Zone textuelle
Zone textuelle
Zone textuelle
SITE
SITEName
rapportDemo/res1-siteDemo
SITEWebDocumentReports
SITEreportFile1
index.htm
SITEreportFile2
page1.htm
SITEreportFile3
page2.htm
SITEFile
SITEFileName
rapportDemo/res1(siteDemo)/report_on_index.html
SITEFileMeta
SITEFileStructure
SITEFileTxtBrut
Welcome...
Page 1
Page 2
...Bon surf !!
SITEFile
SITEFileName
rapportDemo/res1(siteDemo)/report_on_page2.html
SITEFileMeta
SITEFileStructure
SITEFileTxtBrut
Welcome sur la page 1...
retour page 1
...Au revoir !!
SITEFile
SITEFileName
rapportDemo/res1(siteDemo)/report_on_page1.html
SITEFileMeta
SITEFileStructure
SITEFileTxtBrut
Welcome sur la page 1...
Lien externe 1
Image interne 1
Lien externe 2 + Image externe 1
Lien externe 3
...Au revoir !!

Requête Xpath sur corpus Typweb

- **Recherche des zones de texte après une balise BOLD (<B>)**

```
//tagHTML[@TAGType='B']/following-
  sibling::SITEFileTxtBrut[1]
```

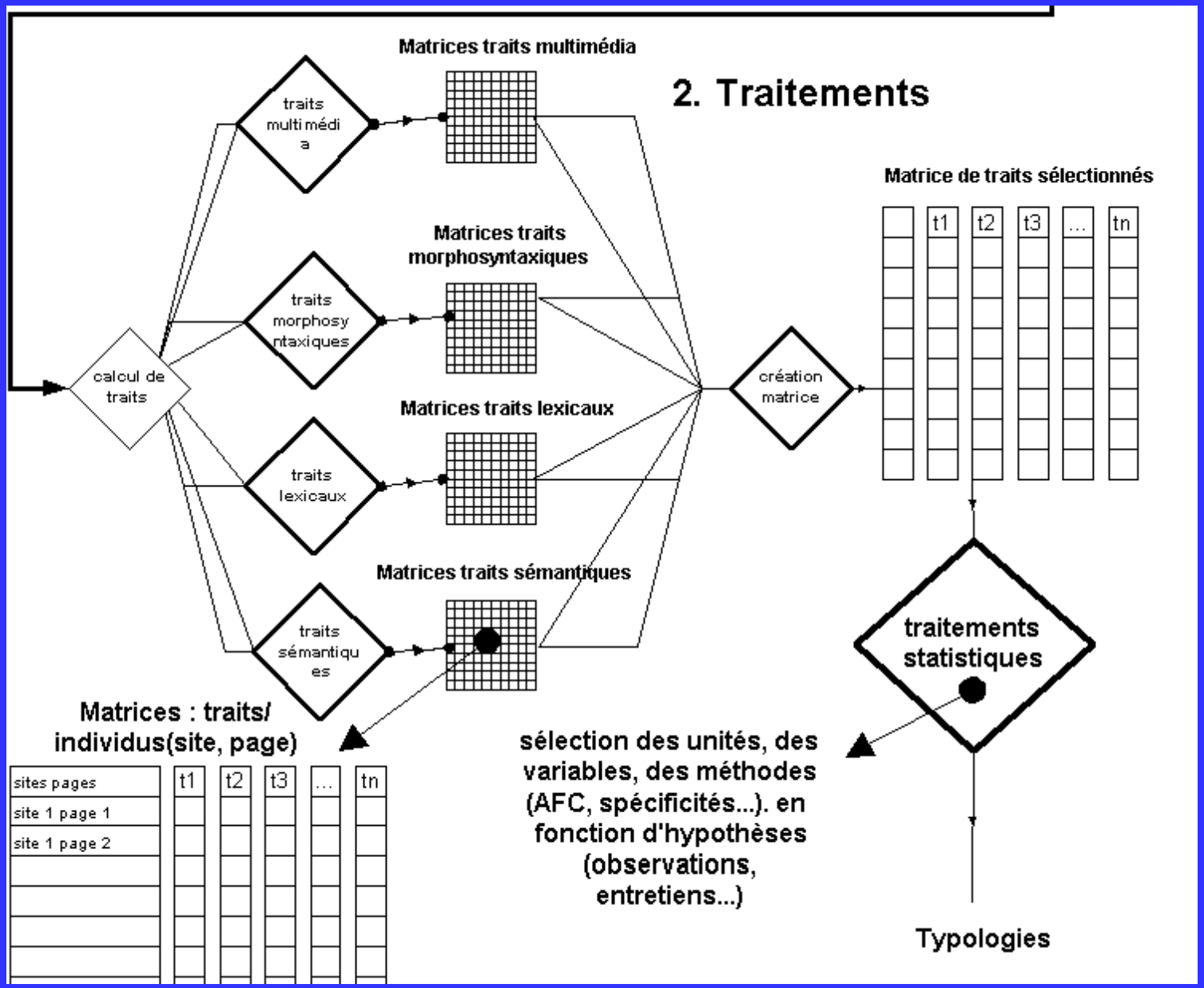
- **Recherche des zones de texte situées après une balise FONT (<FONT>)**
  - Zones textuelles avec POLICE TAILLE=+4 et FACE=Trajan

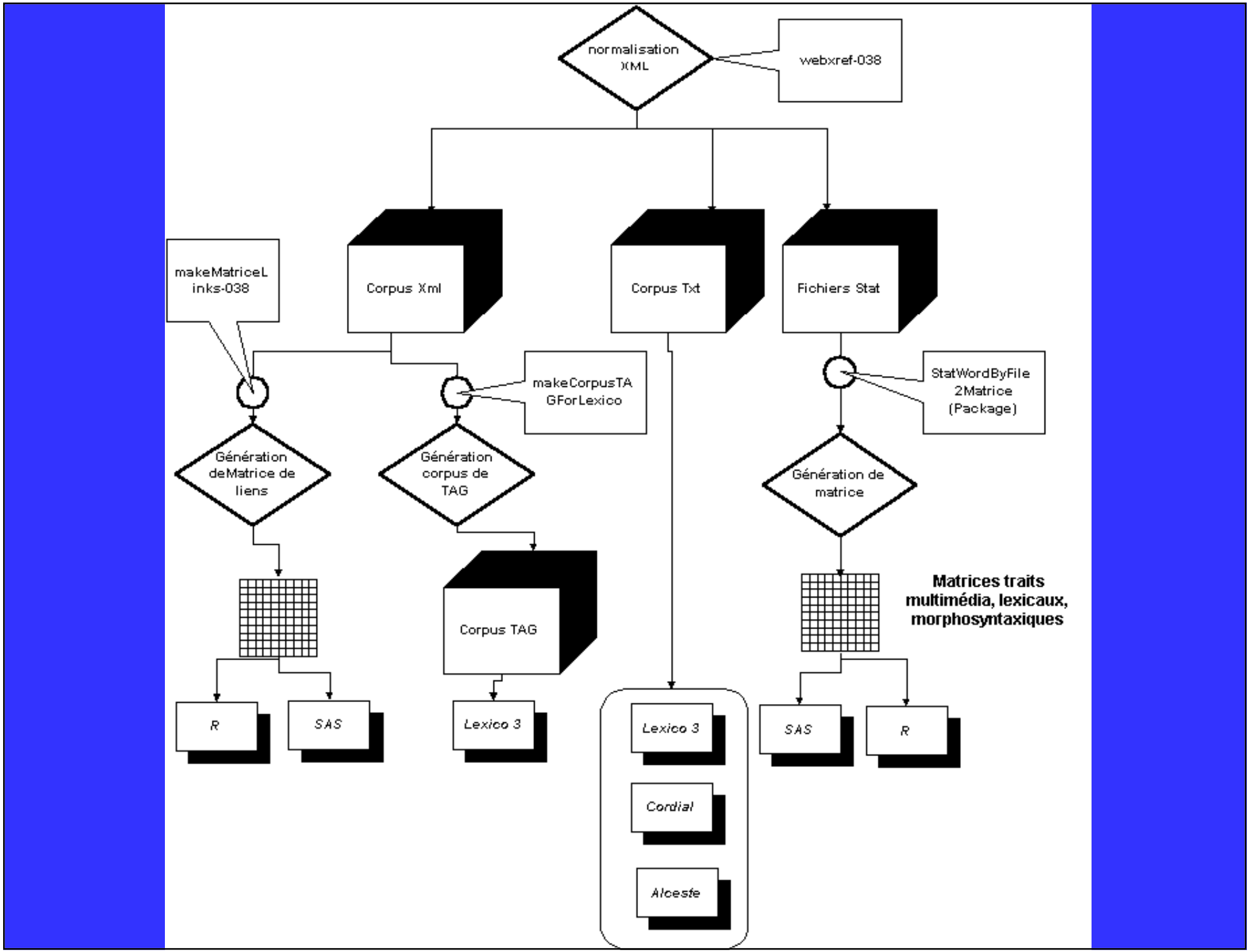
```
//tagHTML[@TAGType='FONT'][(./tagHTMLAttr[@ATTRType='F
  ACE' and @VALUE='Trajan']) and
  (./tagHTMLAttr[@ATTRType='SIZE' and
  @VALUE='+4'])]/following-sibling::SITEFileTxtBrut[1]
```

- **Recherche des zones de texte situées après une balise LIEN (<A>)**
  - Zones textuelles contenues dans une définition de lien de type HREF vers un fichier nommé "conseil_municipal.htm

```
//tagHTML[@TAGType='A'][./tagHTMLAttr[@ATTRType='HREF'
  and @VALUE='conseil_municipal.htm']]/following-
  sibling::SITEFileTxtBrut[1]
```

Traitements (1)




normalisation XML
webxref-038
makeMatriceLinks-038
Corpus Xml
Corpus Txt
Fichiers Stat
makeCorpusTAGForLexico
StatWordByFile2Matrice (Package)
Génération deMatrice de liens
Génération corpus de TAG
Génération de matrice
Corpus TAG
Matrices traits multimédia, lexicaux, morphosyntaxiques
R
SAS
Lexico 3
Lexico 3
Cordial
Alceste
SAS
R

Données analysées

- De la base XML aux corpus construits et adaptés aux outils d'analyse
  - des corpus construits à partir de sélections
    - corpus « textuel »
    - corpus de TAG HTML
    - États statistiques : liens, tags
  - Analyses réalisées
    - statistiques générales : R, SAS
    - statistiques textuelles : Lexico3, Alceste, Tropes
    - étiqueteurs : Cordial, TILT

Résistance et complexité des données

- Grand nombre de traits « bruts »
  - **de l'ordre de 200 000 sur l'ensemble des corpus**
  - **construction de matrice de traits ingérable par les outils disponibles**
- Complexité des éléments structurels/présentationnels
  - **exemple : élément `FONT` (police) peut être caractérisé par le type de police, par sa taille, sa couleur…**
  - **reste à trouver pour chaque attribut de ce type les découpages pertinents**
- Complexité des éléments textuels
  - **éclatement local du texte par les balises (lettrines par exemple)**
  - **« sauts » d'une langue à l'autre (surtout avec l'anglais)**
  - **faible proportion de texte dans bon nombre de pages**

Des traits « bruts » aux traits « pertinents »

- **Traits structurels / présentationnels**
  - complexité des sites
  - redondance des liens
- **Traits textuels**
  - répartition des pronoms personnels
  - spécificité de vocabulaire par hébergeur

Exploration de corpus de sites marchands et personnels

- Choix des sites
- Présentation du corpus
- Traitements et résultats

Sélection des sites

Sites Marchands

- **enjeu majeur pour l'économie d'internet**
- **secteur en évolution permanente**
- **étude qualitative de grande ampleur menée sur les sites marchands (Licoppe 2000)**

Sites Personnels

- **rôle essentiel dans les interactions électroniques (Beaudouin & Velkovska 1999) : lieu stable de présentation de soi et du réseau de relations**
- **services d'hébergement de pages personnelles : très visités (données NetValue)**
- **contenu et forme moins connus**

sélection de catégories de sites très contrastées
pour mettre à jour leurs spécificités

Corpus

Corpus de sites personnels

- 539 sites wanadoo des membres actifs d'un forum (été 99)
- 568 sites wanadoo visités en mars 2000 par des internautes du panel NetValue (été 2000)
  - pas de recouvrement avec le premier corpus

Corpus de sites marchands

- environ 50 sites marchands
- 4 états du corpus : 1999/2000, été 2000, mars 2001 et début 2002
- diverses thématiques des corpus : assurance, voyage, vente par correspondance, informatique, etc.

Difficultés de constitution (sites marchands)

- volume des données collectées
- difficultés d'accès : règles *no robots*, scripts cgi, fichiers flash, etc.

Données quantitatives

	Sites Personnels		Sites Marchands		
	PP-Ete99	PP-mars00	SM-99-00	SM-été00	TOTAL
nb de sites	539	568	29	16	1 162
nb de pages	11 006	24 938	29 199	5 726	96 885
moy. pages/site	20	44	1 007	358	83
nb occurrences	3 878 647	10 577 421	3 090 399	1 284 664	18 831 131
nb de formes	148 360	348 092	66 635	53 805	616 892
nb d'éléments HTML					13 882 836
Nb de formes HTML					349
fichiers XML (en ko)	292 074	1 029 274	450 433	159 434	

Traitements

- **Objectif : discriminer des catégories de sites connues *a priori***
  - Unité minimale : le site
  - Répartition sites marchands versus sites personnels
- **Méthodologie**
  - Formulation d'hypothèses en fonction de connaissances externes aux corpus
  - Identification de traits pertinents
  - Traitements statistiques
- **Résultats : identification de traits discriminants**
  - Traits structurels et présentationnels : complexité des sites et redondance des liens
  - Traits textuels : répartition des pronoms personnels
- **Difficultés rencontrées**
  - Taille des corpus → difficultés et longueurs des traitements
  - Complexité de découpage des éléments structurels et présentationnels
  - Complexité des éléments textuels → saut de langue, faible proportion de texte, etc.

Complexité des sites

- La structure des sites marchands est plus complexe que celle des sites personnels
  - en moyenne 20 fois plus de pages dans les sites marchands ➔ donc plus de liens
  - même nombre moyen de liens par page
- La structure des sites personnels visités par un panel d'internautes est plus complexe que celle des sites des « habitués » d'un forum
  - 44 pages par site contre 25
  - 1,9 liens externes par page contre 1,2
  - 10 liens internes contre 6
  - même nombre moyen de liens par page

Corrélation entre taille des sites et audience

Redondance des liens

- sur les sites marchands : contraste important entre pages d'accueil et les autres pages

	Page d'accueil	Autres pages
Liens internes	11	4
Liens externes	3	0,1
Redondance des liens	1,25	1,1

  - Calcul redondance (R=nombre de liens total divisé par le nombre de liens différents)
- sur les sites personnels : pas de contraste pages d'accueil et autres pages

Site marchand : portail qui donne accès à un espace fermé sur l'extérieur

Site personnel : espace ouvert interconnecté avec d'autres sites personnels

Répartition des pronoms personnels

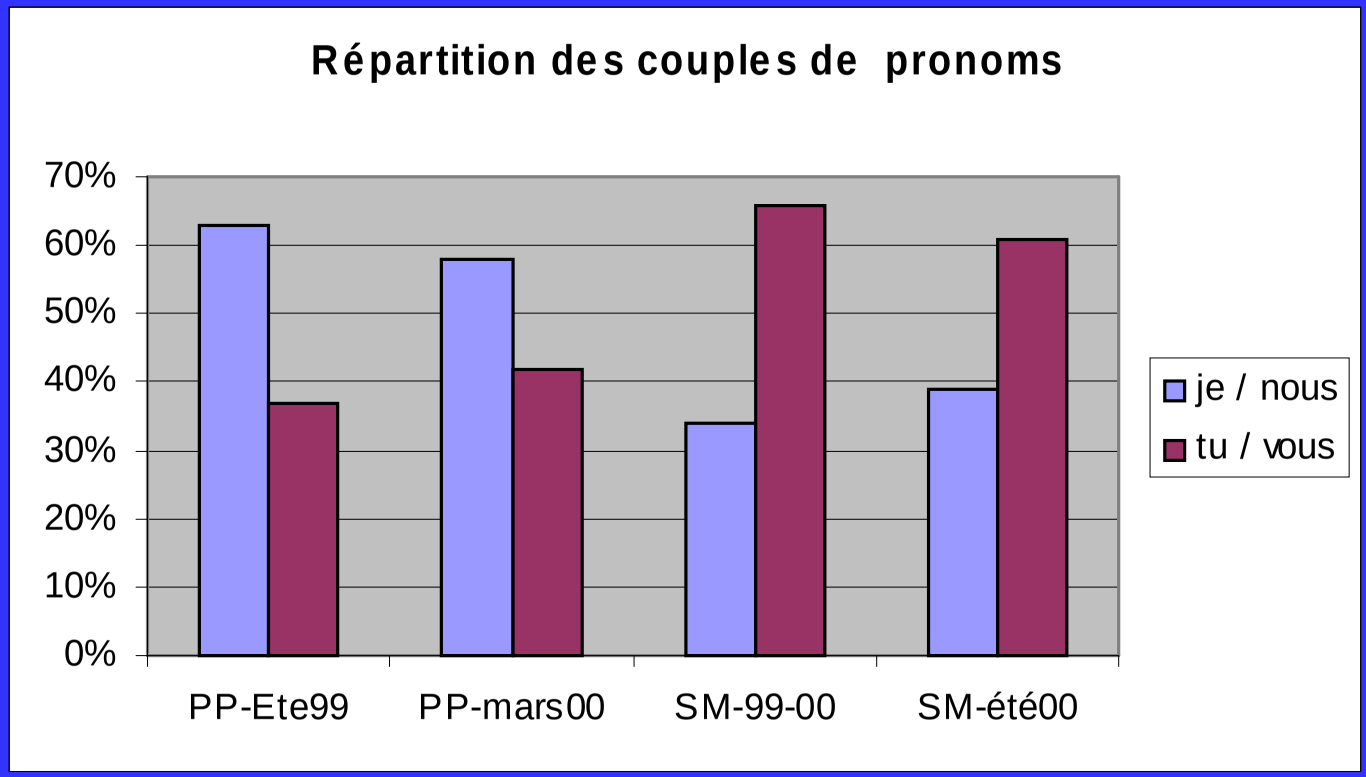
- **Utilisation très différenciée selon les types de sites**
- **6 catégories de pronoms personnels fondées sur le nombre et la personne : choix**
  - **absence des pronoms *le*, *la*, *les* trop ambigus**
  - ***on* et *soi* classés dans la 3ème pers. du singulier**
  - **sites contenant au moins 10 pronoms**

	TOTAL (705 sites)	PP-Ete99 (239 sites)	PP-mars00 (430 sites)	SM-99-00 (22 sites)	SM-été00 (14 sites)
je/me/moi	24	26	23	13	14
tu/te/toi	5	5	5	2	3
il/elle/on/lui/leur/soi	34	35	34	25	33
nous	10	10	10	11	11
vous	21	18	20	44	34
ils/elles/leurs	6	6	6	5	5
Total	100	100	100	100	100

Répartition des pronoms personnels (en %)

Répartition de couples de pronoms

- **Répartition des couples de pronoms des 1ère et 2ème pers**
  - **nette rupture entre sites marchands et personnels**
  - **mise en évidence des spécificités des parties de corpus**



Sites Personnels Sites Marchands

Corpus textuel PP-SM pour Lexico

Volume CorpTxt

```
allSitesTxt2
27418209 27418209 523721 19499580 522600 718570 258375 5000000 18 2 1102 0 0
*** Résultat de la segmentation du fichier: allSitesTxt2.TXT ***
Délimiteurs .,:;!?/_-\"'()[]{}ß$
    nombre des occurrences :    19499580
    nombre des formes      :      522600
    frequence maximale     :      718570
    nombre des hapax       :      258375
    nombre des clés(type)  :           2
    nombre des clés(ctnu)  :        1102
*** Fin de la segmentation du fichier:      allSitesTxt2.TXT ***
```

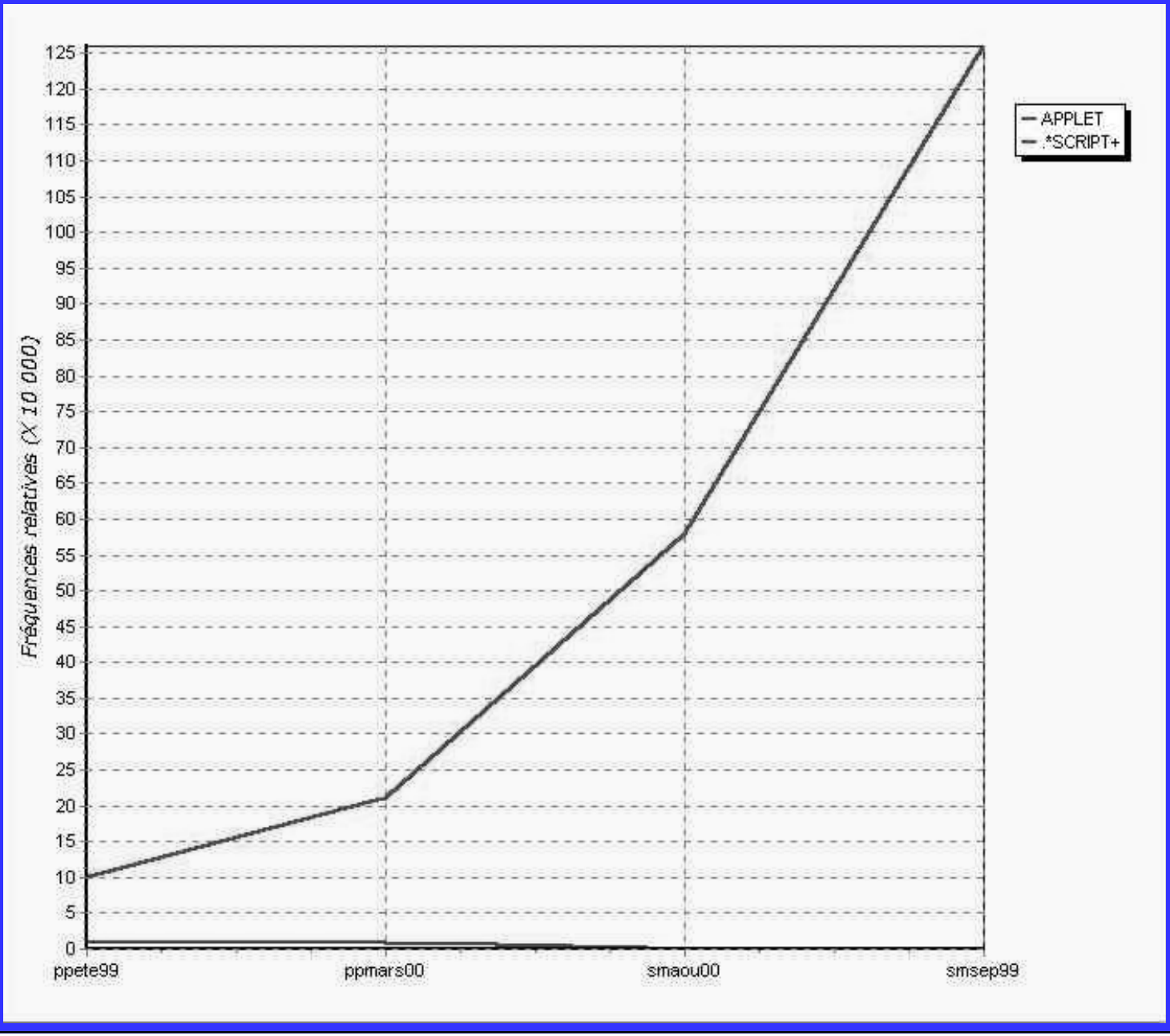
	Partie	Occurenc	Formes	Hapax	Fmax	Forme
✓	ppete99	3990797	170072	76030	172132	de
✓	ppmar00	10808752	407366	215078	371402	de
✓	smaou00	1343898	64115	26147	51367	de
✓	smfev00	9534	3031	1533	288	de
✓	smsep99	3346599	87960	27754	123381	de

Corpus PP-SM de TAG HTML pour Lexico

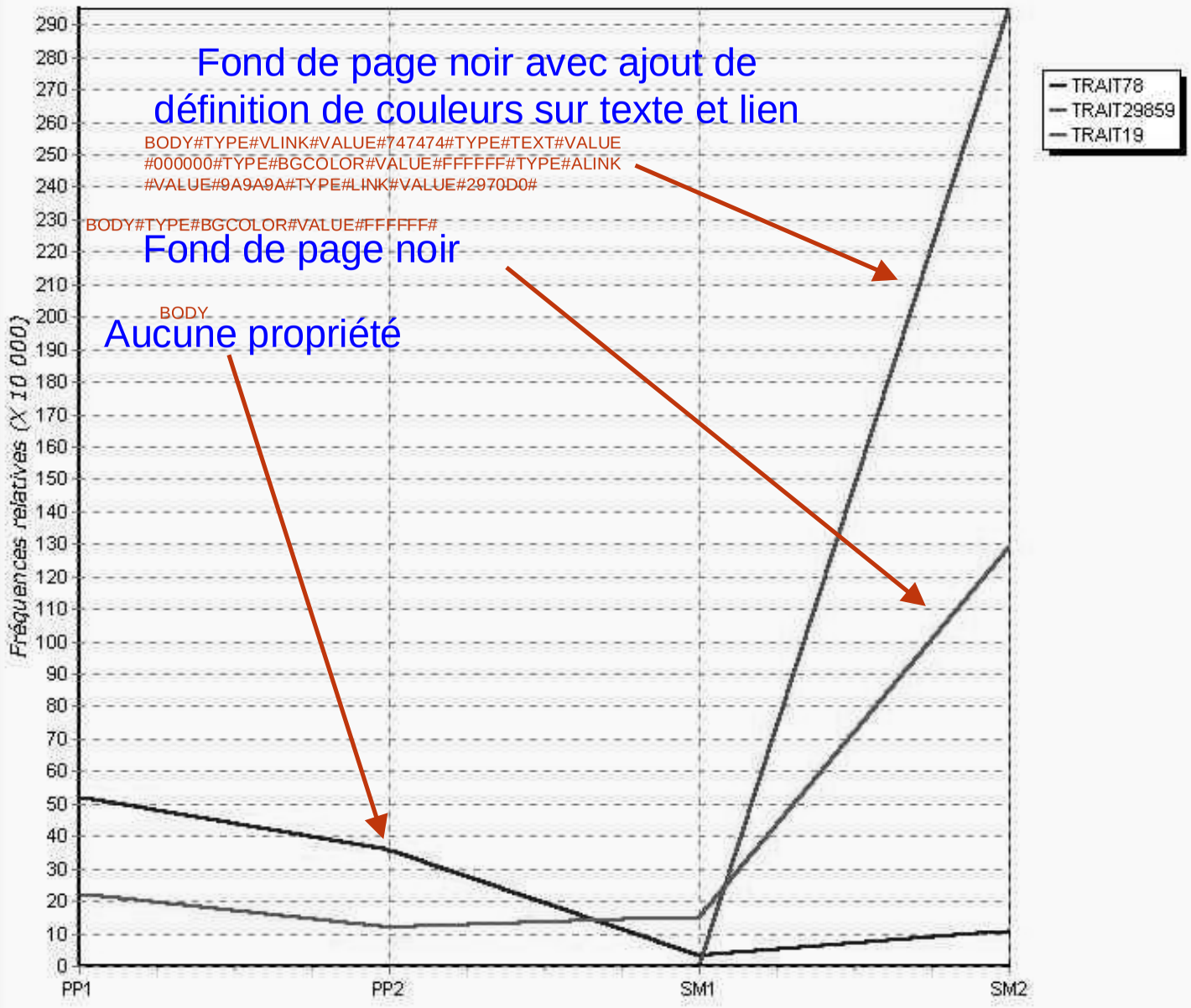
```
Volume CorpTag
allSitesTag2
15465075 15465075 60235 13882836 349 1426259 96 5000000 2 3 59882 0 0
*** Résultat de la segmentation du fichier: allSitesTag2.TXT ***
Délimiteurs .,:;!?/_\"'()[]{}ß$
    nombre des occurrences :    13882836
    nombre des formes      :          349
    frequence maximale     :      1426259
    nombre des hapax       :           96
    nombre des clés(type)  :            3
    nombre des clés(ctnu)  :        59882
*** Fin de la segmentation du fichier:      allSitesTag2.TXT ***
```

	Partie	Occurenc	Formes	Hapax	Fmax	Forme
✓	ppete99	2250398	215	37	197792	FONT
✓	ppmars00	7578785	287	65	789008	TD
✓	smaou00	1019390	163	23	89611	BR
✓	smsep99	3034263	147	13	365494	TD

Applet *vs* Script



Qualification des propriétés des pages : balise BODY



Le trait BODY seul est un trait plus fréquent sur les sites perso. Plus ce trait est pourvu d'attributs, plus sa répartition passe du côté des sites SM.

Perspectives : de TypWeb à SensNet

- Identification d'autres jeux de traits pertinents pour discriminer des catégories de sites
- Utilisation d'autres ressources pour décrire les pages : les descriptifs des annuaires, les commentaires des internautes sur les sites des autres…
- Amélioration des méthodes de catégorisation : combiner approche inductive et classification supervisée

SensNet

« *Catégorisation sémantique des usages et des parcours sur le web* », labellisé par le RNRT

`http://www.telecom.gouv.fr/rnrt/projets/res_01_39.htm`

- Objectif : Analyser les usages d'Internet en qualifiant finement le type de service proposé, en intégrant la dimension formelle de la page (hypermédia), en situant la page vue dans un parcours, en qualifiant les contenus
- Partenaires : France Télécom R&D, NetValue, LIMSI-CNRS et Paris III