Chapitre 3

De l'URL au contenu

L'analyse des parcours Web passe nécessairement par une première étape de description des contenus visités. Si l'on peut souhaiter disposer d'une description fine au niveau des différents éléments qui composent chaque page afin de les agréger dans des descriptions plus larges au niveau de la session ou de l'utilisateur, le postulat de la primauté de la tâche sur les contenus visités nous oblige à relativiser cette approche compositionnelle. En particulier, il n'est pas certain que la description seule des pages permette une description des sessions : il est fort possible qu'aux niveaux méso et macro-analytique se jouent des phénomènes qui inscrivent les contenus visités dans des dynamiques qui en modifient profondément le sens. Il importera, dans ce cadre, d'évaluer la pertinence des descriptions disponibles selon le niveau d'analyse auquel on se place et selon la granularité du résultat que l'on souhaite obtenir. De ce fait, les enjeux de la caractérisation des contenus au niveau de la page répondent à l'objectif principal d'évaluer les différentes méthodes qui permettent d'identifier et de qualifier ces contenus, problème loin d'être évident en lui-même. Il appartiendra aux autres paliers d'analyse d'une sémantique des parcours d'apprécier l'utilisation qui peut être faite de ces descriptions aux niveaux *méso* et *macro*. Nous traiterons donc dans cette section les différentes techniques que nous avons envisagées pour qualifier les pages visitées et les problèmes qu'elles posent.

3.1 Les URL, porteuses d'informations

Dans les données de trafic de base dont nous disposons, les URL sont en elles-mêmes porteuses d'informations au niveau *micro* : type de protocole utilisé, contenu dynamique et noms de fichiers sont autant de renseignements qui, pour minimaux qu'ils soient, peuvent être pris en compte pour une description élémentaire du contenu ou, *a minima*, du type de contenu des pages visitées. Nous évaluons cette approche minimaliste en nous appuyant sur les données de trafic du panel SensNet en 2002, le plus représentatif et le plus volumineux avec 3 398 internautes observés pendant dix mois (voir chapitre 5.1, « Description des panels » pour une vue plus détaillée des données et des panels).

3.1.1 Des informations techniques aux indices d'usages

Nous l'avons vu, une URL est l'assemblage, suivant une syntaxe particulière, de plusieurs éléments : protocole, nom de domaine ou adresse IP, chemin vers la ressource, fichier demandé et, éventuellement, paramètres passés à la requête (méthode GET). Nous cherchons à voir ici si ces informations ne sont pas en elles-mêmes exploitables et ne fournissent pas des indices valorisables pour l'analyse d'usages.

Protocoles

Le protocole HTTP tend à s'imposer comme protocole standard, et à être le support de tâches et de modes d'interaction jusqu'alors réservés à FTP, POP/SMTP, ICQ, IRC, etc. : on trouve ainsi du WebMail, du WebChat, du téléchargement de fichiers à partir de serveurs Web. En conséquence, HTTP ne peut être un indicateur de contenu fiable au contraire, sinon dans sa version sécurisée, HTTPS, dont l'utilisation par les serveurs montre la nécessité de crypter les données échangées. L'utilisation de HTTPS est souvent associée à des transactions d'ordre financier, où la confidentialité des données est rigoureusement indispensable : achat en ligne (courses, voyage, tout ce pour quoi le numéro de carte bancaire sert à la transaction), services financiers (consultation de compte en banque, bourse en ligne), WebMail pour certains serveurs, ou plus généralement services personnalisés (auprès de son fournisseur d'accès, de prestataires de services, etc.). Dans tous les cas, il est question de sécuriser un échange d'information où l'identification personnelle de l'utilisateur est capitale ; ainsi, c'est surtout la « personnalisation » que l'utilisation de HTTPS dénote. FTP est plus clair à interpréter : il est question de télécharger des fichiers pour en « faire quelque chose » par la suite, et non de les visualiser et d'interagir avec leur contenu comme dans le cas de HTTP. Avec FTP, on est clairement dans une logique de récupération de ressources dont l'usage n'est pas immédiat, pour des volumes souvent bien supérieurs à ceux échangés par HTTP. On pourrait résumer cela en disant qu'avec HTTP, on est plutôt dans le « à consommer sur place » tandis que FTP nous met du côté du « à emporter ».

Dans nos données, nous ne disposons de traces sur le FTP qu'en 2000, la sonde NetMeter de NetValue ayant cessé de recueillir ces informations par la suite ; cela étant, à cette époque, pour un panel représentatif de 1140 individus, le FTP n'était présent que dans 400 sessions sur près de 130 000. On peut raisonnablement penser que, même si cette présence augmente en 2002, elle reste faible et son absence dans les données n'est pas gênante.

Dans les données SensNet 2002, HTTP est nettement majoritaire, en nombre d'URL vues comme en nombre d'URL distinctes (voir Tableau 3.1). Nous notons également la présence non négligeable du protocole AOL, protocole propriétaire réservé aux abonnés de ce fournisseur d'accès, et dont le contenu s'apparente à du contenu Web.

Tableau 3.1. Protocoles utilisés par le panel SensNet de janvier à octobre 2002

Protocole	Nombre d'URL distinctes	Nombre d'URL vues	Présence dans les sessions
AOL	2,3 %	8,9 %	17,7 %
HTTP	94,7 %	88,2 %	95,2 %
HTTPS	3,0 %	2,9 %	16,0 %

Comment faut-il interpréter ces éléments ? L'absence de correspondance directe entre protocole et contenu rapatrié, due en particulier à la disparité des services personnalisés accessibles par HTTPS, interdit d'exploiter ces données seules. Toutefois, elles pourront être mobilisées en renfort d'autres traitements, comme indice d'une action particulière. Par exemple, sur un type de site d'achat de billets d'avion comme Opodo, l'observation d'une séquence amenant un passage par le HTTPS peut être un indice d'engagement vers un acte de réservation ou d'achat. Dans le cadre de recherche de logiciel sur un site comme `www.telecharger.com`, l'usage du FTP peut, de manière similaire, attester le téléchargement d'un logiciel. C'est donc à une échelle plus fine d'analyse que l'on peut mobiliser l'information relative au protocole utilisé dans la navigation Web, celle-ci étant trop générale hors de tout contexte.

Domaines

Le nom de domaine fournit également des informations sur les contenus visités : le rattachement à un domaine de premier niveau (*Top Level Domain*, ou TLD : `.com`, `.org`, `.fr`, etc.) est, dans certains cas, un indice du type de site et de la langue des documents visités. Nous renvoyons à la lecture, dans l'Annexe 2, du chapitre présentant les principes d'organisation en domaines et sous-domaines. Nous retiendrons ici que, pour lier domaine et contenus, il faut distinguer les deux grandes familles de domaines de premier niveau, les TLD génériques (*Generic TLD*, ou gTLD), et les TLD nationaux (*Country Code TLD* ou ccTLD). Pour les premiers, le domaine correspond en principe à un regroupement thématique et fonctionnel :

- *org* : organisations à but non lucratif
- *edu* : organismes éducatifs américains
- *mil* : organismes militaires américains
- *com* : organismes à but lucratif
- *net* : organismes chargés de l'administration du réseau
- *gov* : organismes gouvernementaux américains
- *int* : organismes internationaux

De nouveaux TLD génériques sont apparus en 2001 et 2002 :

- *biz* : destiné au *Business*
- *info* : usage illimité
- *name* : pour les particuliers
- *pro* : comptables, juristes, médecins, et autres professionnels
- *aero* : industrie des transports aériens
- *coop* : pour les Coopératives
- *museum* : musées

Les conditions d'accès à ces TLD sont variables, ce qui vient parasiter la correspondance entre TLD et type de contenu. Les TLD apparus en 2000 et 2001 sont encore très peu répandus, et au sein des autres TLD, les `.com`, `.net`, `.org`, `.name`, `.biz` et `.info` sont dans les faits accessibles à tout un chacun. Impossible, dans ces conditions, d'exploiter ces deux domaines, le contenu des pages étant complètement indéterminé, tant dans la nature que dans la langue des sites.

Les ccTLD sont plus exploitables : en premier lieu, ils renseignent de manière relativement fiable sur la langue générale des sites. Si rien n'empêche un site en `.fr` de publier des pages dans des langues autres que le français, ce site reste globalement rattaché à l'univers francophone[1]. Par contre, les conditions d'accès aux ccTLD sont gérées individuellement par chaque pays, et pour la France, l'accès à une adresse en `.fr` est peu aisé, ce qui pousse bon nombre de webmestres à investir dans le « dot com ». En outre, l'information de contenu est quasi-nulle : si le ccTLD contient certains sous-domaines réservés, comme le `.asso.fr` pour les associations, ou le `.st.fr` pour les sociétés, ces conventions sont peu utilisées et permettent de décrire peu de sites.

L'examen des TLD et ccTLD réservés accédés dans les données vient confirmer ces éléments. Dans les données SensNet 2002, les adresses en `.com` et en `.fr` représentent 83 % des URL distinctes et 79 % des URL vues (voir Tableau 3.2). L'évolution sur trois ans auprès des trois panels mobilisés dans les projets TypWeb et SensNet montre par ailleurs une certaine stabilité de cette situation (voir Tableau 3.3).

Tableau 3.2. TLD et ccTLD réservés dans les données SensNet 2002

Domaine	% des URL distinctes	% des URL vues
`com`	52,8 %	44,0 %
`fr`	30,6 %	35,2 %
`net`	6,0 %	3,9 %
`org`	1,3 %	1,0 %
`tm.fr`	0,3 %	0,9 %
Adresse IP	3,0 %	3,4 %
`gouv.fr`	0,3 %	0,4 %
`de`	0,4 %	0,2 %
`asso.fr`	0,2 %	0,2 %
`be`	0,3 %	0,2 %
`cc`	0,1 %	0,1 %
`ch`	0,2 %	0,1 %
`it`	0,1 %	0,1 %
Autres	4,4 %	10,3 %

[1] Quelques exceptions existent : d'une part, les conditions d'accès à chaque ccTLD sont définies par chaque pays, et certains peuvent être choisis par des webmestres étrangers pour leur prix ou leur facilité d'accès. D'autre part, pour certains ccTLD de petits pays, l'extension correspondante a une signification dans d'autres langues et peut être rattachée à une détermination thématique : par exemple, les Iles Tuvalu ont une extension en `.tv`, ce qui a amené des chaînes de télévision à acheter des noms de domaine sur ce ccTLD (par exemple : la chaîne française « Cuisine TV » est accessible à l'adresse `www.cuisine.tv`).

Tableau 3.3. Évolution des TLD et ccTLD réservés dans les données SensNet

Domaine	2000	2001	2002
com	47,5 %	45,0 %	44,0 %
fr	36,2 %	38,6 %	36,9 %
net	4,2 %	2,8 %	3,5 %
org	1,7 %	1,2 %	1,1 %
tm.fr	0,4 %	1,8 %	1,0 %
Adresse IP	2,6 %	1,2 %	0,7 %
gouv.fr	0,6 %	0,6 %	0,4 %
de	0,4 %	0,3 %	0,3 %
asso.fr	0,4 %	0,2 %	0,2 %
be	0,2 %	0,1 %	0,2 %
ch	0,2 %	0,1 %	0,1 %
it	0,1 %	0,1 %	0,1 %
Autres	5,3 %	7,9 %	11,5 %

Ces informations ne sont pas inintéressantes en elles-mêmes, mais renseignent bien plutôt sur la production des contenus Web : la gestion des noms de domaines de premier niveau, leur structuration et leur organisation sont l'objet d'enjeux économiques et stratégiques, et l'importance du `.com` montre la prévalence d'un TLD « fourre-tout » où les webmestres vont préférentiellement inscrire leur nom de domaine. Pour l'analyse des usages, nous pouvons tirer bien peu de conclusions de ces éléments, les utilisateurs ne choisissant pas d'aller sur tel ou tel TLD mais sur des sites en fonction de contenus qui les intéressent. De ce point de vue, un `.com` trop large et un `.fr` hétérogène ne constituent pas des indices exploitables pour la qualification des contenus visités par les internautes ; on tentera tout au plus de voir de manière différentielle entre plusieurs groupes d'internautes comment l'accès à certains domaines minoritaires mais discriminants, comme le `.edu` ou le `.gouv.fr`, peut être un signe de centres d'intérêt particuliers.

Types de fichiers, types de contenus ?

Les types de fichiers peuvent fournir des indications sur le contenu des documents : une image ne se « lit » pas de la même manière qu'un fichier PDF, le HTML permet des interactions que ne permet pas le format MS Word. Les types de fichiers permettent également de savoir si les contenus sont dynamiques ou non, en examinant si l'URL renvoie vers un script ou vers un format statique. Pour utiliser cette information, nous avons créé une grille d'analyse associant les extensions de fichiers, qui permettent d'identifier leur type lorsque cette extension existe, et les types généraux de fichiers et de contenus associés :

Type principal	Sous-type	Extensions
Document	HTML	htm, html, dhtml, xhtml, etc.
	texte	txt, dat
	PDF	pdf
	Post Script	ps
	Word	doc, rtf
	Excel	xls, csv
	XML	xml
Multimédia	audio	wav, ram, mp3, m3u, etc.
	audio/vidéo	rm, mpg, mpeg, avi, mov, etc.
	image	gif, jpeg, jpg, bmp, png, etc.
Script	-	asp, php, pl, cgi, etc.
Archive	-	zip, rar, etc.
Outil	-	exe, jar, rpm, ico, etc.
Autre	-	Copernic, ini, css, etc.

En projetant cette grille sur les URL visitées dans les données SensNet 2002, nous trouvons que sur 6,7 millions d'URL distinctes (représentant plus de 27,2 millions d'URL vues), près de 5,9 millions ont un fichier spécifié. Sur ces URL, nous avons extrait l'extension de fichier et examiné si celle-ci correspond à un type référencé dans la grille ci-dessus[1].

Tableau 3.4. Types de fichiers pour les URL pointant vers un fichier avec extension

Type	% des URL distinctes	% des URL vues
script	42,41 %	44,86 %
document	34,37 %	39,55 %
Pas d'extension	21,07 %	10,38 %
Non classé [2]	1,68 %	4,29 %
multimédia	0,23 %	0,44 %
autres	0,18 %	0,41 %
outil	0,03 %	0,07 %
archive	0,01 %	0,01 %

Au terme de cette analyse, hormis les 20 % de fichiers sans extension, la répartition des catégories de contenu montre une prédominance forte des types « script » et « document » (voir Tableau 3.4). À la catégorie « script », qui représente

[1] Rappelons que les sondes utilisées pour recueillir les données de trafic n'enregistrent pas les requêtes pointant vers des fichiers de type image (extensions 'jpg', 'gif', etc.).

[2] L'extraction d'extension est faite sur la base d'une expression régulière (l'expression `/\.([^.])$/`). Cette méthode renvoie toutes sortes de chaînes de caractères, y compris des extensions qui n'en sont pas réellement mais font partie d'un nom du fichier comprenant un point (ex : le fichier `browser_menu.lasso` dans l'URL : `http://www.geneaguide.com/a-store/browser_menu.lasso?st=&lng=&cat=3&act=rub&or=GGIX`). En conséquence, nous avons créé une catégorie « Non classé » pour distinguer ces extensions suspectes qui ne renvoient à aucun type de fichier et s'apparentent à des fichiers sans extension.

près de 42 % des URL vues contenant une extension, il faut sans doute ajouter les fichiers sans extensions, qui correspondent très probablement à des scripts également ; au total, sur l'ensemble des URL vues, ce sont ainsi près de 63 % des requêtes aboutissant vers des script côté serveur (21 % de fichiers sans extension, 42 % de type « script »).

Ceci montre l'importance des contenus dynamiques sur le Web : interrogation de bases de données, requêtes sur des moteurs de recherche, examen d'espaces personnalisés sont autant de requêtes qui engagent une interaction avec l'utilisateur, et la production de contenus en fonction de sa requête. On notera que la part des contenus dynamiques est en augmentation par rapport à l'année 2000 : sur les URL vues par le panel NetValue cette année, les contenus dynamiques représentaient environ 52 % du total des URL visitées (21 % de fichiers sans extension, 31 % rattachés au type « script »), contre 63 % en 2002.

Pour autant, nous ne savons pas le type de contenu renvoyé par ces scripts, et rien ne permet de le déterminer sur la base des données de trafic[1]. Si l'on postule que les scripts renvoient globalement les mêmes types de fichiers que les requêtes vers des fichiers statiques, le HTML est alors le format standard majoritaire. En effet, hormis les scripts, le type « document » est largement majoritaire dans les URL demandées, les fichiers d'archives et multimédia restant négligeables ; au sein du type « document », le HTML est présent dans 98 % des cas, devançant de loin tous les autres formats (voir Tableau 3.5).

Tableau 3.5. Audience des types de documents en 2002

Type de document	% des URL distinctes	% des URL vues
HTML	98,17 %	97,26 %
texte	1,10 %	2,19 %
Word	0,41 %	0,11 %
XML	0,30 %	0,43 %
PDF	0,02 %	0,01 %
Excel	0,00 %	0,00 %
Post script	0,00 %	0,00 %

On constate ainsi que le format HTML constitue le support majoritaire de la communication sur le Web, et ce d'autant plus que nous pouvons supposer que les résultats de l'exécution de scripts côté serveur sont très majoritairement dans ce format, auxquels il faut très certainement ajouter les requêtes ne pointant vers aucun fichier, les serveurs les redirigeant majoritairement vers un fichier `index.html`. Nous pouvons ainsi estimer que près de 95 % des fichiers récupérés par les internautes sont au format HTML, ceux-ci pouvant bien entendu inclure des éléments non textuels.

[1] Il faudrait pour cela que les sondes de recueil de trafic extraient, dans les en-têtes HTTP renvoyées par les serveurs, le champ « Content-type » ou, mieux, examinent les en-têtes de fichiers eux-mêmes, le « Content-type » HTTP étant souvent peu fiable.

Ici encore, comme pour l'étude des domaines de premier niveau visités, ces éléments intéressent plus l'analyse de la production que celle des usages : l'accès au non-HTML est intéressant à noter en termes d'usages, mais la part écrasante du HTML rend cette information si rare qu'elle peut à peine être exploitée. Par ailleurs, le format HTML est en quelque sorte l'arbre qui cache la forêt, car il peut contenir toutes sortes d'éléments : audio, vidéo, animations, etc. Dans ces conditions, l'exploitation du type de fichier pour la caractérisation des contenus visités ne pourra être faite que ponctuellement et avec parcimonie, pour repérer des phénomènes bien précis.

> ***Synthèse.*** *Les informations sur les protocoles et les types de fichiers accédés sont trop pauvres pour être exploitées efficacement comme descripteurs de contenu des pages visitées.*

3.1.2 Noms de répertoires

Les indices d'ordre techniques fournis par les URL sur les modes d'accès aux documents et aux contenus se révèlent en définitive assez peu productifs, mais l'exploitation de l'URL ne s'arrête pas là. Parallèlement, nous avons tenté de déployer une approche plus linguistique utilisant les noms de répertoires comme indications de contenus.

Principe et hypothèses

Nous avons constaté au fil de l'examen des URL que leur simple lecture nous permettait bien souvent de déduire le contenu qu'elles recouvrent. Quelques exemples extraits des données illustrent ce propos :

- sur Yahoo, l'ensemble des différents services du portail est organisé en sous-domaines de yahoo.com, et préfixé par service et par pays. Ainsi, `http://fr.finance.yahoo.com/` regroupe l'ensemble des pages de Yahoo France traitant de la bourse, `http://fr.news.yahoo.com` les pages d'actualité, `http://fr.games.yahoo.com/` les jeux ;
- les recherches dans le catalogue de l'université de Strasbourg se font à l'aide d'un script situé dans un répertoire nommé « catalogue » : `http://www-bnus.u-strasbg.fr/catalogue/cgi-bin/boutons.asp` ;
- sur les sites de type annuaires présentant des liste de liens classés, la structure logique en catégories et sous-catégories se retrouve souvent dans la structure des répertoires. Ainsi, on trouve sur l'annuaire du Web Nomade des adresses de la forme : `http://www.nomade.fr/cat/mes_courses/artisans_profession/artisanat_art/travail_des_textiles/` ; sur `www.ressources-web.com`, les sites dédiés au recrutement se trouve sous `http://www.ressources-web.com/RH/emploi/recrutement/`.

Sur cette base empirique, nous avons voulu quantifier la présence de mots de la langue dans les noms employés pour nommer les répertoires. Nous écartons les noms de domaines, qui correspondent la plupart du temps à des noms de marques,

ainsi que les noms de fichiers qui répondent à des normes et des impératifs qui les rendent peu productifs, comme nous avons pu le constater manuellement. L'analyse porte donc sur les noms de répertoires, et vise à évaluer la présence de graphies correspondant à des mots anglais ou français, sous forme canonique ou fléchie. Nous voulons tester ici l'hypothèse selon laquelle le nommage, hormis certains cas où des impératifs techniques ou conventionnels prévalent, correspond à une désignation des contenus, et ce par l'emploi de mots ou de composition de mots de la langue. Cette recherche ne présage pas de l'exploitation éventuelle de ces résultats (utilisation de thésaurus, de lexiques par domaines, etc.) : il s'agit d'une première étape d'évaluation de la description de contenus par les noms de répertoires, avant d'envisager d'aller plus loin.

Pour vérifier cette hypothèse, nous avons extrait dans le *chemin éditorial*[1] des URL visitées les noms des répertoires utilisés, et examiné si ceux-ci correspondent à des graphies répertoriées dans des dictionnaires de formes françaises et anglaises. Pour cela, nous avons utilisé le dictionnaire de l'ABU pour le français[2], qui contient 290 000 formes fléchies, et un dictionnaire anglais qui propose 111 000 formes fléchies. Une première étape a consisté à extraire les noms de répertoires ; ensuite, ces noms ont été normalisés, c'est-à-dire que les codages Unicode des caractères non supportés par HTTP ont été transcrits en iso-latin-1. Nous avons minusculisé les noms de répertoire, et ainsi obtenu 676 614 noms uniques de répertoires, se retrouvant au total dans 5,2 millions d'URL (représentant 22,4 millions de pages vues). Ensuite, nous avons dressé une liste des noms de répertoires « techniques », c'est-à-dire ceux dont le nom, du fait des conventions et valeurs par défaut des serveurs, est fixé à l'avance.

Tableau 3.6. Répertoires techniques et pages visitées en 2002

Nom	Nb. URL distinctes	Nb. URL vues
asp	1,3 %	1,7 %
bin	7,9 %	22,3 %
cgi	0,8 %	0,5 %
cgi-bin *et dérivés*	57,8 %	30,1 %
exec	2,0 %	1,0 %
html	7,6 %	8,7 %
include	0,7 %	5,2 %
jsp	2,8 %	2,1 %
local	0,1 %	0,0 %
perl	1,7 %	2,6 %
php *et dérivés*	2,8 %	1,8 %
pub	1,5 %	2,5 %
scripts *et dérivés*	9,0 %	5,3 %
servlet / servlets	4,1 %	16,3 %
Total	100 %	100 %

[1] Correspond au chemin après l'identification des *sites éditoriaux* ; voir 2.2.2, « Traitement des URL » p. 57 pour une description détaillée de cette opération.

[2] ABU : Association des Bibliophiles Universels ; voir http://abu.cnam.fr/DICO/mots-communs.html.

Au total, nous avons identifié manuellement une trentaine de noms de répertoires correspondant à ces critères, présents dans 25 % des requêtes retenues, soit 27,4 % des pages vues retenues. La présence des répertoires techniques dans les URL visitées en 2002 confirme qu'il s'agit bien de scripts, les noms 'bin', 'cgi-bin', et 'servlet' arrivant en tête (voir Tableau 3.6 ci-dessus).

Des résultats décevants

Nous avons ensuite confronté la liste des noms de répertoire, expurgée de cette liste de noms techniques, aux dictionnaires français et anglais dont nous disposions. Nous avons calculé le nombre d'URL distinctes comportant le nom extrait, ainsi que le nombre de pages vues correspondant, sachant qu'une adresse peut comporter plusieurs noms (répertoires et sous-répertoires) ; ce calcul porte sur l'ensemble des URL vues en 2002 par le panel SensNet (Tableau 3.7).

Tableau 3.7. Couverture des noms de répertoires en 2002

	Nombre de noms uniques	Nb. URL distinctes	Nb URL vues
Total sans les répertoires techniques	0,6 Mls – 100 %	4,1 Mls – 100 %	17,4 Mls (100 %)
Présent dans le dictionnaire français	1,1 %	30,8 %	30,4 %
Présent dans le dictionnaire anglais	1,9 %	40,3 %	37,3 %
Présent dans les deux dictionnaires	0,5 %	16,8 %	16,9 %

À la lecture de ces résultats, nous constatons que les noms des répertoires sont globalement étrangers à la langue : seulement 1,5 % de ces noms correspondent à des mots de la langue anglaise ou française. De manière plus surprenante, alors que la visite de domaines français est majoritaire dans les pages visitées, l'anglais est plus présent que le français. Si les taux de couverture avec les URL sont malgré cela assez importants (entre 30 et 40 %), la faible diversité des lexies invite à la prudence, de même que le recouvrement important entre listes de mots anglais et français, qui rend difficile l'exploitation des mots extraits.

Cette approche s'avère en définitive peu productive, au même titre que les autres tentatives d'attacher des éléments de contenu aux URL sur la base d'indices techniques. Ce constat met un terme à l'ambition initiale d'une qualification, même à gros grain, des contenus sur la simple base de l'URL ; il souligne de manière évidente la difficulté à qualifier les contenus et la nécessité de recourir à des ressources externes.

> ***Synthèse.*** *Les noms données aux répertoires dans les URL ne correspondent pas assez à des mots de la langue pour être utilisés comme descripteurs du contenu des pages.*

3.1.3 Catégorisation semi-automatique avec *CatService*

L'application *CatService* constitue une voie alternative d'exploitation des URL en y attachant des informations de contenu externes définies par les utilisateurs de l'application. Développée dans le cadre des projets TypWeb et SensNet, *CatService* permet d'attacher à une URL, sur la base d'expressions régulières, des catégories sur

une échelle d'analyse à cinq niveaux. Il s'agit là d'une première voie d'enrichissement complexe des données de trafic qui reste encore proche des données brutes de trafic, et produit des descriptions entièrement paramétrables et facilement exploitables.

Fonctionnement

Le module *CatService* a pour objectif, dans la plateforme SensNet, de qualifier les URL visitées en termes de types de sites et de services. La qualification se fait à cinq niveaux :

- *type de site* : définit le type de site ou de contenus accessibles sur le site, par exemple « portail généraliste », « site de WebMail », « bibliothèque électronique », etc. Dans le système, un site peut être rattaché à plusieurs *types de sites*, ce qui est cohérent avec l'offre de contenu sur le Web : Yahoo est ainsi un portail généraliste, mais également un moteur de recherche et un portail de WebMail, tandis que Google n'est qu'un moteur de recherche.
- *portail* : le site ou le portail auquel l'URL renvoie. Le système permet ainsi de regrouper sous une seule entité les portails répartis sur plusieurs noms de domaines.
- *fournisseur* : le fournisseur de service éventuellement appelé par le portail : par exemple, le portail Free fait appel à Google pour son service de recherche sur le Web.
- *service* : au sein d'un *type de site* donné, une grille de services proposés est définie et appliquée à l'ensemble des sites concernés. Ce fonctionnement permet de créer des catégories comparables au sein d'une analyse portant sur un type de site particulier, et de dépasser les rubriquages définis par chaque site.
- *sous-service* : la catégorie *service* peut être précisée en sous-catégories. Par exemple : dans le service « moteur de recherche », on distingue la recherche de pages Web, d'images et de contributions à des forums, et l'accès à une page de recherche avancée.

Pour fonctionner, *CatService* a besoin d'un ensemble de ressources qui sont stockées dans des tables d'un SGBD relationnel :

1. la table contenant les URL à catégoriser ;
2. le référentiel à cinq niveaux ;
3. les règles de *pattern matching*, construites à l'aide du formalisme des expressions régulières. Ces règles permettent d'associer à une classe d'URL (décrites à l'aide d'expressions régulières) un couple portail-fournisseur, un service et un sous-service donnés.

Le rattachement d'une URL à un service se fait sur la base d'expressions régulières construites manuellement après examen des différentes adresses relatives à un portail et vérification du contenu des pages vers lesquelles elles pointent. Les expressions régulières portent distinctement sur le nom de domaine et la suite de l'adresse, et peuvent être enrichies d'une expression régulière « négative » qui exclut du résultat les URL la vérifiant. En outre, deux traitements spécifiques sont opérés sur certains types de services :

- pour les URL correspondant à des requêtes auprès des moteurs de recherche, une procédure extrait et normalise les mots-clefs de la requête, et repère également la navigation dans les pages de résultat suivantes ;
- pour les URL accédant à des services de WebMail, l'outil repère, lorsque cela est possible, les actions de login, de lecture et d'écriture des messages.

Trois exemples de règles illustreront bien ce mécanisme :

- *Règle pour un moteur de recherche*

RegExpHost	`^(www\.)?google\.(com|fr|be|ch|de|co\.jp|it)$`
RegExpReste	`(search|custo|advanced_search)`
MotClef	`(&|\?)q=`
Navigation	`start=`
Portail	`Google`
Fournisseur	`Google`
Service	`Moteur`
Sous-service	`Web`

- *Règle de WebMail*

RegExpHost	`(u|w{3})[0-9\-]*\.caramail\.lycos\.(fr|com).*(Compose|[Aa]fficheBody|ActionMail|cgi-bin/contenu\?FOLDER=)`
Écriture	`Compose`
Lecture	`[Aa]fficheBody`
Login	*NA*
Portail	`Caramail`
Fournisseur	`Caramail`
Service	`Communication`
Sous-service	`Mail`

- *Règle générale, sur un portail de e-commerce*

RegExpHost	`www\.fnac\.(fr|com)`
RegExpReste	`^/default\.asp`
Négatif	`(NID=%2D[1-4]&|Account)`
Portail	`Fnac`
Fournisseur	`Fnac`
Service	`Page Accueil`
Sous-service	`Accueil`

L'ensemble du référentiel et des règles est construit manuellement par les utilisateurs de l'application. Ce travail nécessite un investissement important en temps, mais *CatService* permet alors de qualifier avec précision la part de chaque service utilisé et de dépasser ainsi la simple mesure d'audience. En particulier, la création d'un référentiel de services au sein d'un type de portails donné ouvre la voie de la comparaison des usages entre différents portails ; cette homogénéisation a ainsi permis de comparer l'usage des différents services sur les portails généralistes en 2000 dans [Beaudouin *et al.* 2002].

Référentiel utilisé

En juillet 2003, date à laquelle nous avons exploité cette base pour nos données de trafic, plus de 1 800 règles avaient été créées au sein de *CatService*[1]. Elles identifient quinze types de contenus et de sites sur plus de 230 portails identifiés, dont le Tableau 3.8 donne la liste complète. Comme le référentiel permet d'attribuer plusieurs types de portail ou de contenus à un site donné, certains sites apparaissent plusieurs fois, par exemple le site de La Poste qui propose à la fois des services bancaires (type « e-commerce / Banque – Bourse ») et un service de messagerie (type « WebMail »). C'est au niveau des services et sous-services que l'on repère par la suite dans les données le type de contenu visité par l'internaute sur le portail considéré.

Tableau 3.8. Types de portails et portails référencés dans CatService (juillet 2003)

Type de portail	Nb. portails	Portails répertoriés
Bibliothèque électronique	24	ABU, Alex Catalogue, American Memory, Arob@se, Athena, Berkeley DL, Bibelec, Bibliopolis, Bibliothèque de Lisieux, BN Canada - Numérique, BNF, Gallica, ClicNet, CNUM, Electronic Text Center, eLibrary, Gutenberg project, INALF, LiNuM, Mozambook, NZ Digital Library, Online Books Page, Revues.org, UMDL
e-commerce / banque, bourse	16	BanqueDirecte, BanquePopulaire, BNP, Boursorama, BRED, CaisseEpargne, CIC, Crédit Agricole, Crédit Lyonnais, Crédit Mutuel, Direct Finance, Fimatex, La Poste, Selftrade, Smcaps, Société Générale
e-commerce / biens culturels	17	Alapage, Amazon, Barnes & Noble, Chapitre.com, CNRS Editions, Cylibris, Edibook, Eyrolles, Fnac, Galaxidion, Imprimermonlivre.com, Les Introuvables, Librissimo, LibrisZone, Litraweb, Livre-rare-book, Numilog
e-commerce / courses	4	c-mescourses, Houra, ooshop, Telemarket
e-commerce / tourisme	7	AirFrance, Degriftour, Ebookers, NF, Promovac, SNCF, Travelprice
Forum	45	2037.biz, 24rollers, Aceboard.net, Adobe, AdultForums, Afterdawn, AideOnLine, Air-radiohead, AOL, Atari.org, AtomicForum, Aufeminin, AutoJournal, Boursorama, Chez, Clubic, DynDns, EnseignantsDuPrimaire, EuropeanServer, Fimatex, Forum 2CV, Hardware.fr, HitParade, HomepageTools, i! France, JeuxVideo.com, JudgeHype, Lagardere Interactive, LesForums.com, Libertysurf, Loftstory, Lycos, M6, MadStef, Ondelette.com, Presence PC, QuickWeb, Respublica, Smcaps, Telecharger.com, Voila (fr), VVLR.com,

[1] Cette base résulte de la contribution de l'ensemble des personnes engagées dans les projets TypWeb et SensNet ; seul ce travail collectif a permis de constituer un jeu de règles précis et large tel que celui que nous employons, et nous remercions à cette occasion tous les contributeurs à ce travail.

		Wanadoo (fr), Wordox, Yahoo (fr)
Généalogie	17	123genealogie, AFG, Ancestry.com, Ancetres.com, CGFA, FamilySearch, GeFrance, GeneaLand, Genealogie-standard, Genealogy.tm.fr, GenealoJ, GeneaNet, GénéaStar, GenLink, Histoire-Généalogie, Ma-Genealogie, Notre Famille
Média / presse	12	AutoJournal, L'Express, Le Figaro, Le Monde, Le Parisien, Le Point, Les Echos, Libération, New York Times, Nouvel Obs, Paris Match, Telerama
Media / Radio	9	Chérie FM, Contact FM, Europe 1, Fun Radio, NRJ, Radio France, RFI, RTL, Skyrock
Media / TV	7	France Télévision, France2, France3, France5, Loftstory, M6, TF1
Moteur	32	AllTheWeb, Altavista.com, Altavista.fr, BlueWindow, Carrefour.net, Club-internet, Ctrouve, Dmoz, Ecila, Euroseek, Excite, Excite (fr), Francité, Free, Google, Goto, Grolier, Kartoo, Lokace, Looksmart, Lycos, Metacrawler, MSN, Netscape, Nomade, NorthernLight, Toile, Voila (fr), Wanadoo (fr), Webcrawler, Yahoo (com), Yahoo (fr)
Portail Généraliste	13	Altavista.com, Altavista.fr, Club-internet, Free, Libertysurf, Lycos, MSN, Noos, Tiscali, Voila (fr), Wanadoo (fr), Yahoo (com), Yahoo (fr)
Portail Pages Perso	44	Altavista.com, Altern, Angelfire, AOL, Aufeminin, Bluewin, Chez, CiteWeb, Claranet, Claranet (fr), Club-internet, Cybercable, Forez, Fortunecity, Free, Freesurf, i! Belgique, i! France, i! Québec, i! Suisse, Icq, Infonie, Le Village, LeVillage, Libertysurf, Lycos, Mageos, Multimania, Noos, Nordnet, Pagesweb, Pandora, Populis, Respublica, Skynet, Swing, Tripod (com), Tripod (fr), VirtualAvenue, Voila (fr), Wanadoo (fr), Wanadoo Pro, Worldonline, Yahoo (com)
WebChat	31	asterochat, Boulimie, Canalchat, Caramail, Club-internet, Free, Fun Radio, GOA, hiwit, LeVillage, Libération, Libertysurf, Lycos, Meetic, MSN, nokiagame, Nomade, Notre Famille, NRJ, onconux, Prizee, radioespace, Respublica, Skyrock, tchatche, TF1, Voila (fr), Wanadoo (fr), Worldonline, Yahoo (com), Yahoo (fr)
WebMail	36	AOL, Aufeminin, Bigmailbox, Boursorama, Caramail, Club-internet, Compuserve, Excite (com), Excite (fr), Fnac, Francemail, Free, Freesbee, Freesurf, i! France, La Poste, Lemailparisien, Libertysurf, Lycos, Mageos, MSN, Multimania, Netclic, Netcourrier, Netscape, Nomade, Noos, Oreka, Populis, Respublica, Voila (fr), Wanadoo (fr), Worldonline, Yahoo (com), Yahoo (fr)

Au sein de la catégorie « Portail généraliste », où l'offre de contenus est la plus diversifiée, 17 services distincts sont identifiés, chacun étant détaillé en un nombre de sous-services variable selon l'importance du service (voir Tableau 3.9).

Tableau 3.9. Services et sous-services référencés pour la catégorie « Portail généraliste »[1]

Service	Sous-services associés
Achat	Enchères, Logiciels, Offre d'Emploi, Petites-annonces, Téléchargement de sonneries, Vente en Ligne, Voyages
Annuaire	Annuaire, Local, Mail, Page Personnelle, Web, Webring
Bourse	Infos
Communication	Carte, Club, Débat, Forum, Groupe de Discussion, Invitation, Liste De Discussion, Mail, Messager, Minitel, Mobile, Rencontres, SMS, WebChat
Divers	Aide, Family Filter, Flash, Outils Web, Provider, Référencement, Traduction
Généralités	Aide, Contact, Jeux, Promo
Information Produit	Abonnement, Voyage
Information Service	Information Abonnés, Informations Pratiques, Présentation des Entreprises, Présentation des Services
Informations	Auto/Moto, Charme, Cinéma, Encyclopédie, Enseignement, Événement, Famille, Féminin, Finance, Horoscope, Informations, Junior, Loisir, Météo, Multimédia, Musique, Plan / Itinéraire, Pratique, Programme TV, Senior, Sport, Tourisme, Trafic, Voyage
Loisir En Ligne	Jeux, Serveur de Jeux
Moteur	Forum, FTP, Images, Web, Options
Page Accueil	Accueil
Page Perso	FTP, Hébergement, Outils, Profiles, Recherche, Référencement, Site Perso
Personnalisation	Affiliate, Agenda, Album Photos, Carnet d'Adresses, Compte Utilisateur, Fidélisation, My Yahoo, Personnalisation, Photos, Profiles, Suivi Consommation, Wanadoo et Moi
Aide	-
Loisir En Ligne	-
Non catégorisé	-

Intérêt et mobilisation de *CatService* pour l'analyse des parcours

La catégorisation des services a une grande valeur pour notre travail, en particulier pour les portails généralistes : ces sites drainent la majorité de l'audience sur le Web,

[1] Nous incluons ici également les services liés aux moteurs (« Moteur » - « Web ») et au WebMail (« Communication » - « Mail »), présents sur la plupart des treize portails généralistes identifiés.

et occupent une place incontournable dans les données de trafic. La description fine de *CatService* permet non seulement de distinguer, dans l'audience de chaque portail, les différents services utilisés (moteur, WebMail, etc.), mais aussi de rendre comparables ces éléments d'un portail à l'autre.

Elle introduit également une importante notion de services, et permet de faire une distinction entre les pages dont le contenu textuel prime et celles où leur fonction (le service proposé) prend le dessus d'un point de vue descriptif. À titre d'exemple, il semble plus pertinent pour l'analyse des pages de Yahoo, d'un point de vue utilisateur, de retenir que telle URL fournit de l'information en continu plutôt que d'examiner le contenu des informations fournies dans la page, contenu dynamique et en perpétuel changement. Il est alors possible d'opérer des traitements différenciés en termes de description entre pages « à lire » et pages de services et d'outils.

De ce point de vue, si *CatService* ne décrit pas l'ensemble du Web, le choix des sites manuellement catégorisés répond à la nécessité de décrire ceux qui sont les plus visités par les panels. On couvre, avec les catégories « portail généraliste », « moteur », « WebMail », « forum » et « e-commerce », les sites qui attirent le plus d'internautes et s'imposent comme des points de passage incontournables. La description des parcours en termes de services connaît ainsi une base solide et large, qui correspond aux principales activités sur le Web : information, communication, achat, services bancaires et divertissement.

Tableau 3.10. Couverture des données de trafic par CatService

	Nb URL distinctes	Nb URL vues
SensNet 2002	29,4 %	27,8 %
SensNet 00-02	27,5 %	29,4 %
BibUsages	30,5 %	32,3 %

En termes de couverture, la catégorisation avec *CatService* décrit entre 28 % et 32 % des URL distinctes selon les données de trafic, avec des chiffres similaires en termes d'URL vues (voir Tableau 3.10). C'est surtout en termes de sessions qu'elle montre son utilité : *CatService* décrit des pages dans 79 % des sessions des données BibUsages. Ceci ne doit pas nous surprendre : le fournisseur d'accès ou un portail généraliste figure souvent en page de démarrage automatique des navigateurs. Mais l'effet « page de démarrage » n'explique pas tout, et cette proportion doit également beaucoup au fait que les sites identifiés par l'application sont des nœuds de passage fréquemment visités par les internautes. Disposer d'informations précises sur ces nœuds est un atout précieux pour l'analyse des parcours.

En outre, *CatService* permet également d'identifier l'usage de certains types de sites particuliers : l'extraction des mots-clefs dans les requêtes adressées aux moteurs de recherche permet ainsi une étude poussée des usages des différents moteurs et de la reformulation des requêtes[1]. Dans le même ordre d'idées, c'est dans ce cadre que la

[1] Voir l'étude menée dans [Assadi & Beaudouin 2002], qui montre en particulier les spécificités des moteurs de recherche en fonction des requêtes qui leur sont adressées et du profil de leurs utilisateurs.

catégorie « bibliothèques électroniques » a été créée, que nous mobiliserons par la suite dans l'analyse des usages de ce type de sites. Ces deux exemples, qui seront développés par la suite, montrent que *CatService* permet non seulement une approche globale des sessions en pointant des types de contenus particuliers nécessitant des traitements adaptés, mais aussi une approche spécifique fine de certains types de sites dont on souhaite étudier les usages. La catégorisation semi-automatique apparaît, de ce point de vue, comme une approche très productive et efficace : si elle n'a pas vocation à couvrir l'ensemble des parcours, elle permet de sélectionner les sites que l'on souhaite décrire et de disposer, pour ces sites, d'informations précises et maîtrisées facilement exploitables par la suite.

> ***Synthèse.*** *L'application* CatService *permet d'appliquer des expressions régulières à des URL pour les rattacher à un référentiel à cinq niveaux : type de portail, portail, fournisseur, service, sous-service. L'outil autorise une description thématique ou fonctionnelle des contenus des sites Web. Sur les grands portails généralistes,* CatService *permet d'avoir une description synthétique et unifiée des différents contenus de ces sites.*

3.2 Aspiration de pages

Le deuxième axe de recherche que nous avons exploré pour qualifier les contenus des parcours consiste à analyser le contenu des pages visitées. Nous présentons ici la mise en œuvre de cette méthode de description endogène, les problèmes techniques qu'elle pose, l'exploitation que l'on peut en faire et le corpus que nous avons constitué autour des données issues du projet BibUsages.

3.2.1 Intérêt de la méthode, choix des outils

L'aspiration des pages visitées par les panélistes est une des pistes majeures pour qualifier le contenu des parcours. Elle apparaît comme la plus intuitive : en effet, pour prendre connaissance du contenu de navigation et en dégager la logique, le mieux n'est-il pas d'aller examiner les pages visitées ? Dans cette optique, la consultation des pages semble incontournable, ce que confirme l'examen manuel des pages formant le parcours *via RePlay* (voir 4.1.1, « Rejouer les parcours »).

Apports et contraintes de l'aspiration de pages

Du côté de la production des pages Web, les serveurs HTTP peuvent dans leur fonctionnement élémentaire renvoyer le contenu de fichiers statiques, figés, mais ils ont aussi la capacité de produire des contenus dynamiquement, c'est-à-dire que les données renvoyées au poste client seront générées pour chaque requête. En outre, le client a la possibilité de passer des paramètres à la requête : typiquement, l'interrogation d'un moteur de recherche revient à faire exécuter par le serveur un programme de recherche dans une base de données qui aura, entre autres paramètres, les mots-clefs demandés par l'utilisateur, et la page de résultat aura été composée pour répondre à cette requête particulière. À cela s'ajoute la capacité du serveur à créer des données persistantes du côté du client et à les interroger, par le mécanisme

des *cookies* ou des sessions, qui peut être vu comme un paramètre supplémentaire dans la composition des pages renvoyées. Cette double dynamique, exécution de programmes côté serveur et passage de paramètres par le client, fait des contenus Web des objets potentiellement très évolutifs, et leur confère une dimension de péremption intrinsèque. Ces éléments introduisent dans l'aspiration de pages des facteurs de complexité, voire d'infaisabilité, qu'il importe de décrire et de quantifier.

La première difficulté tient au fait même de reproduire *a posteriori* les requêtes adressées par les utilisateurs, ce qui induit plusieurs biais :

- Obsolescence ou renouvellement des pages liées au différé : les pages consultées peuvent avoir été modifiées ou bien avoir une fréquence de rafraîchissement très élevée, de sorte que nous ne voyons pas exactement ce qu'a pu voir l'utilisateur. Ainsi, une requête sur la page d'accueil du site du journal *Le Monde* (`www.lemonde.fr`) ne produira pas, à quelques heures d'écart, le même résultat. À cela s'ajoute le fait que, pour bien des cas, ce biais est très difficile à évaluer : même s'il est souvent possible d'identifier les requêtes HTTP pointant vers des scripts sur la base des extensions de fichiers et du passage de paramètres (*php*, *asp*, *java*, *perl*, etc.), rien ne garantit que le programme exécuté sur le serveur produise exactement le même contenu qu'au moment de sa consultation par l'utilisateur. Pour prendre un exemple concret, il est aujourd'hui possible à n'importe qui de mettre en place un forum Web, avec *PhpBB*[1] par exemple : toutes les requêtes adressées au forum pointeront vers des fichiers php, et renverront un contenu dynamique ; mais si personne n'intervient sur le forum, le résultat des requêtes sera toujours le même. À l'inverse, un fichier statique (html le plus souvent) peut avoir subi une ou plusieurs transformations, dont nous ne pouvons connaître l'étendue.
- Contenus personnalisés : certains services exécutent des opérations en fonction de paramètres soumis par l'utilisateur. C'est par exemple le cas des moteurs de recherche et plus généralement de tous les outils de recherche. Deux problèmes se posent : d'une part, les paramètres ne nous sont pas toujours connus *via* l'URL, lorsque ceux-ci sont envoyés au serveur via la méthode POST (voir en Annexe 2, Requêtes Web : mille-feuille technique). D'autre part, il n'est pas certain que la requête produira le même résultat que pour l'utilisateur ; dans l'exemple d'un moteur de recherche, sa base d'indexation en perpétuelle mise à jour implique que, passé un certain délai, une même requête donnera un résultat différent.
- Accès restreints : les protocoles sécurisés (HTTPS) et les accès par mot de passe interdisent d'accéder à certaines pages visitées par les panélistes. Cela concerne en particulier les transactions financières au sens large (banque en ligne, achat en ligne), le WebMail, et les profils personnalisés (du type *MyYahoo*, etc.), où l'identification de l'utilisateur est un passage indispensable.

[1] Système de gestion de forum sur le Web ; voir : `http://www.phpbb.com/`.

Face à cette série de problèmes, nous déplorons finalement qu'un module de récupération et de copie des contenus visités ne soit pas intégré aux sondes de recueil de trafic, qui permette de disposer d'une copie exacte des contenus effectivement vus par l'utilisateur. Si cette fonctionnalité pose en elle-même des problèmes techniques (le rapatriement des données en particulier) et de confidentialité (nous aurions accès aux correspondances privées, aux relevés de compte bancaires, etc.), elle résoudrait tous ceux que nous venons d'énumérer et qui, *in fine*, grèvent l'analyse.

À cela, s'ajoutent des problèmes propres à la structure des pages Web et à leur production : l'utilisation des *frames*, en particulier, fait que certaines pages sont très pauvres en contenu (typiquement les *frames* de navigation). D'autres pages peuvent être de simples bandeaux de navigation, ou des formulaires d'authentification : elles prennent leur sens dans la globalité de l'interface de navigation et dans la dynamique de la consultation. Ces pages sont difficiles à identifier, et il n'est pas toujours évident de les replacer dans leur contexte de visualisation (celui de l'interface, mais aussi de la séquence d'action dans laquelle elles peuvent se trouver). Un exemple illustre cette difficulté : pour accéder aux jeux en ligne de Yahoo, il faut passer par une étape d'authentification (voir Figure 3.1).

Figure 3.1. Interface d'authentification de Yahoo Jeux France

Cette page n'a pas de sens en elle-même, et même si son vocabulaire est relatif à l'univers du jeu, elle ne constitue pas une finalité mais une étape de nature technique dans la navigation. On rejoint ici la question du typage des contenus Web, qui se déclinent autant en contenus « à lire » qu'en services, où la page s'inscrit dans une procédure où sa fonction est le déterminant majeur de sa place dans le parcours.

Enfin, il faut mentionner l'ensemble des éléments non textuels, ou textuels mais non codés comme tels : les images et animations multimédia (au format Flash en particulier), qui peuvent être très riches en termes de contenu, échappent à notre analyse. Dans ce cas, même si l'on peut avoir une copie des pages visitées dans leur ensemble (fichier HTML et ensemble des images, sons, etc. qui la composent), on ne dispose pas à ce jour d'outils permettant de traiter cette complexité dans son ensemble.

Le choix de l'outil

L'utilisation des résultats d'aspiration apparaît, malgré ces difficultés, comme une piste intéressante, et qu'il importe, sinon d'exploiter, du moins de tester pour évaluer précisément les problèmes qu'elle pose. Pour ce faire, nous avons besoin d'une solution logicielle capable de rapatrier des pages Web à partir d'une liste d'URL et de les stocker.

Cette opération, pour naturelle et simple qu'elle puisse paraître, n'est pas évidente ; elle se complique même sensiblement si l'on souhaite aspirer des sites entiers et non seulement des pages[1]. Néanmoins, dans la perspective de la construction de corpus pour l'analyse des parcours Web, deux éléments peuvent motiver cette ambition : d'une part, si une page visitée se révèle pauvre en contenu, une solution peut être de ne pas utiliser cette page seule, mais le site entier. D'autre part, il apparaît intéressant de comparer les contenus visités par un internaute avec l'ensemble du site : on peut ainsi voir, à ce point de rencontre entre l'offre de contenu et sa réception, ce que l'internaute prend et ce qu'il laisse de côté.

Les impératifs techniques de formats de stockage, de temps d'aspiration et d'exhaustivité nous ont amené dans un premier temps à développer en Java un logiciel spécifique capable de pratiquer plusieurs aspirations simultanées (*multithreading*), de supporter les protocoles HTTP et FTP, et de garder trace d'une série d'informations sur le contenu des pages et le déroulement de l'aspiration données dans les en-têtes HTTP. Le développement de cette application, est apparu indispensable pour constituer un corpus directement exploitable ; l'outil était en outre destiné à être réutilisé dans les autres applications de traitement de données de trafic, et à cette fin, il a été conçu comme relativement générique et facilement paramétrable.

Par la suite, nous avons laissé de côté cette solution logicielle spécifique qui s'est avérée incomplète et surtout inapte à l'aspiration de sites. En lieu et place de cet outil, nous avons utilisé le module d'aspiration intégré à la plateforme de traitement de données de trafic développée dans le cadre du projet SensNet, *SensNetAspi*.

La fonction du module d'aspiration *SensNetAspi* est de créer une copie locale des pages d'un site Web ou d'un parcours d'internaute selon certains critères de sélection dans le système SensNet afin d'en traiter le contenu ultérieurement. Ce module est

---

[1] Voir les problèmes de récupération et de stockage des contenus décrits dans [Beaudouin *et al.* 2001], notamment pour les sites marchands.

entièrement intégré aux outils de la plateforme (voir Figure 3.2), et directement interfacé aux données de trafic, ce qui est une de ses principales valeurs ajoutées.

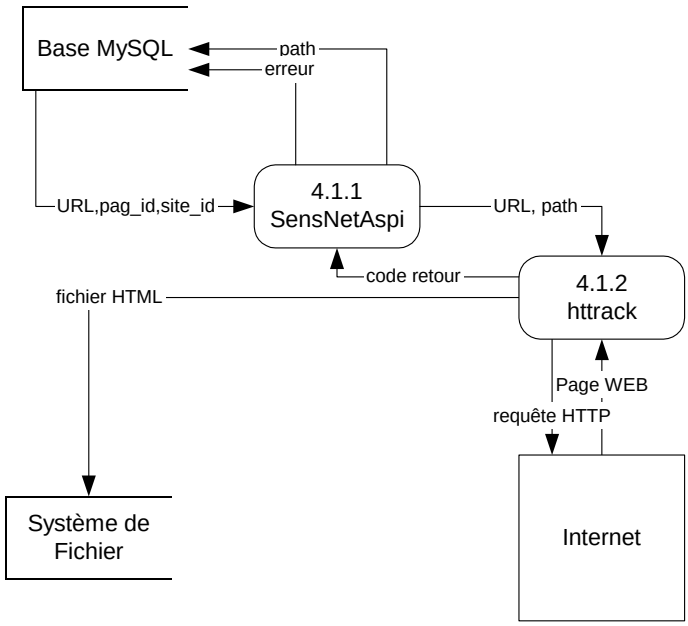


Figure 3.2. Fonctionnement du module SensNetAspi

Le module *SensNetIndic* permet trois types d'aspirations :

- l'aspiration d'un site Web,
- l'aspiration d'un parcours d'utilisateur, sur la base d'un identifiant de session,
- l'aspiration de page, à partir d'un identifiant d'URL ou d'une URL seule.

L'aspiration proprement dite encapsule un aspirateur existant sous licence GPL, HTTrack[1], dont les performances et la souplesse ont motivé le choix.

Ce module d'aspiration est adossé à un module de conversion des corpus HTML en XML, *SensNetXRef*[2], et un autre module de constitution et de manipulation de sous-corpus et d'indicateurs, *SensNetCorpus*. À partir de pages ou sites HTML, *SensNetXRef* construit des corpus en format XML fournissant une représentation structurée et exhaustive de l'ensemble des informations relatives au contenu, à la structure et à la forme d'une page ou d'un ensemble de pages donné. Les différents traits produits, outre le texte de la page ou du site, peuvent se compter par centaines : images contenues, liens externes et internes, texte des liens, formulaires, polices utilisées, taille de caractères, etc. Sur cette base, le module *SensNetCorpus* extrait et

[1] Voir http://www.httrack.com.

[2] Ce composant se base sur le logiciel Webxref, modifié par Serge Fleury dans le cadre du projet TypWeb ; voir http://www.cavi.univ-paris3.fr/ilpga/ilpga/sfleury/outilsSensnet.htm.

formate des sous-corpus pour les analyser afin d'y déceler les indicateurs qui les caractérisent le mieux.

Nous disposons ainsi, avec ce système complet et directement interfacé aux données de trafic, d'un outillage performant pour la description des pages et l'analyse des parcours. Il importe ensuite de sélectionner les éléments descriptifs pertinents au sein des sorties de *SensNetXRef* pour l'exploitation des corpus en vue de la description de sessions et des parcours d'internautes.

> ***Synthèse.*** *L'aspiration des pages visitées par les internautes à partir des données de trafic permet de connaître exactement le contenu des parcours, même si elle se heurte au double biais de l'accès différé et des accès restreints. Le module* SensNetAspi *développé dans le cadre du projet SensNet permet de constituer un tel corpus de pages à partir des données de trafic, et d'en extraire l'ensemble des informations de contenu, de structure et de mise en forme des pages.*

3.2.2 Exploitation de corpus de sites et de pages

Nous ne sommes bien évidemment pas les premiers à tenter d'exploiter des corpus constitués à partir du Web, quoiqu'il ne nous semble pas qu'aucun corpus ait déjà été constitué dans l'objectif d'analyser les parcours sur la Toile. Dans cette perspective, les travaux déjà menés peuvent nous être d'une grande utilité, notamment les études autour du typage des pages et des sites Web qui peuvent nous permettre d'envisager de rattacher nos pages ou sites aspirées à des éléments descriptifs plus généraux et plus synthétiques.

Genres du Web

Les travaux sur les genres du Web trouvent leur origine dans les travaux de Karlgren ([Karlgren & Cutting 1994]), et la présentation d'un prototype de logiciel, *Easify* (voir [Karlgren *et al.* 1998] et [Bretan *et al.* 1998]), qui permet de classer des documents issus du Web selon une série de paramètres, y compris le genre. Les genres sont ici rapprochés de la notion de « variation stylistique », et sont opposés au plan du contenu. Ces éléments stylistiques peuvent, nous dit l'auteur, être trouvés aux niveaux « lexical, syntaxique ou textuel : chacun a peu d'importance en lui-même, mais prises ensembles, leurs variations indiquent des différences systématiques »[1]. Une « palette de genres » est définie à partir des « impressions » des internautes, qui regroupe onze genres spécifiques aux pages Web :

- pages personnelles (« informal, private : personal home pages ») ;
- sites commerciaux (« public, commercial : home pages for the general public ») ;

[1] « Stylistic items can be found on any level of linguistic abstraction: lexical, syntactic, or textual; each is of little importance in itself, but taken together their variation indicate systematic differences. »

- pages interactives (« interactive pages: pages with feed-back : searchable indexes, customer dialogue ») ;
- matériel journalistique (« journalistic materials: press: news, editorials, reviews, e-zines ») ;
- rapports (« reports: scientific, legal and public materials ; formal text) ;
- autres textes (« other texts ») ;
- FAQ (« FAQs ») ;
- pages de liens (« link collections ») ;
- autres tableaux et listes (« other listings and tables ») ;
- forums de discussion (« discussions ») ;
- messages d'erreur (« error messages »).

On ne s'étonnera pas de retrouver parmi les éléments servant à classer les documents des éléments que nous qualifierons de paratextuels, liés au support HTML des documents et à leur dimension hypertextuelle : nombre de liens, liens interne ou externe au site, nombre et proportion d'images, etc. Un algorithme de classement permet de ranger les documents dans telle ou telle catégorie.

Dans « Web Genre Visualization » ([Dimitrova *et al.* 2002]), Dimitrova propose un autre prototype d'outil permettant « d'aider l'utilisateur à trouver rapidement des documents d'un genre approprié »[1]. En 2002 également, Rehm présente dans « Towards automatic Web genre identification » ([Rehm 2002]) une analyse en corpus du genre « Academic Personal Homepage » qui devrait permettre une identification automatique et l'extraction des informations que contiennent les pages qui en relèvent. Plus tôt, Crowston s'intéressait aux pages relevant du genre *Frequently Asked Questions* ([Crowston & Williams 1999]) et s'interrogeait sur les éléments constitutifs de ce genre.

De manière générale, ces travaux nous paraissent à la fois intéressants et insuffisants. Nous leur reprochons surtout d'utiliser abusivement la notion de genre : dans [Karlgren & Cutting 1994], les éléments définitoires des genres sont hétérogènes, mêlant des oppositions d'ordre technique (page personnelle *vs.* commerciale), de contenu (« journalistic materials » *vs.* « reports »), etc. ; dans [Dimitrova *et al.* 2002], les genres sont réduits à « une classification du document selon des dimensions comme le degré d'expertise du document, le degré de détails qu'il contient, ou selon que le document rapporte essentiellement des faits ou des opinions »[2] ; dans [Rehm 2002] les genres sont rapprochés d'une taxinomie et un document sur le Web peut relever de plusieurs genres.

Cela étant, ces travaux partent d'un constat simple, l'observation de régularités dans certains « types » de pages, et l'on y retrouve souvent des oppositions entre les pages personnelles et les autres types, ou une attention marquée pour certaines pages semblant répondre à des règles de composition marquées (les FAQ, par exemple).

---

[1] « We propose a simple visualization tool that helps users rapidly find genre-appropriated documents. »

[2] « We define the 'genre' of a document to be a classification of the document according to dimensions such as the degree of expertise assumed by the document, the amount of detail presented, or whether the document reports mainly facts or opinions. »

Face à ce constat indiscutable, l'approche typologique propose une démarche inductive.

Approche typologique

Les analyses typologiques de pages et de sites Web ont fait le même constat mais sans parler de genres : les différents travaux d'Amitay, par exemple, tournent l'analyse vers le typage, exploitent l'ensemble des éléments spécifiques aux contenus Web (mise en forme, éléments multimédia, liens), et relient les logiques de composition de pages à des systèmes de conventions ([Amitay 1997], [Amitay 1999]).

En 2000, Amitay et Paris présentent, dans « Automatically summarising Web sites - Is there a way around it? » ([Amitay & Paris 2000]), un outil, *InCommonSense*, qui utilise les descriptions accompagnant les différents liens vers une page donnée pour décrire cette page. Le système est capable de sélectionner, dans les différentes descriptions de sites ainsi obtenues, la plus représentative et la plus pertinente.

Plus récemment, dans « The connectivity sonar » ([Amitay *et al.* 2003]), Amitay *et alii* proposent une méthode de classification fonctionnelle des sites sur la base de leur structure interne, en dehors de toute analyse de contenu. Les auteurs font l'hypothèse que le type d'un site est étroitement lié à sa structure (sa taille, l'organisation de ses pages en répertoires et sous-répertoires, les liens internes et externes), et que celui-ci peut être retrouvé à partir de celle-ci[1]. Pour le vérifier, 296 sites sont classés manuellement dans les huit catégories : « corporate sites, content & media sites, search engines, Web hierarchies & directories, portals (both general Web portals and community-specific portals), E-stores, virtual hosting services and universities ». Ensuite, à partir de 73 indicateurs structurels de base, 16 modalités synthétiques sont calculées pour rendre compte de la structure des sites. Une classification faite à l'aide de la méthode des arbres de décision permet de comparer les classes faites manuellement et automatiquement. La précision finale obtenue est de 55%, et les auteurs proposent d'associer à l'analyse structurelle des éléments de contenu pour obtenir de meilleurs résultats, en y incluant des heuristiques spéciales et propres aux propriétés de chaque type de site.

Si nous pouvons opposer à cette étude de n'avoir pas mieux motivé le choix de ses classes de sites, dont certaines semblent se recouper (portail et moteur, en particulier), elle n'en montre pas moins la corrélation forte entre éléments structurels et type de sites, et la nécessité de tenir compte de l'ensemble des éléments structurels propres au Web.

Dans une perspective différente, et plus large en ce qui concerne les traits retenus pour décrire les sites et les pages, Ivory et Hearst proposent aux concepteurs « amateurs » de sites un outil permettant d'améliorer leur site en le comparant à la

[1] « Since sites are created for different purposes and by different people, it should come as no surprise that they sport different designs: the sizes of the sites, the organization of the pages in directories and subdirectories, the internal linkage patterns within the site's pages and the manner in which the sites link to the rest of the Web »

structure et à la forme de sites de qualité ([Ivory & Hearst 2002]). Pour cela, ils ont analysé au sein de corpus de sites récompensés aux « Webby Awards », la répartition de 157 traits formels regroupés au sein de neuf catégories :

- Éléments textuels : volume, qualité et complexité du texte ;
- Liens : nombre et types de liens ;
- Éléments graphiques : nombre et types d'images ;
- Formatage du texte : polices utilisées, mise en forme (taille, casse, etc.) ;
- Formatage des liens (couleur, souligné, etc.) ;
- Mise en forme des éléments graphiques : taille des images, place occupée dans l'ensemble de la page ;
- Mise en forme de la page : utilisation de couleurs, de polices, feuilles de style ;
- Performances de la page : volume, temps de chargement, erreurs dans le code HTML ;
- Architecture du site : profondeur, taille, importance des différents éléments.

Un classement manuel en bons, moyens et mauvais sites est appliqué à plus de 300 sites. Un sous-corpus permet d'entraîner le système pour l'application d'un algorithme de classification par arbres de décision (« Classification and Regression Tree algorithm »). Les 144 règles extraites sont ensuite appliquées au reste du corpus afin de vérifier leurs performances à décider si les sites se rapprochent ou non de sites « de qualité », et en quoi ils pourraient être améliorés. Le système obtient des bons résultats, avec une précision de 94 %.

Ce travail montre la capacité des indicateurs formels et structurels à rendre compte du rendu visuel et de l'ergonomie des sites, qui rentrent fortement dans leur évaluation. Il permet également de supposer, à l'inverse, que les sites développés par des webmestres professionnels et des « amateurs avertis » sont identifiables sur la base de ces traits structurels et formels. Sans présager du contenu précis des sites, de tels éléments donnent des indices sur l'ambition des webmestres et l'audience supposée des sites qu'ils administrent, et tendent à distinguer les sites conçus pour un faible public et ceux destinés à une plus large audience.

Plus près de nous, les projets TypWeb et SensNet suivent une démarche similaire dans la description des contenus. L'objectif est de parvenir à « faire émerger, de manière inductive, des typologies sur la base des corrélations observées entre des indicateurs portant sur l'outillage grammatical et le lexique, sur la structuration textuelle et hypertextuelle, et sur l'aspect multimédia. » ([Beaudouin *et al.* 2001]). Le projet s'appuie sur le logiciel WebXRef, modifié par S. Fleury pour produire des corpus normalisés au format XML rendant compte de l'ensemble des éléments textuels, structurels et formels de pages et de sites pour leur analyse à l'aide de traitements matriciels (voir Figure 3.3 ci-dessous).

L'analyse de corpus volumineux de sites personnels et commerciaux constitués en 2000 et 2001 montre en premier lieu, outre les difficultés techniques liées à l'aspiration de sites, la difficulté de faire émerger des traits saillants au sein des milliers de traits descriptifs générés. Au-delà de ces difficultés, ces travaux ont mis en lumière des oppositions fortes entre sites marchands et sites personnels autour de la taille des sites, de l'emploi des pronoms personnels (le « vous » chez les premiers

s'opposant au « je » des seconds) et du nombre de liens internes et externes (voir [Beaudouin *et al.* 2003b]). Ces écarts traduisent des logiques d'ouverture et d'interaction avec les visiteurs des sites bien différenciées, qui se prolongent au sein du corpus de sites personnels, où s'opposent les sites à tendance professionnelle et les sites de webmestres amateurs (en cela, cette étude rejoint les éléments mis en évidence par [Ivory & Hearst 2002]). L'étude a également montré des spécificités thématiques des différents hébergeurs de sites personnels sur la base du contenu textuel des pages hébergées.

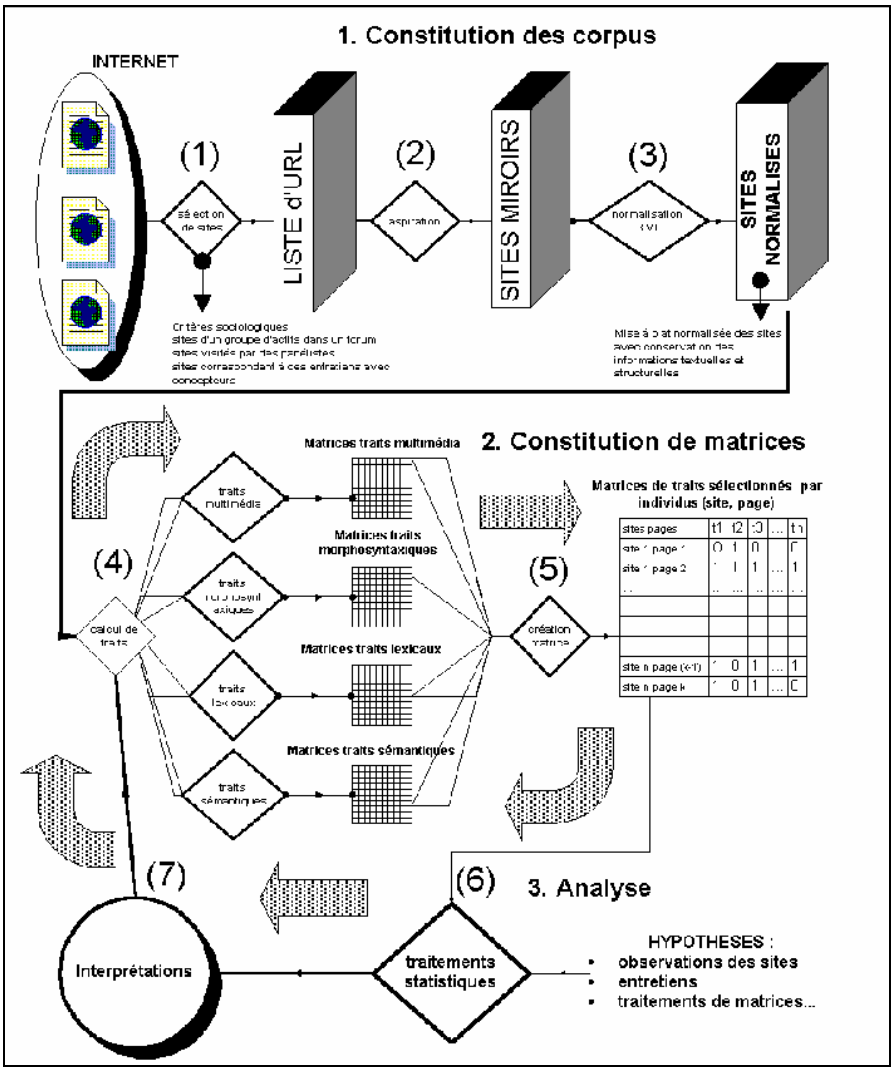


Figure 3.3. Système de catégorisation de sites et de pages dans TypWeb (présenté dans [Beaudouin et al. 2001])

En outre, ce travail pointe les différences fortes entre types de pages à l'intérieur d'un site, fondées sur leur fonction dans le cadre de la navigation, donc spécifiques aux contenus Web :

> Nous pouvons en effet identifier dans notre corpus une ligne de partage entre les pages à contenu et les pages qui facilitent la navigation à l'intérieur du site. Dans l'ensemble de ces pages d'orientation (qui correspondent à 15 % des pages visitées), peuvent être distinguées : les pages de redirection qui pointent vers la nouvelle localisation du site (nous avons vu que cette pratique est loin d'être négligeable) ; les pages de menu qui donnent accès aux différentes rubriques du site (elles peuvent se présenter comme une page autonome ou être inscrites dans une page à contenu) ; les pages de listes qui regroupent des pointeurs vers d'autres pages du site.[1]

L'implication est double : en premier lieu, pour l'analyse des corpus, un traitement différencié s'impose entre pages « à contenu » et pages « d'orientation », en particulier pour la mise à contribution du contenu textuel des pages. Ensuite, le repérage de ces pages fournit une information intéressante en elle-même pour une analyse des parcours tirant profit des descriptions fonctionnelles des pages.

Corpus de sites et de pages pour l'analyse des parcours

Que devons-nous retenir de ces différents travaux pour l'analyse des parcours ? Tout d'abord, la distinction intuitive de différents types de sites se retrouve dans les différents niveaux descriptifs des contenus proposés : lexique, grammaire, volume total et textuel, images, liens externes et internes, solutions techniques mobilisées par les concepteurs, etc. Ces différents éléments peuvent, pris isolément, servir à opposer certains types de sites, ce sont surtout les approches multi-critères qui obtiennent les meilleurs résultats, en intégrant dans leurs analyses les spécificités des contenus disponibles sur le Web. D'autres ont fait ce constat avant nous : les moteurs de recherche, par exemple, ont connu une amélioration significative de leurs résultats lorsqu'ils ont dépassé une analyse bornée au lexique pour inclure les méta-données et surtout la dimension hypertextuelle de la Toile comme indice de « vote » pour une page et un site donné[2]. Si l'analyse des parcours suit des objectifs différents de ceux des moteurs, il s'agit bien de voir quels éléments doivent être retenus dans les pages en exploitant le substrat technique et hypertextuel sous-jacent à la publication sur le Web.

Cela étant, on constate également que la mise en avant de tel ou tel trait est étroitement liée à la finalité du traitement. Le passage au sein de TypWeb d'une approche inductive pure à une approche hypothético-déductive basée sur des oppositions de types socialement attestés en est pour nous le signe le plus évident : devant la multiplicité des traits disponibles, il devient nécessaire de formuler des hypothèses sur les catégories cibles pour la constitution nécessaire d'agrégats au sein des descripteurs.

Dans cette perspective, les catégories mises en avant par les travaux que nous venons de citer ne nous satisfont pas tout à fait. Nous avons affaire soit à des oppositions locales (sites personnels *vs*. sites commerciaux), soit à des classes hétérogènes (genres du Web), soit à des classes trop faiblement motivées pour être convaincantes. En outre, notons que, quand bien même les éléments de classification

---

[1] [Beaudouin *et al.* 2001], p. 41.

[2] En particulier la technologie PageRank développée par Google.

nous satisferaient, nous serions bien en peine de mettre en œuvre les techniques proposées, les auteurs n'explicitant jamais précisément les algorithmes de calcul et les poids accordés aux différents traits entrant dans la description des pages.

Dès lors, l'établissement de méthodes de typologie de pages dépassant le cadre de notre travail, nous tenterons à défaut une approche purement linguistique, basée sur le contenu textuel des pages visitées dans les parcours. À l'aide de *CatService*, nous pouvons également gérer l'opposition entre pages « à contenu » et pages « de navigation » constatée dans [Beaudouin *et al.* 2003b], en distinguant les pages correspondant à des services de communication, de recherche ou d'information sur les portails généralistes des autres. Ces éléments alimentent une analyse de la thématique des sessions sur la base du lexique des contenus visités, modérée par un filtre fonctionnel. Cette approche basique et exploratoire permet d'évaluer la capacité du contenu des pages à représenter les thématiques des sessions, et leur possible mobilisation dans une description plus vaste des parcours sur le Web, ce que nous avons mis en œuvre à partir de données issues du projet BibUsages.

> ***Synthèse.*** *La description des pages et des sites Web en termes de genres ou de types permet d'appréhender la diversité et la spécificité des contenus Web, et de les replacer dans la perspective de logiques de production particulières. Toutefois, aucune des classifications proposées à ce jour n'est tout à fait satisfaisante, et aucun outil ne permet de distinguer automatiquement un type de page d'un autre. Ceci nous interdit de mettre en œuvre ces approches pour l'analyse des parcours, et l'on tente une exploitation purement textuelle d'un corpus de pages vues.*

3.2.3 Expérience : corpus BibUsages

Le corpus que nous avons constitué est issu des données BibUsages. Ce panel étant constitué de personnes s'intéressant de manière générale aux contenus « à lire » en ligne[1], nous comptons obtenir dans leur trafic une part importante de pages au contenu textuel consistant et exploitable par la suite.

Critères de sélection

Si l'option maximaliste nous invitait à aspirer *in extenso* l'ensemble du trafic Web du panel BibUsages, les contraintes techniques (temps et espace disque en particulier) nous ont invité à tempérer cette première intention. Face à cette contrainte, deux options se posent : l'approche par sessions, et celle par individus.

Dans la première, on sélectionne certaines sessions dont on suppose qu'elles apporteront des pages au contenu textuel consistant, sur la base de leur longueur et du type de sites et de services visités en particulier. Cette approche pose le problème des critères de sélection à appliquer pour le choix des sessions à aspirer. Si elle se révèle intéressante lorsque l'on souhaite étudier des parcours répondant à certains

[1] Voir chap. 5.1, « Description des panels » pour une description complète du mode de constitution du panel BibUsages.

critères particuliers en termes de contenu, de services, de durée, etc., elle s'avère problématique si l'on souhaite traiter les sessions sans filtre *a priori* : on se voit alors contraint de postuler que telles pages, présentes dans certaines sessions, sont plus susceptibles d'être exploitables par la suite, ce que l'on ne peut affirmer sans l'avoir effectivement vérifié.

Dans cette perspective, nous avons choisi de constituer un corpus de pages par le filtre de l'individu, c'est-à-dire d'aspirer l'ensemble des pages visitées par un individu sur l'ensemble de la période d'observation. Cette seconde approche jouit de deux avantages : en premier lieu, aspirer l'ensemble des parcours de certains individus est cohérent avec l'approche centrée-utilisateur qui est la nôtre. Il est alors possible d'appréhender la diversité des contenus visités par chaque internaute, et de répondre à une série de questions connexes : quels sont les effets de régularité observables d'une session à l'autre ? La même diversité est-elle observable d'un individu à un autre ? Dans des contextes similaires, des individus ont-ils le même profil de session, en d'autres termes, la tâche prime-t-elle sur les déterminations individuelles ? En second lieu, l'approche par individus renseigne sur la constitution de corpus de pages visitées en général, et permettra de voir quels types de pages renvoient des contenus exploitables. En ce sens, elle apparaît comme un préalable à l'approche par sessions, à laquelle elle fournit un cadre général pour sa mise en œuvre.

Enfin, la sélection d'aspirations par individus n'est pas complètement exclusive de l'approche par le contenu des sessions : il est tout à fait possible de sélectionner des individus ayant accédé à certains types de contenus ou de services en particulier. C'est d'ailleurs un peu ce que nous faisons en constituant notre corpus : en travaillant sur les données BibUsages, nous centrons l'analyse sur des internautes enclins à visiter des fonds numérisés et plus généralement des « contenus à lire ».

Le choix des individus au sein des 72 participants à l'expérimentation BibUsages a pour sa part été guidé par un souhait méthodologique légitime : nous avons retenu les seize panélistes qui ont été interviewés (voir Tableau 5.8, p. 178). Dans ce cadre, nous pouvons mettre à profit la complémentarité des différentes approches (recueil de trafic, qualification des contenus, éléments de description des panélistes, entretiens) afin de disposer pour ces individus de l'ensemble des descriptions de contenus et de pratiques que nous pouvons mettre en place.

Phase d'aspiration

Le trafic total des seize interviewés de BibUsages, nettoyé des bannières publicitaires et autres requêtes non sollicitées, représente pour six mois d'activité plus de 451 000 requêtes, pour près de 210 525 URL uniques, sur 13 300 sites différents au cours de 6 005 sessions. L'aspiration des pages visitées a été effectuée en avril 2003, pour un trafic réalisé entre juin et décembre 2002. Au terme de cette première phase réalisée à l'aide du module *SensNetAspi* de la plateforme SensNet, 183 873 aspirations (87,3 % du total demandé) sont lancées et tracées dans le système, laissant de côté 26 652 pages considérées comme des échecs d'aspiration.

Sur ces 183 873 aspirations ayant été réalisées, 152 134 contenant au moins un fichier aspiré, soit 83,8 %, ce qui correspond aux aspirations réussies. Pour les 31 739 autres, l'aspiration est un échec, et le fichier de *log* produit par HTTrack nous permet d'identifier la source de cet échec :

- pour 20 428 pages (un tiers des échecs d'aspiration), nous n'avons pas de code retour HTTP. L'erreur renvoyée par HTTrack est ventilée comme suit :

Message	Fréquence	Pourcentage
Receive Time Out	9 822	48,1 %
Unable to get server's address	5 151	25,2 %
Connect Time Out	3 340	16,3 %
Receive Error	1 539	7,5 %
No data (connection closed)	395	1,9 %
Autre	181	0,9 %

La principale cause d'erreur est due au dépassement de délai de réponse, que ce soit du côté du client ou du serveur.

- pour 11 311 pages, la requête est correctement traitée et le serveur Web renvoie un code retour HTTP, réparti ainsi :

Code HTTP	Signification	Fréquence	Pourcentage
204	Pas de contenu	113	1,0 %
30x	Redirection	17	0,2 %
40x (400, 401, 403, 404, 408)	Requête incorrecte	8 882	78,5 %
5xx (500, 501, 502, 503, 508, 514)	Erreur interne du serveur	2 299	20,3 %

Dans le détail, ce sont surtout les codes de la famille 400 qui sont sources d'erreur, au sein desquels le 404 (« fichier non trouvé ») tient la plus grande place, ce qui nous renvoie directement au problème du décalage entre l'aspiration et le moment où le trafic a été produit.

Nous obtenons donc 152 134 aspirations réussies, soit 72,2 % du nombre total de documents que nous souhaitions initialement récupérer. Au sein de ce corpus, tous types de fichiers sont présents, il importe donc de voir quels fichiers sont exploitables dans le cadre de la création d'un corpus textuel. Pour cela, nous avons développé un module qui analyse les résultats d'aspiration et examine pour chacune d'entre elles le nombre et les types de fichiers récupérés. Nous constatons ainsi que si, à l'exception des fichiers de *cookies*, le nombre moyen de fichiers est de 6,3, la moitié des aspirations ne renvoient qu'un seul fichier, au format HTML dans 60 % des cas.

L'examen de la répartition globale des types de fichiers par aspiration (voir Tableau 3.11 ci-dessous) montre par ailleurs la prédominance générale des documents HTML dans l'ensemble, avec 76,5 % des aspirations renvoyant (au moins) un fichier de ce type. En ce qui concerne les images, on note la prédominance du format GIF, présent dans 63,7 % des aspiration contre 17,2 % pour le JPEG ; les fichiers Flash (extension « swf ») sont comparativement très peu présents (1,8 % des aspirations), sans doute plus réservés aux concepteurs professionnels (coût du logiciel, expertise nécessaire pour construire les objets Flash).

Tableau 3.11. Présence des types de fichiers dans les aspirations de pages

Type de fichier	Nombre d'aspirations incluant ce type	Part du total des aspirations
html	116 411	76,5 %
gif	96 936	63,7 %
cookies	65 141	42,8 %
jpeg	26 161	17,2 %
autres	3 134	2,1 %
swf	2 759	1,8 %
txt	2 650	1,7 %
png	2 240	1,5 %
js	2 200	1,4 %
pdf	1 471	1,0 %
zip	1 288	0,8 %
Autres	1 669	1,0 %

Clef de lecture : sur les 152 134 aspirations réussies, 26 161 contiennent au moins un fichier de type JPEG. Le nombre de fichiers par aspiration pouvant être supérieur à 1, le total dépasse le nombre d'aspirations réalisées.

Du corpus d'aspirations au corpus textuel

Notre corpus se compose donc de 152 134 aspirations effectivement réussies en reproduisant les requêtes effectuées par les 16 individus interviewés dans le cadre de l'expérimentation BibUsages. Pour l'ensemble de ces documents, nous avons conçu et appliqué un outil permettant d'extraire le texte contenu dans les documents suivant les règles suivantes :

- pour les fichiers au format texte brut, le contenu est copié tel quel ;
- dans le cas des fichiers au format HTML nous utilisons le navigateur en mode console Lynx[1], auquel nous passons les options nécessaires à l'élimination des traces de liens hypertextes, d'images et de formulaires[2] ;
- pour les fichiers PostScript et PDF, nous utilisons respectivement les outils `ps2ascii`[3] et `pdftotext`[4]. Dans les deux cas, l'extraction de texte n'est pas assurée, car ces deux formats peuvent tout à la fois encoder du texte en tant que tel ou en mode image. Dans ce deuxième cas, nous ne récupérons pas le contenu textuel des documents ; c'est le cas pour les documents téléchargeables sur Gallica, le site présentant les fonds numérisés de la Bibliothèque Nationale de France.

Au terme de ce traitement, quel corpus textuel obtenons-nous ? Sur près de 152 134 téléchargements réussis de documents visités, 20,9 % (31 740 aspirations) ne

---

[1] disponible sur `http://lynx.browser.org/`.

[2] soit les options : '`-dump -nocolor -nolist -hiddenlinks=ignore -pseudo_inlines -verbose=off`'.

[3] `ps2ascii` est inclus dans GNU Ghostscript ; voir `http://www.cs.wisc.edu/~ghost/`.

[4] `pdftotext` est inclus dans XPdf ; voir `http://www.foolabs.com/xpdf/index.html`.

peuvent renvoyer aucun contenu textuel : images, fichier compressés, etc. Restent alors quelques 120 400 documents pour lesquels nous pouvons avoir du texte : fichiers aux formats HTML, TXT, PDF et POSTSCRIPT. Nous ne sommes pas pour autant tirés d'affaire : ce corpus doit encore être nettoyé car il contient quelques scories :

- pages de redirection *via* l'en-tête HTML, mal interprétées par LYNX, qui renvoient un texte du type :

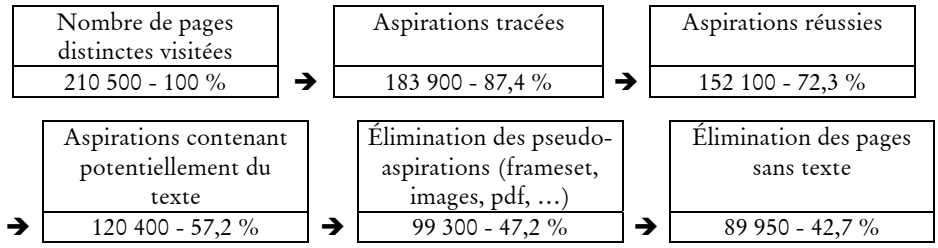
```
REFRESH(0 sec): http://sunearth.gsfc.nasa.gov/evsecef.htm
Click here...
```

  Ces pages sont filtrées en excluant les extractions de textes incluant la chaîne de caractères 'REFRESH([:num:] sec):' ; nous excluons ainsi 10 819 documents.
- Documents PDF en mode image, qui commencent par 'PDF-1.1' : 874 documents.
- Les *frameset*, reconnus via le repérage du texte 'FRAME:' répété dans le document, en conservant les documents contenant 'IFRAME:' (*frame* interne à une page), ce qui exclut 5 537 documents.
- images aux formats GIF (en-tête 'GIF89a' ou 'GIF87a') ou JEPG (en-tête contenant, en iso-latin-1, 'ÿØÿà') enregistrées par HTTrack comme fichiers HTML : 3 899 documents.

Au terme de ce nettoyage, nous obtenons 99 296 documents valides, c'est-à-dire qui comportent *a priori* du texte, sans présager de l'exploitabilité de celui-ci. Pour nous en assurer, nous avons compté le nombre de mots dans chaque extraction du texte du document[1] : au terme de ce calcul, sur près de 99 000 pages, 9,4 % ne contiennent aucun mot, et sont dès lors réputées inexploitables dans l'optique qui nous intéresse.

Restent ainsi 89 950 documents aspirés contenant plus d'un mot, issus de 9 042 sites différents, qui constituent en définitive notre corpus textuel exploitable. Nous savons par ailleurs qu'il reste quelques scories dans ce corpus : certaines aspirations ont renvoyé des javascript, des feuilles de style, des listes de répertoires ; cependant, ces dernières « pollutions » du corpus restent minimes, et c'est finalement sur ces presque 90 000 pages que nous travaillons par la suite.

L'ensemble du processus de sélection, aspiration, nettoyage, filtrage et transformation peut être résumé ainsi :



[1] La méthode utilisée ici est assez sommaire, mais suffisante pour le résultat qui nous intéresse : un mot est, pour ces comptages, défini comme une suite de caractères alphabétiques (expression régulière : « [A-Za-zéèêëâàîïùûôç]+ »).

Profil du corpus textuel

Une fois réalisée l'identification des aspirations exploitables dans le cadre de la constitution d'un corpus textuel, il s'agit d'examiner le profil de ce corpus. En termes de nombre de mots, les 90 000 documents contiennent en moyenne 436 occurrences, mais les plus longs d'entre eux pèsent lourd dans ce calcul, et la médiane s'établit à 171 mots, avec un quart du corpus contenant moins de 49 mots (voir Tableau 3.12).

Tableau 3.12. Caractéristiques générales du corpus de pages aspirées exploitables

		Nombre d'occurrence	Nombre de formes	Nombre de formes minusculisées
Moyenne		436	185	170
Médiane		171	112	105
Minimum		1	1	1
Maximum		15 512	5 324	4 497
Quartiles	25	49	38	37
	50	171	112	105
	75	480	234	214

On retrouve ici les ordres de grandeur déjà observés dans [Beaudouin *et al.* 2001] sur des sites personnels : deux corpus de sites personnels et deux corpus de sites marchands, tous francophones, avaient alors été constitués en 1999 et 2000. Pour les sites marchands, le nombre moyen d'occurrences par page était de 105 et 224, tandis que les sites personnels comptaient en moyenne 352 et 424 occurrences par page pour chaque corpus.

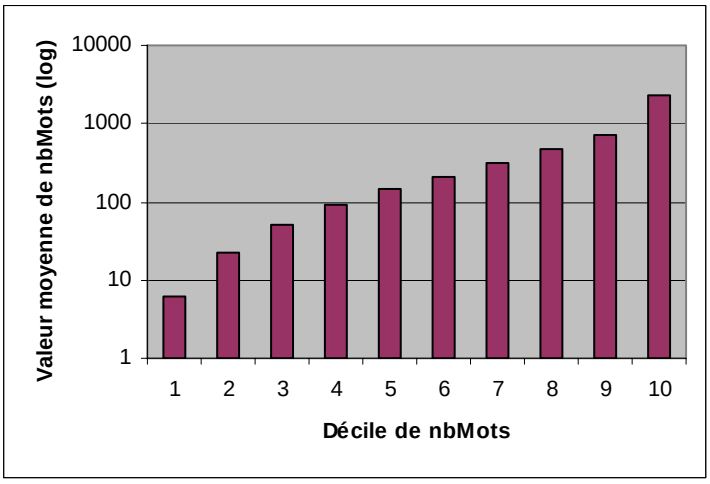


Figure 3.4 . Nombre moyen de mots par décile du nombre de mots (échelle logarithmique)

Si l'on scinde le corpus de documents en déciles du nombre d'occurrences qu'ils contiennent, on constate que la valeur moyenne du nombre d'occurrences pour chaque décile croît rapidement. La Figure 3.4, qui présente l'évolution de cette moyenne par décile en échelle logarithmique pour l'ordonnée, fait apparaître une

progression exponentielle du nombre d'occurrences lorsque l'on progresse vers les gros documents. On a donc beaucoup de documents au contenu textuel relativement court, et un faible nombre de documents très longs.

En termes de vocabulaire, on observe une distribution classique des formes et des occurrences pour un corpus textuel (voir Tableau 3.13) : nous avons plus de 913 000 formes différentes, au sein desquelles 46,3 % d'hapax font 1,1 % des occurrences, tandis que les 9,9 % de formes les plus fréquentes représentent 94 % des occurrences.

Tableau 3.13. Corpus BibUsages : répartition du vocabulaire

Valeur du nombre d'occurrences	Nombre de formes	Part de l'ensemble des formes	Nombre d'occurrences	Part du total occurrences
1	422 710	46,3%	422 710	1,1%
2	142 232	15,6%	284 464	0,7%
3	71 457	7,8%	214 371	0,5%
4 à 6	95 735	10,5%	457 264	1,2%
7 à 18	91 099	10,0%	988 656	2,5%
19 et plus	90 207	9,9%	36 841 663	94,0%
Total	913 440	100,0%	39 209 128	100,0%

La répartition des formes dans les documents suit une distribution similaire : si l'on examine le nombre de documents où est représentée chaque forme, en excluant les hapax, on constate que 13 % des formes sont présentes dans un document seulement, tandis que 18 % le sont dans plus de 13 documents, soit un peu plus de 88 000 formes (voir Figure 3.5).

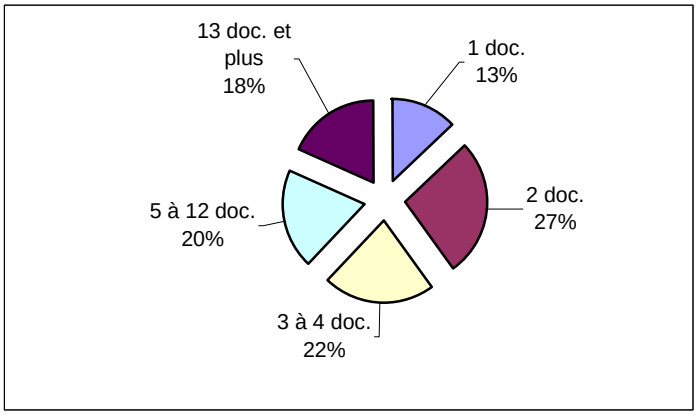


Figure 3.5. Représentation des formes de fréquence supérieure à 1 dans les documents

Couverture du corpus avec les parcours

Les 89 950 contenus textuels que nous obtenons *in fine* couvrent 42,7 % des 210 500 pages que nous souhaitions initialement ; cependant, en termes de pages

vues, elles représentent 183 550 requêtes sur 451 000 au total, soit 40,7 % du total des pages vues par le panel.

La couverture des sessions par le corpus constitué varie d'une session à l'autre : sur l'ensemble des sessions, 7,1 % ne sont pas couvertes du tout, 3,9 % le sont au contraire complètement, et le reste, 89 % l'est en moyenne à 43,4 % en nombre d'URL et 46,1 % en durée (voir Tableau 3.14).

Tableau 3.14. Couverture des sessions par le corpus

		Taux de couverture en nombre d'URL	Taux de couverture en durée
Moyenne		42,6 %	45 %
Médiane		42,9 %	42,8 %
Minimum		0 %	0 %
Maximum		100 %	100 %
Quartiles	25	25 %	17,7 %
	50	42,9 %	42,7 %
	75	57,9 %	70,9 %

Les moyenne et médiane de la couverture par session sont assez similaires sous l'angle de la durée et du nombre d'URL, entre 42 % et 45 %, mais la distribution du taux de couverture est très différente dans les deux cas : pour le calcul en nombre d'URL, elle suit presque une loi normale, tandis qu'elle est très uniforme pour le calcul basé sur la durée (Figure 3.6).

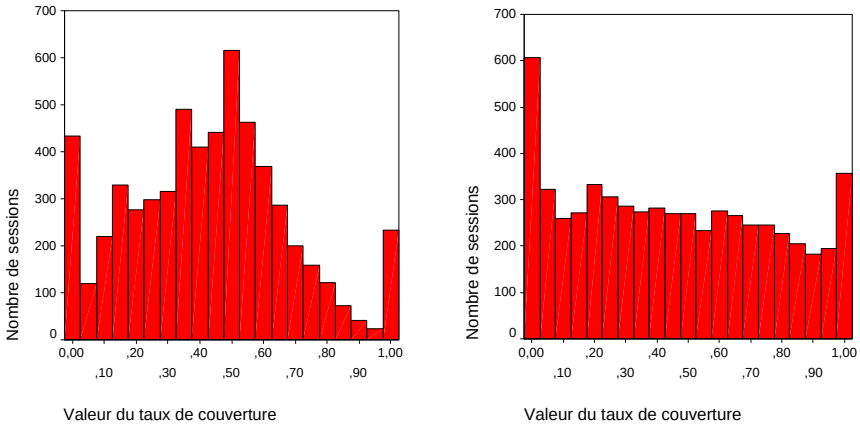


Figure 3.6. Répartition des valeurs du taux de couverture par session en nombre d'URL (à gauche) et en durée (à droite)

Ce contraste illustre bien la différence d'approche entre le travail basé sur les URL et celui fondé sur les durées. Pour l'heure, notons que cette différence devra être prise en compte au moment d'utiliser le corpus textuel pour qualifier les parcours : d'une part, nous pourrons être amenés à ne sélectionner que quelques sessions suffisamment couvertes pour mener nos analyses, auquel cas il faudra arbitrer entre couverture en nombre d'URL et couverture en durée. D'autre part, on pourra tenter de pondérer les descriptions textuelles de certaines phases d'un parcours en fonction

du temps ou du nombre de pages vues, et ces deux calculs interviennent encore dans ce cas.

Afin d'améliorer ces résultats, nous avons tenté de regrouper les différentes pages vues consécutivement sur un site, et de les décrire par la ou les pages de cette séquence. On suppose ici que les pleins comblent les creux, et que plusieurs pages vues sur un même site traitent globalement de la même chose, une page suffisant alors à décrire l'ensemble.

Nous avons donc obtenu près de 137 000 séquences, et projeté le corpus de pages aspirées sur ces séquences : les résultats ne sont finalement pas très différents de ceux obtenus à l'échelle de la page. Près de 48 % des séquences ne sont pas décrites du tout, et sur les 52 % restant, les deux tiers sont complètement décrites (soit 35 % de l'ensemble), le reste étant couvert pour moitié en moyenne. Ces résultats ne sont en réalité pas surprenants, car très peu de pages sont vues consécutivement sur un même site en moyenne : 70 % des séquences ne contiennent qu'une URL, seules 10 % en contiennent plus de 6. La description par séquences rejoint donc celle à l'échelle de la page, et le gain de performance est assez faible pour que l'on préfère travailler au niveau de la page par la suite.

Enfin, notons que la couverture globale finale des parcours des seize panélistes interviewés de BibUsages par le corpus constitué est globalement plutôt satisfaisante, dans la mesure où nous avons reproduit les requêtes effectuées avec un décalage allant de trois à neuf mois.

Questions de représentativité des données

Nous avons vu que le taux de couverture moyen par session est de 42,6 % en nombre d'URL, et de 46 % en durée, avec surtout des distributions très différentes pour les deux modes de calcul. Calculés par individu, les taux peuvent varier du simple au double (voir Figure 3.7).

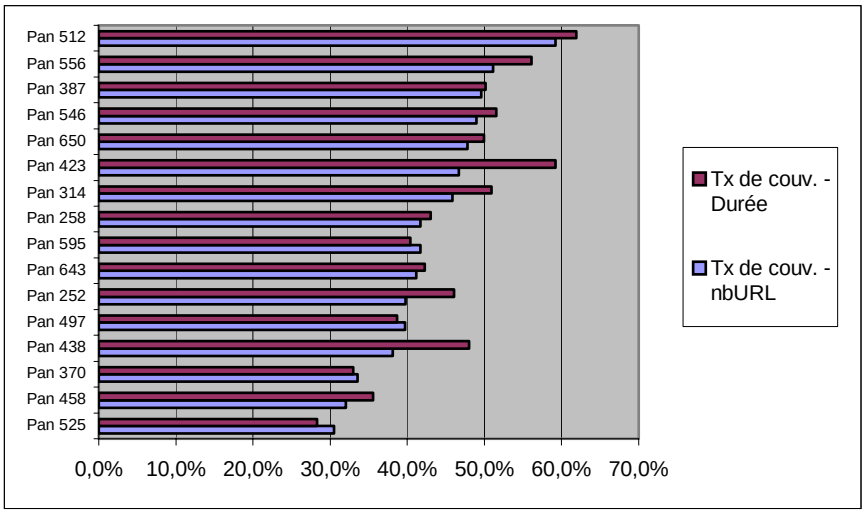


Figure 3.7. Couverture moyenne des sessions de chaque panéliste BibUsages

Bien évidemment, le taux de couverture implique que les sessions ne seront que partiellement décrites, en moyenne à moitié, parfois complètement et parfois pas du tout. Ce silence pose le problème de la représentativité des données textuelles dont nous disposons : quelle est la distorsion induite par cette couverture partielle ? Il faut ici distinguer les pages que nous n'avons pas pu recueillir de celles qui, correctement aspirées, n'ont pas de contenu textuel. Pour le second cas, on ne peut pas parler de distorsion, bien au contraire : la nature des contenus (images, programmes, contenus techniques comme les *frameset*, etc) est telle qu'il n'y a pas de texte, nous collons ici complètement au texte des pages visitées. Par contre, c'est dans la réussite des aspirations que le bât blesse : si les erreurs impliquant une réponse des serveurs (de type 404 : « fichier non trouvé », ou 500 : « erreur interne du serveur ») sont acceptables, les autres le sont moins, en particulier les dépassement de délai (« connect time out » ou « server time out »), et surtout les erreurs non tracées par l'application. Ainsi, sur 210 500 aspirations programmées, 26 600 sont « oubliées » par le système, et 20 500 renvoient une erreur non HTTP, ce qui au total représente 22,3 % des URL visitées. Pour cette partie du trafic, nous avons bel et bien un point aveugle ; cependant, pour les aspirations réussies, seules 59 % contiennent du texte récupérable, si bien que ce ne sont en définitive que 13 % de textes de l'ensemble des pages visitées qui nous manquent[1], proportion acceptable à notre sens.

Autre élément de suspicion à l'encontre de la représentativité de ce corpus, l'évolution des pages entre leur visite par les panélistes de BibUsages et la période d'aspiration : six à neuf mois séparent les deux événements, au cours desquels les contenus sont tout à fait susceptibles d'avoir évolué. Dans le cas de la disparition des pages, nous ne risquons pas de distorsion, nous ne récupérons qu'une erreur lors de l'aspiration, ce que reflètent les 8 882 erreurs de type 404 renvoyées par les serveurs Web. Pour les autres pages, dont les auteurs et les modes de production font que nous observons un contenu différent ; plus encore, il est quasiment impossible d'évaluer cette évolution, et d'en mesurer l'importance (page entièrement altérée ou bien modifiée à la marge). Néanmoins, nous restons sereins : on peut raisonnablement supposer que si le contenu d'une page évolue, il est dans une certaine continuité thématique et, partant, lexicographique, avec ses états antérieurs. En somme, les biais induits par l'aspiration différée des parcours ne nous semblent pas rédhibitoires, et diminuent la couverture des parcours sans en altérer profondément la représentation.

Impossible approche lexicale ?

Nous avons dans un premier temps retenu une approche lexicographique simple pour exploiter le contenu de ces aspirations de parcours. Nous faisons l'hypothèse qu'à l'échelle de la session, l'ensemble des textes contenus dans les pages Web sont à même de donner une représentation des thématiques des sessions. Nous avons

[1] On suppose, hypothèse acceptable, que les aspirations ratées contiennent autant de texte que celles qui ont été menées à bien.

cherché à classifier les différentes sessions pour chaque individu pris séparément[1], chaque session étant décrite par l'ensemble des textes disponibles qui la composent.

Les résultats de cette approche se sont révélés très mauvais : les classes construites par Alceste sont ininterprétables, le vocabulaire spécifique de chaque classe se révélant trop hétérogène pour renvoyer à une thématique particulière. Ce résultat ne nous a pas vraiment surpris : en incluant l'ensemble des pages visitées, on fait fi de la différence notable entre pages « orientées services », et pages « orientées lecture » ; et inversement, on inclut dans les descriptifs textuels des sessions les textes littéraires qu'ont pu visualiser les internautes de BibUsages, grands consommateurs de bibliothèques électroniques. Que viennent faire pêle-mêle *Les orientales* de Victor Hugo, la page de login de Yahoo et la liste des numéros de pages de Gallica pour décrire le contenu d'une session ?

Nous avons donc reconstruit nos corpus en excluant les pages relatives aux portails généralistes ainsi qu'aux bibliothèques électroniques ; en outre, nous avons exclu les sessions où moins de cinq sites ont été visités afin de travailler sur des unités élémentaires de taille comparable, et d'avoir un matériau textuel assez consistant pour l'analyse. Malgré ce double filtre, les résultats sont toujours aussi décevants : classes déséquilibrées et impossibles à interpréter.

Comment comprendre cet échec de l'approche lexicographique ? Nous avançons une explication qui renvoie à la spécificité des pages Web. À la différence des corpus textuels « classiques » manipulés en ingénierie linguistique, les pages Web agrègent dans un même objet des éléments de nature très différente. Les textes issus des pages Web sont particuliers par leur taille, leur vocabulaire comportant des lexies spécifiques au Web, les problèmes de grammaire et d'orthographe qu'ils posent, et la coexistence au sein d'une même unité d'éléments aux statuts bien différenciés. Dans l'exemple de la page d'authentification de Yahoo Jeux donné précédemment, nous pouvons distinguer plusieurs zones (voir Figure 3.8 ci-dessous) :

1. bandeau général de Yahoo Jeux : contient l'intitulé de rubrique dans Yahoo, ainsi que deux liens vers l'accueil général du portail et l'aide de la rubrique ;
2. bandeau contextuel informant l'utilisateur qu'il doit effectuer la procédure d'authentification pour poursuivre sa navigation ;
3. zone d'information destinée aux internautes accédant pour la première fois à cette interface et en explicitant la nécessité (il faut s'enregistrer pour accéder aux services de jeux) ;
4. partie spécifiquement technique : formulaire d'authentification proprement dit ;
5. pied de page générique au portail Yahoo, d'ordre juridique, comportant des informations sur la propriété intellectuelle qui régit la page et la confidentialité des informations fournies par l'utilisateur.

Au niveau strictement textuel, dans le cadre de corpus textuels « classiques » cette situation reviendrait à traiter conjointement texte et paratexte ; pour des corpus d'articles de journaux par exemple, chaque article serait parasité par le nom du journal, le bandeau contenant, dans l'édition papier, le prix du journal, la date et le

[1] Nous avons utilisé le logiciel d'analyse textuelle Alceste.

nom du rédacteur en chef, et l'ours. Le plus difficile dans cette situation est qu'il est difficile de repérer de manière automatique ces éléments dans les pages Web, où ils prennent des formes et des proportions variables.

Figure 3.8. Les différentes zones de l'interface d'authentification de Yahoo Jeux France

Dans notre corpus, nous avons tenté d'éviter ce biais en écartant les pages vues sur les portails généralistes ; mais en examinant de plus près l'extraction textuelle du corpus restant, force est de constater que ce comportement se retrouve dans une très grande part des pages Web. Comme le montre l'exemple donné Figure 3.9 ci-dessous, les éléments de navigation et d'information, les listes, les séries de liens occupent dans le lexique une place considérable, et rendent très difficile l'analyse textuelle hors de tout filtrage préalable.

```
                Vous possédez un Palm(TM) (ou compatible) ?

 Vous n'en possédez pas ? (mais vous vous intéressez à cette fabuleuse
petite machine) Venez nous rejoindre au prochain rendez-vous du


                           Club Palm(TM) de REIMS


           Inscription conseillée par e-mail : palmreims@fr.st

     Groupe de discussion: http://fr.groups.yahoo.com/group/palm-reims

     Et pour ne pas oublier le rendez-vous... demandez le programme !!!

Nouveau ! Téléchargez les horaires des cours de La Plage pour Palm !
   Logo "La Plage" Horaires de La Plage de Reims
```

```
FranceMap' : départements + vacances scolaires !

   Capture d'écran FranceMap     Capture d'écran FranceMap     Capture
   d'écran FranceMap

   J'ai adapté (avec son aimable accord) le programme FranceMap de Lars
   Empacher pour qu'il intègre les dates des vacances scolaires par
   zones pour la métropole, jusqu' en 2004.

   N'hésitez pas à télécharger cette nouvelle version : FranceMap' 1.2,
   qui est comme toujours gratuite !
   (dernière mise à jour : 23 juillet 2002)
```

Figure 3.9. Exemple de contenu textuel d'une page du corpus

Dans ces conditions, en l'absence d'un outil capable de repérer au sein de pages Web différentes « zones » aux fonctions distinctes, la caractérisation des sessions par le vocabulaire des pages nous semble impossible. Un tel constat va sans doute à l'encontre de l'engouement actuel de la communauté du TAL[1] pour les corpus Web : en 2003, dans l'introduction au numéro spécial de *Computational Linguistics* titré « Web as Corpus », Kilgarriff et Grefenstette s'enthousiasment pour cette inépuisable ressource :

> Le Web contient d'énormes quantités de textes, dans de nombreuses langues et divers types de langues. À nos yeux, le Web constitue un fabuleux terrain de jeu pour les linguistes. Nous espérons que ce numéro spécial [de *Computational Linguistics*] vous encouragera à vous joindre au jeu ![2]

En fait de jeu, nous aurons pour notre part fait le constat des difficultés que présentent les corpus issus du Web constitués sur la base des visites effectives des utilisateurs, dans toute la spécificité sémiotique de la production HTML et l'hétérogénéité générique et fonctionnelle des contenus proposés. Faute de disposer des outils et du temps suffisant pour décrire et représenter fidèlement et synthétiquement les contenus Web sur corpus, nous laisserons de côté l'approche par aspiration de pages pour la suite de ce travail.

> ***Synthèse.*** *En constituant un corpus de pages vues pour seize panélistes de BibUsages, on perçoit les difficultés que pose l'aspiration de pages* a posteriori *: un quart des aspirations échoue, et le corpus textuel représente moins de la moitié des pages souhaitées initialement. Le profil statistique du corpus s'apparente à celui d'un corpus textuel classique, mais les spécificités sémiotiques des pages Web (bandeaux de navigation, pages de formulaires, énumérations, etc.) rendent impossible l'exploitation strictement lexicale des aspirations. Dès lors, le typage des pages s'avère indispensable à la mobilisation d'un corpus de pages pour l'analyse des parcours.*

---

[1] Traitement Automatique des Langues.

[2] « The Web contains enormous quantities of text, in numerous languages and language types, on a vast array of topics. Our take on the Web is that it is a fabulous linguists' playground. We hope the special issue will encourage you to come on out and play! » ([Kilgarriff & Grefenstette 2003], p. 345).

3.3 Utilisation des annuaires

La troisième approche retenue pour décrire les contenus des parcours consiste, à la différence de l'aspiration, à faire appel à des données exogènes aux pages. Il s'agit d'exploiter les descriptions de pages ou de sites faites par les annuaires du Web, qui peuvent s'apparenter à des méta-données textuelles structurées. Nous détaillons ici la méthode utilisée et sa mise en œuvre, et donnons une description fine des annuaires utilisés dans leurs composantes textuelles et structurelles, étape indispensable à leur mobilisation dans l'analyse par la suite.

3.3.1 Méthode

Qu'est-ce qu'un annuaire du Web ?

S'inscrivant dans l'offre d'outils d'aide à la recherche de contenus et de services sur le Web, un annuaire propose à l'internaute un classement hiérarchisé de sites regroupés dans des catégories thématiques.

Par opposition aux moteurs de recherche, les annuaires Web proposent aux internautes un classement commenté de sites, lesquels sont organisés en catégories et sous-catégories. Parmi les plus connus, on trouve Yahoo, Nomade, ou encore Voila ; on y retrouve, pour chaque site ou page indexé, son adresse, son titre et une description de son contenu en quelques lignes. La Figure 3.10 présente un exemple de catégorie d'annuaire pour Yahoo France.

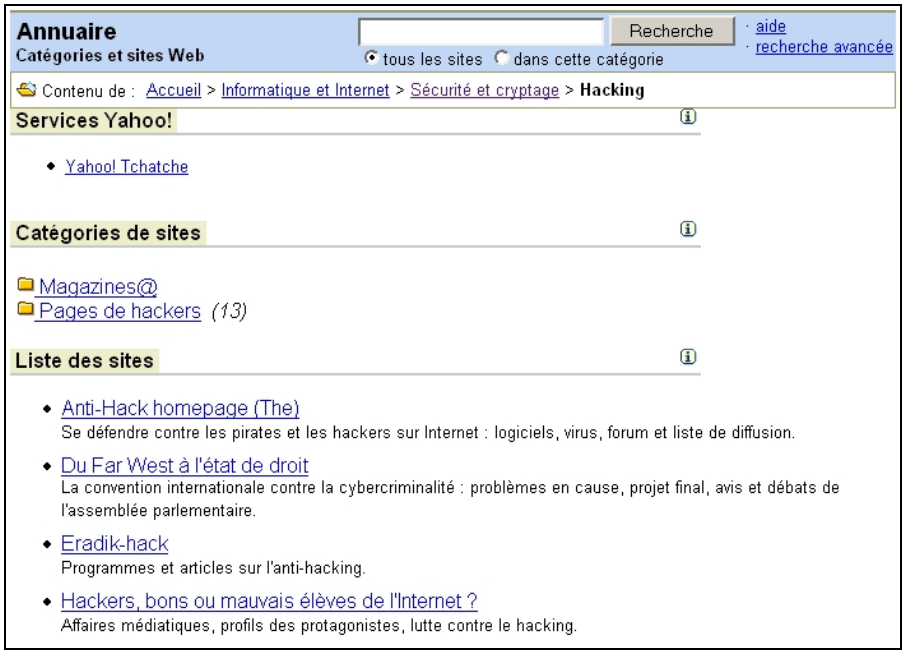


Figure 3.10. Un exemple d'annuaire : Yahoo France

En termes structurels, l'ensemble de ces catégories forme un arbre dont la racine est la page d'accueil et les nœuds, les différentes catégories de l'annuaire ; dans ces catégories, sont placés les sites ou pages indexés, qui sont accompagnés d'une description plus ou moins détaillée de leur contenu. La structure d'un annuaire peut être définie par le croisement de trois éléments :

- Multi-indexation : certains annuaires indexent la même URL dans plusieurs catégories ; une même adresse peut ainsi figurer plusieurs fois dans un même annuaire, à des endroits différents. Dans l'exemple de la Figure 3.10 ci-dessus, le site « Hackers, bons ou mauvais élèves de l'Internet ? » est également indexé par Yahoo dans la catégorie « Technologies de l'information et de la communication » :

- Position des URL indexées dans l'arbre : certains annuaires proposent des URL dans l'ensemble de leurs catégories, d'autres ne les classent que dans les catégories terminales (qui n'ont pas de catégorie-fille). Dans Yahoo, les URL sont réparties tout au long de l'annuaire : la catégorie « Hacking » ci-dessus n'est pas terminale, puisqu'elle contient une sous-catégorie « Pages de Hackers », et propose quatre URL.
- Utilisation des renvois : certains annuaires proposent, à l'intérieur d'une catégorie, des liens vers des catégories qui ne sont pas situées directement en dessous dans l'arbre, mais à un tout autre endroit de l'annuaire. Dans notre exemple, Yahoo propose des renvois, signalés par le signe « @ » à la fin du nom de la catégorie visée. Le lien noté « Magazines@ » pointe vers la catégorie « Hacking » dans :

Chaque annuaire du Web est, structurellement, une combinaison de ces trois éléments, et ces choix influencent sa taille et sa structuration globale.

Intérêt des informations fournies par les annuaires et mise en oeuvre

L'objectif est ici d'utiliser la description textuelle du site ou de la page indexée dans l'annuaire, ainsi que sa position dans les catégories et sous-catégories, afin de caractériser son contenu de manière thématique et fonctionnelle. Cette méthode de caractérisation des contenus présente plusieurs avantages :

- il n'est pas nécessaire d'aspirer les pages visitées par les panélistes, ce qui permet de s'affranchir des problèmes nombreux liés à l'aspiration que nous avons déjà évoqués ;
- les pages et les sites indexés sont situés dans une structure du monde en domaines et sous-domaines, forcément imparfaite, mais dont on peut se servir pour un typage des domaines vus par les utilisateurs ;

- les descriptifs de sites et de pages sont vérifiés manuellement par les indexeurs des annuaires : ils sont donc susceptibles d'être justes et précis.

À l'inverse, cette approche présente certains inconvénients, dont le plus important est d'indexer majoritairement des sites et non des pages : les descriptions faites concernent dans la plupart des cas un site dans son ensemble, dont elles ne fournissent qu'une présentation générale. Ainsi, si un utilisateur visite plusieurs pages à l'intérieur d'un même site, nous n'aurons pas accès au contenu spécifique de chaque page.

Deux campagnes de collecte d'information ont été menées, la première en février 2001, et la seconde un an plus tard. La première a concerné six annuaires francophones, dont cinq annuaires généralistes :

- Nomade (`http://www.nomade.fr`),
- MSN France (`http://search.msn.fr/`),
- la partie francophone de l'annuaire libre Open Directory (`http://dmoz.org/World/Fran%E7ais/`),
- Voila (`http://guide.voila.fr`),
- Yahoo France (`http://fr.yahoo.com`),

ainsi qu'un annuaire spécialisé dans les sites personnels, Voila Pages Perso (`http://annuaire-pp.voila.fr/Nav/nav`). Il semblait en effet intéressant de disposer d'informations sur ce type particulier de sites, qui sont souvent peu indexés par les annuaires généralistes, alors qu'ils représentent une part non négligeable des URL visitées par les panélistes : ainsi sur les 6,8 millions d'URL distinctes vues par le panel SensNet en 2002, près de 385 000 renvoient à des sites personnels. Lors de la deuxième collecte d'informations, en février 2002, nous avons actualisé les données recueillies en 2001 ; constatant, sur la base des données de février 2001, que les annuaires se recoupent peu entre eux, nous y avons adjoint deux annuaires généralistes, Lycos France (`http://www.lycos.fr/dir/`) et Looksmart France (`http://www.looksmart.fr/explore.jsp?lan=fr&path=176866,182901`). De la sorte, nous couvrons les sept grands annuaires généralistes les plus importants et les plus populaires du Web francophone ainsi que le plus important annuaire de sites personnels, et nous pouvons ainsi espérer avoir une couverture satisfaisante des URL visitées par les internautes des différents panels dont nous disposons.

Dans les deux campagnes de collecte de données, nous avons « aspiré » les annuaires pour en extraire les informations sur leur structure (arbre des catégories et renvois) et les sites qu'ils indexent (URL, titre, description). À l'exception d'Open Directory, nous avons développé pour chaque annuaire un logiciel adapté capable de reconnaître sélectivement, dans chaque page affichant le contenu d'une catégorie, les liens vers les sous-catégories et les URL de sites indexés. Pour cela, nous avons conçu pour chaque annuaire et chaque millésime de l'annuaire un aspirateur de pages capable d'extraire ces informations sur la base d'expressions régulières spécifiques ; les données ainsi extraites ont été formatées et stockées sous forme de base de données (voir Figure 3.11).

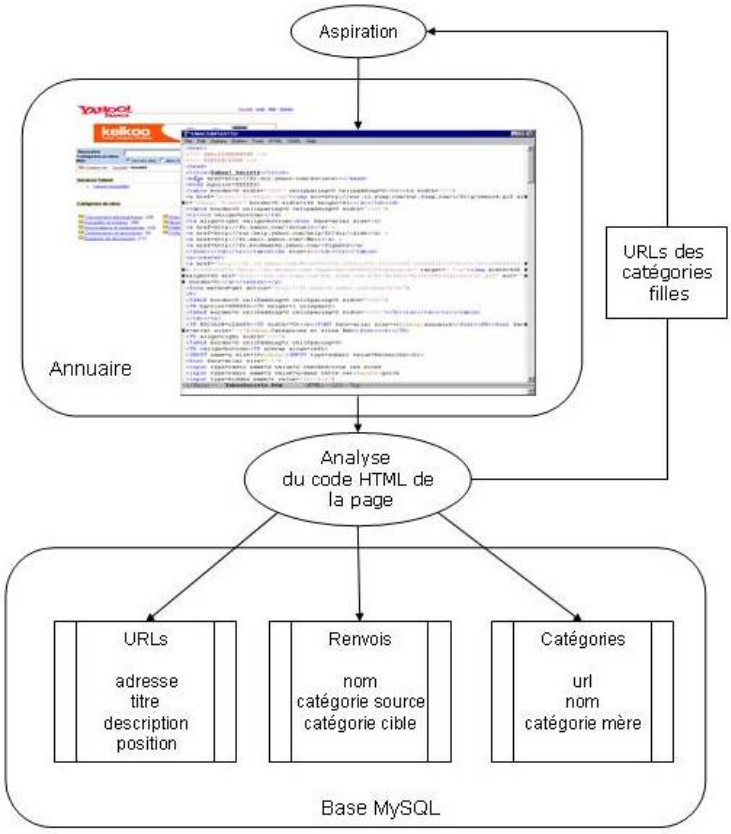


Figure 3.11. Aspiration des annuaires : fonctionnement général

Dans le cas d'Open Directory, annuaire « libre » dont la structure et le contenu sont téléchargeables sous forme de deux fichiers formatés en RDF, nous avons récupéré ces fichiers ; nous avons ensuite écrit un programme de transcription de ce format vers notre schéma de stockage et de sélection de la partie francophone d'Open Directory.

> ***Synthèse.*** *Les annuaires Web fournissent une description à vocation universelle des sites de référence sur la Toile. Il s'agit de descriptions textuelles structurées sous forme d'arbre en domaines et sous-domaines, et adaptées à la spécificité des contenus du Web, qui permettent d'envisager leur mobilisation systématique pour la description du contenu des parcours. Pour cela, nous exploitons sept annuaires généralistes et un annuaire des pages personnelles.*

3.3.2 Des différences de taille et de structure

Structurellement, chaque annuaire est une combinaison des trois éléments que nous avons présentés : multi-indexation, position des URL et utilisation des renvois. Le Tableau 3.15 ci-dessous présente de manière synthétique ces éléments pour les

huit annuaires étudiés en 2002 ; il est à noter qu'aucun des annuaires étudiés en 2001 n'avait modifié ces éléments structurels en 2002.

Tableau 3.15. Description structurelle des annuaires

	Multi-indexation	Les URL ne sont indexées que dans les catégories terminales	Utilisation de renvois
Looksmart	✔	✖	✖
Lycos	✔	✖	✔
MSN	✔	✖	✖
Nomade	✔	✖	✔
Open Directory	✖	✖	✔
Voila	✔	✔	✔
Voila Pages Perso	✔	✖	✖
Yahoo	✔	✖	✔

À ces différences de structure, s'ajoute une spécificité de conception pour Voila Pages Perso : cet annuaire est géré de manière complètement automatique par « auto-inscription »[1], tandis que pour les autres, chaque soumission de site par son concepteur est examinée manuellement par l'équipe éditoriale de l'annuaire ; en cas d'acceptation de la soumission, le site est inséré dans l'arborescence de l'annuaire accompagné d'un descriptif, selon des règles propres à chaque annuaire.

Multi-indexation des sites

La multi-indexation des sites dans les annuaires représente un avantage non négligeable pour l'utilisateur. En effet, le fait de pouvoir atteindre le même site en empruntant des chemins différents dans l'annuaire permet à l'utilisateur de s'affranchir d'un point de vue particulier (et unique, celui du documentaliste ayant classé le site en question) pour atteindre l'information recherchée. Ainsi, dans l'exemple du site « Hackers, bons ou mauvais élèves de l'Internet ? » cité ci-dessus, un premier chemin permet d'atteindre le site selon une classification thématique (point d'entrée : « Informatique et Internet ») alors qu'un deuxième chemin permet de l'atteindre selon un point de vue de localisation géographique (point d'entrée : « Exploration géographique »).

Dans ce cadre, la description d'un annuaire du point de vue du nombre d'URL indexées doit tenir compte de la multi-indexation : si un annuaire peut en effet faire figurer la même URL à plusieurs endroits, il présentera à l'utilisateur plus d'adresses qu'il n'en indexe effectivement, c'est pourquoi il est important de distinguer le nombre d'URL présentées du nombre d'URL uniques indexées. Yahoo France présente ainsi un plus grand nombre d'URL aux internautes que Nomade, mais il contient moins d'URL uniques que celui-ci (voir Tableau 3.16) ; Looksmart est quant

[1] Voir http://annuaire-pp.voila.fr/info pour une description du fonctionnement de Voila Pages Perso.

à lui l'annuaire utilisant le plus la multi-indexation, puisqu'une URL y figure en moyenne plus de 9 fois : ceci est dû au fait que, n'utilisant pas les renvois, Looksmart duplique des pans entiers de son annuaire, ce qui explique sa taille en nombre d'URL présentées comme en nombre de catégories. Cela étant, Looksmart s'impose comme l'annuaire le plus important en nombre d'URL uniques, avec plus de 160 000 adresses répertoriées.

Tableau 3.16. Nombre d'URL indexées et multi-indexation en février 2002

	Nombre total d'URL présentées	Nombre d'URL uniques	Taux de répétition moyen des URL
Looksmart	**1 552 553**	**162 730**	**9,54**
Lycos	75 401	67 168	1,12
MSN	137 097	76 773	1,78
Nomade	**179 575**	**143 461**	**1,25**
Open Directory	32 496	32 496	1
Voila	202 269	62 467	3,24
Voila PP	67 447	39 690	1,70
Yahoo	**238 873**	**130 393**	**1,83**

Les annuaires ont connu des taux de croissance très divers entre 2001 et 2002 (voir Tableau 3.17) : si Open Directory, Nomade et Voila n'ont presque pas changé de taille, MSN, Yahoo et Voila Pages Perso ont sensiblement augmenté leur nombre d'URL indexées. La part des URL indexées en 2001 encore présente dans l'annuaire l'année suivante nous renseigne sur l'effort consacré à la mise à jour : MSN a ainsi supprimé 44 % de ses adresses de 2001, tandis que Yahoo n'en a supprimé que 14 %.

Tableau 3.17. Nombre d'URL uniques en 2001 et évolution en 2002

	Nombre d'URL uniques en 2001	Taux de répétition moyen des URL en 2001	Évolution du nombre d'URL 2001-2002	Part des URL de 2001 présentes en 2002
MSN	46 137	1,35	**+ 66,4 %**	**56,5 %**
Nomade	138 832	1,32	+ 3,3 %	71,9 %
Open Directory	32 496	1	pas d'évolution	100,0 %
Voila	59 744	2,25	**+ 4,5 %**	**72,1 %**
Voila PP	27 923	1,81	+ 42,1 %	58,0 %
Yahoo	106 832	1,8	**+ 22,0 %**	**86,5 %**

Profondeur des annuaires

Les annuaires varient beaucoup en termes de profondeur, c'est-à-dire de nombre et de position des catégories dans l'arbre, le niveau de profondeur '1' étant l'entrée générale d'un annuaire, équivalente à sa page d'accueil. Une profondeur importante est le signe d'une division fine en domaines et sous-domaines, et garantit la précision des catégories de l'annuaire ; ceci assure à l'utilisateur de trouver ce qu'il recherche avec précision, mais au prix d'un nombre important de « clics » pour arriver à la catégorie qui l'intéresse. À l'inverse, un annuaire peu profond propose des catégories plus grossières, au contenu plus hétérogène, mais l'utilisateur parviendra plus rapidement à la catégorie pertinente pour sa recherche. Entre ces deux extrêmes, les

annuaires tentent de trouver un compromis acceptable entre navigabilité et finesse des catégories.

Looksmart et Yahoo sont les annuaires les plus profonds, avec une profondeur moyenne de 8,1 et 7,6, tandis que Voila Pages Perso, le plus petit de tous, a une profondeur maximale de 5 (voir Tableau 3.18). On note cependant que la profondeur de l'annuaire n'est pas liée au nombre d'URL qu'il présente : Lycos, Nomade, Open Directory et Voila, qui ont les mêmes profondeurs maximales et des niveaux moyens de présentation des URL assez semblables, présentent respectivement 75 000, 180 000, 32 000 et 202 000 URL.

Tableau 3.18. Profondeur des annuaires en 2002

	Nombre de catégories	Profondeur maximum	Profondeur moyenne	Niveau moyen des URL présentées
Looksmart	122 576	17	8,10	8,04
Lycos	7 100	9	4,73	4,51
MSN	15 955	7	4,42	4,19
Nomade	12 318	9	4,96	4,88
Open Directory	5 243	10	5,07	4,36
Voila	12 245	9	4,67	4,66
Voila PP	636	5	2,99	2,70
Yahoo	58 362	16	7,61	6,70

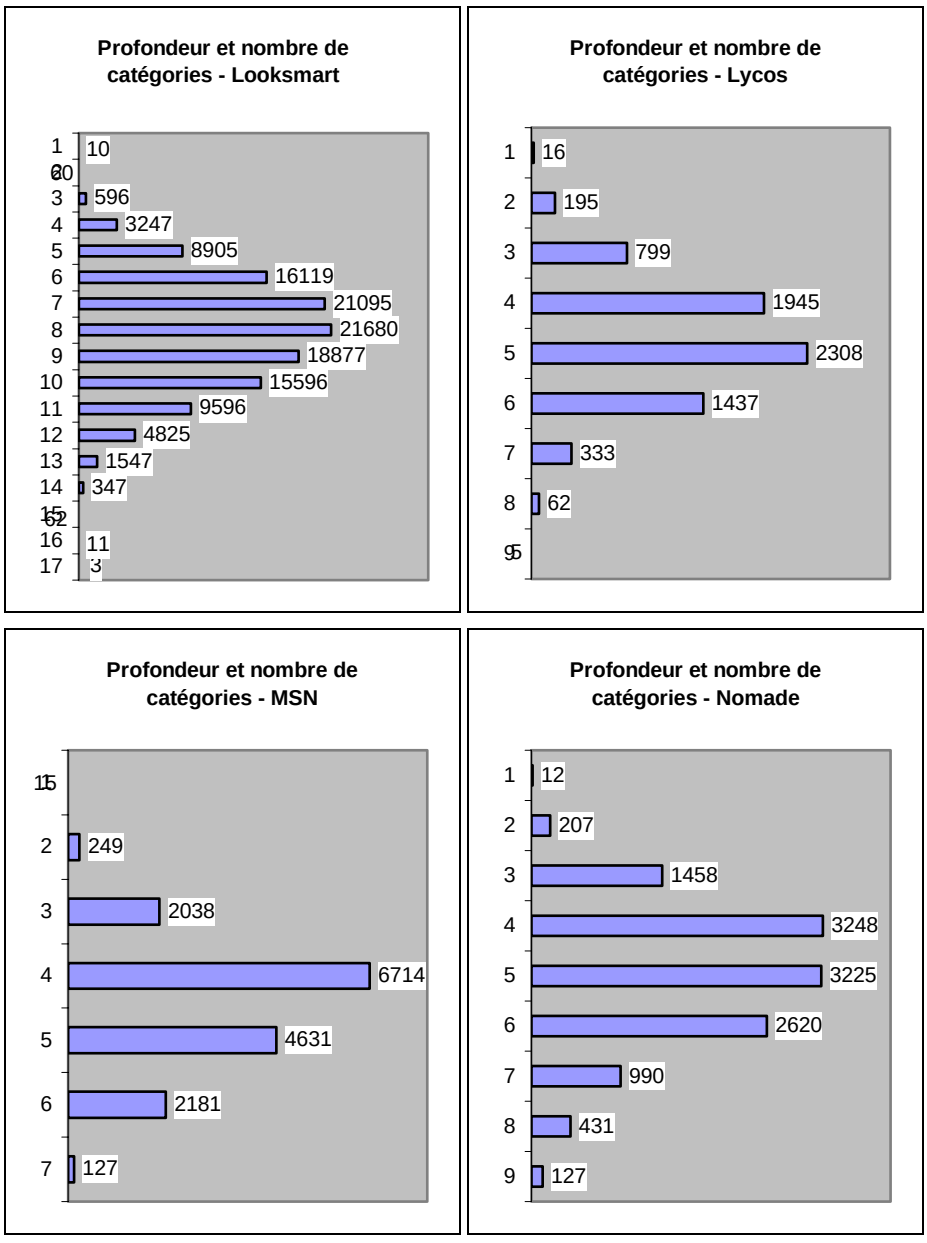
La profondeur d'un annuaire n'est donc pas directement liée au fait d'avoir un nombre important d'URL à présenter, mais semble plutôt résulter d'un choix organisationnel. Cette hypothèse est confirmée par l'examen du nombre moyen d'URL indexées par catégorie comportant au moins une URL (Tableau 3.19) : tandis que Nomade et Voila proposent en moyenne près de 17 URL par catégorie contenant au moins une URL, Lycos, Open Directory, Yahoo et MSN en offrent entre 5 et 10 en moyenne, et Voila Pages Perso près de 112.

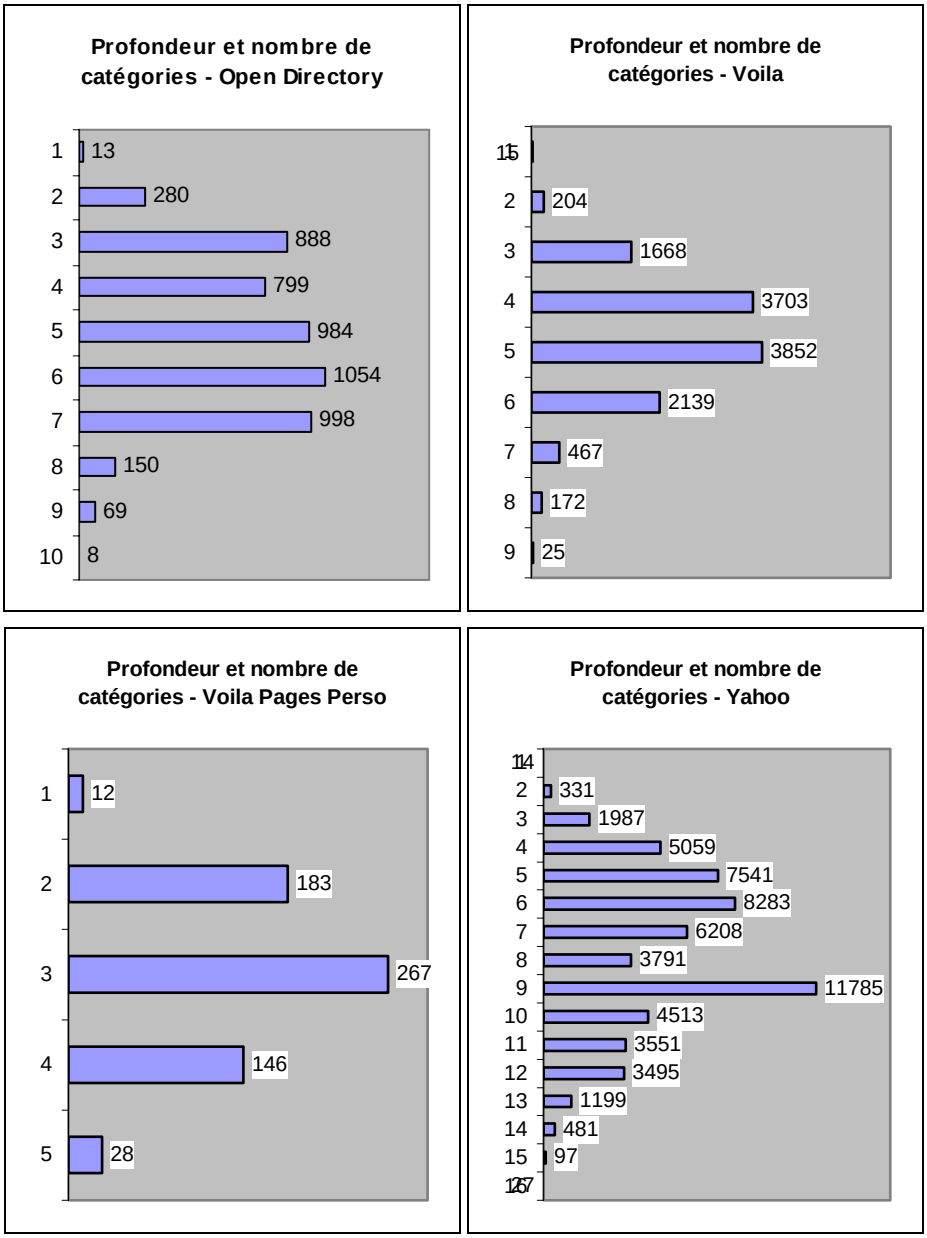
Tableau 3.19. Indexation des URL en 2002

	Nombre total de catégories	Les URL ne sont indexées que dans les catégories terminales	Nombre de catégories comportant une ou plusieurs URL	Nombre moyen d'URL par catégorie comportant au moins une URL
Looksmart	122 576	non	115 780	13,41
Lycos	7 100	non	6 672	11,30
MSN	15 955	non	15 203	9,02
Nomade	12 318	non	11 930	15,05
Open Directory	5 243	oui	4 201	7,73
Voila	12 245	oui	10 823	18,69
Voila PP	636	non	602	112,04
Yahoo	58 362	non	47 657	5,01

En outre, la répartition des catégories dans la hiérarchie des annuaires montre des structures variables ; les graphiques ci-dessous représentent le nombre de catégories

présentes à chaque niveau de profondeur de l'annuaire en 2002. Cette représentation « profilée » des annuaires permet de voir l'homogénéité de la répartition des catégories dans l'arbre : ainsi, Nomade et Open Directory ont la majorité de leurs catégories terminales aux niveaux 4 à 6 et 3 à 7, et sont plus « minces » avant et après. Yahoo, au contraire, connaît une expansion aux niveaux 5 et 6, puis un rétrécissement aux niveaux 7-8, suivi d'un très forte expansion au niveau 9 : on retrouve ici la particularité de Yahoo qui indexe la majorité de ses URL dans la catégorie « Exploration géographique », laquelle connaît là son expansion la plus forte, tandis que les autres catégories de premier niveau sont peu représentées.





Clef de lecture : dans Voila, on compte 3 703 catégories au 4e niveau de profondeur.

Figure 3.12 . Nombre de catégories pour chaque niveau de l'annuaire

Sous-catégories et renvois

Les renvois modifient beaucoup la physionomie de l'annuaire : ils facilitent la navigation pour l'utilisateur, et permettent, pour les créateurs des annuaires, de pallier la rigidité de l'organisation hiérarchique. En introduisant ces renvois, les

concepteurs des annuaires enrichissent les possibilités de navigation hypertextuelle au sein de l'annuaire.

Les cinq annuaires utilisant les renvois (Lycos, Nomade, Open Directory, Voila et Yahoo) n'en font pas le même usage (voir Tableau 3.20) : tandis que Nomade et Voila en font un emploi modéré (seul 1,6 % des catégories de Voila comportent un renvoi, proposant 1,4 renvois en moyenne), Lycos, Open Directory et Yahoo y font massivement appel : ce dispositif concerne près de 20 % des catégories de Yahoo, lesquelles comportent près de 4 renvois en moyenne.

Tableau 3.20. Utilisation des renvois

	Nombre total de catégories	Nombre de catégories avec renvoi	Part des catégories avec renvoi	Nombre total de renvois	Nombre moyen de renvois par catégorie avec renvoi
Looksmart	122 576	-	-	-	-
Lycos	7 100	1 058	14,9 %	3 421	3,23
MSN	15 955	-	-	-	-
Nomade	12 318	666	5,4 %	1 084	1,63
Open Dir.	5 243	527	10,0 %	1 829	3,47
Voila	12 245	215	1,7 %	354	1,65
Voila PP	636	-	-	-	-
Yahoo	58 362	12 847	22,0 %	48 001	3,74

Les renvois rendent les annuaires plus navigables, permettant de passer facilement d'une catégorie à une autre : pour Yahoo, on constate que l'ajout des renvois fait passer de 15 900 à 23 300 le nombre de catégories permettant d'accéder à une autre catégorie, en suivant soit le lien hiérarchique (catégorie-fille), soit le lien de renvoi ; le nombre moyen de liens vers d'autres catégories passe alors de 3,7 à 4,6. Voila et Nomade, au contraire, utilisent très peu les renvois. Looksmart développe une toute autre stratégie, consistant à copier des parties entières de son annuaire à plusieurs endroits, ce qui explique son nombre très élevé de catégories ainsi que le fort taux de répétition des URL.

Des principes organisationnels variés

Nous nous sommes intéressés aux principes qui gouvernent l'organisation et la structuration des annuaires. Il existe plusieurs modèles d'organisation de l'information et des connaissances, qui proviennent de domaines aussi variés que la représentation des connaissances en Intelligence Artificielle, de la construction de thésaurus en documentation et en sciences de l'information, ou de la constitution de répertoires et autres annuaires pratiques (pages jaunes, annuaires professionnels, etc.). Nous pouvons distinguer trois modes d'organisation prototypiques :

- Catégorisation systématique de domaines des activités humaines, des objets de la vie quotidienne, etc. dans une approche de type ontologique. C'est l'approche classique en intelligence artificielle et en documentation (sciences de l'information).

- Catalogage moins systématique, plus pratique, centré sur les activités humaines (activités marchandes, loisirs, formes diverses de sociabilité, etc.), dans une approche du type « pages jaunes » ou annuaire professionnel.
- Catégorisation du « monde de l'Internet » : cartographie des sites et des services disponibles sur Internet, sans avoir de critères précis pour la classification et la catégorisation des objets du monde, des activités humaines, etc. Cette approche a été spontanément mise en œuvre sur différents portails pour organiser l'information selon des catégories propres à Internet (exemples : *chat*, achat en ligne, …).

Ces différents modèles ont été adoptés, de manière plus ou moins consciente et revendiquée, par les annuaires du Web : aucun de ceux que nous avons étudiés ne correspond strictement à l'une ou l'autre de ces catégories et ils s'avèrent assez différents des objets classificatoires habituels (ontologies et thésaurus en particulier).

Tableau 3.21. Répartition dans les catégories de premier niveau des URL présentées, et correspondance entre catégories : Voila et Yahoo

Voila			Yahoo	
38,4 %	Villes, régions, pays	}	Exploration géographique	47,4 %
4,2 %	Tourisme, voyages			
7,2 %	Business, économie	⇔	Commerce et économie	21,7 %
7,9 %	Arts, culture	⇔	Art et culture	8,9 %
5,5 %	Loisirs, sorties	}	Sports et loisirs	5,1 %
5,0 %	Sport, plein air			
5,2 %	Informatique, internet	⇔	Informatique et Internet	2,0 %
3,0 %	Enseignement	⇔	Enseignement et formation	0,7 %
1,8 %	Administrations, politique	⇔	Institutions et politique	0,2 %
1,8 %	Sciences, recherche	⇔	Sciences et technologies	2,8 %
1,8 %	Sujets de société	⇔	Société	5,4 %
1,6 %	Santé, médecine	⇔	Santé	1,2 %
1,4 %	Actualités, médias	⇔	Actualités et médias	2,0 %
			Sciences humaines	1,6 %
			Divertissement	0,8 %
			Références et annuaires	0,2 %
13,7 %	Achats, vie pratique			
1,4 %	Emploi			

Clef de lecture : dans Voila, la catégorie de premier niveau « Achats, vie pratique » contient 13,7 % des URL présentées par cet annuaire, et n'a pas d'équivalent au premier niveau de Yahoo.

À titre d'exemple, l'examen des annuaires Yahoo et Voila révèle des modes d'organisation bien différenciés (voir Tableau 3.21). Yahoo a une approche de classification systématique, révélée par un grand nombre de catégories (58 000 contre 12 000 pour Voila), organisées dans un arbre ayant 16 niveaux de profondeur (contre 9 niveaux pour Voila). Yahoo présente également un réseau très dense formé par un système de renvois entre catégories (48 000 renvois, contre 350 dans Voila). Les catégories de premier niveau les plus importantes dans Yahoo sont « Exploration géographique » et « Commerce et économie », ce qui indique une démarche de classification systématique ; en effet, Yahoo classe de manière privilégiée un site dans

l'une ou l'autre de ces deux grandes catégories, si d'autres classements thématiques sont pertinents pour ce site, le mécanisme des renvois est alors mis en œuvre pour rendre compte de cette multi-classification.

Le côté encyclopédique de Yahoo se manifeste également par la présence de catégories telles que « Sciences humaines » dès le premier niveau. À l'opposé, Voila présente une approche pragmatique, centrée sur les services liés aux différentes activités humaines : activités économiques et sociales, sans oublier les loisirs. Le côté pratique de Voila est manifeste si l'on examine les catégories de premier niveau : nous relevons notamment la présence d'une catégorie « Achat, vie pratique », représentant 13,7 % des sites indexés, qui n'a pas d'équivalent au premier niveau chez Yahoo.

Cette diversité des principes d'organisation des annuaires a déjà été mise en évidence par Van der Walt[1] : pour passer d'une catégorie à ses sous-catégories, un annuaire peut mettre en œuvre simultanément des principes très différents (lien générique-spécifique, lien partie-tout, liste alphabétique, etc.). De fait, les annuaires ne suivent pas rigoureusement les principes issus des disciplines classificatoires telles que les sciences de l'information et de la documentation ou la représentation des connaissances en intelligence artificielle, et leurs principes organisationnels traduisent les contraintes qui ont régi leur mise en place dans un contexte de croissance rapide d'Internet et avec l'obligation d'assurer une large couverture thématique.

Cela étant, les principes de structuration dépendant des tâches et des profils d'usage, il n'est pas évident qu'un principe universel d'organisation puisse répondre à tous les besoins des internautes. Les principes de type thésaurus ont été développés dans un contexte très particulier, celui des bibliothèques, et à destination de publics bien définis (élèves, étudiants, enseignants, chercheurs). Sur Internet, les contenus accessibles sont de nature différente de ceux des bibliothèques, les tâches et les profils des utilisateurs sont très variés, de sorte que les modes d'accès à l'information structurée (sous forme d'annuaire de sites ou sous une autre forme d'ailleurs) devraient tenir compte de cette grande diversité[2].

Les annuaires se recoupent peu

L'ensemble des huit annuaires étudiés comporte près de 421 000 sites ou pages uniques indexés. Nous avons constaté que les annuaires se recoupent peu de manière générale : si l'on exclut Voila Pages Persos pour ne considérer que les sept annuaires généralistes, ceux-ci ont seulement 1 806 URL en commun (0,5 % de l'ensemble), tandis que 62,7 % de l'ensemble des URL indexées ne le sont que par un seul des sept annuaires.

---

[1] Voir [Van der Walt 1998].

[2] Pour une réflexion poussée sur cette question, voir [Assadi & Beauvisage 2002].

Tableau 3.22. Part des URL d'un annuaire A également présentes dans l'annuaire B

↓partage n % de ses URL avec →	Looksmart	Lycos	MSN	Nomade	Open Directory	Voila	Voila PP	Yahoo
Looksmart	100 %	18,6 %	16,1 %	31,1 %	7,0 %	18,3 %	2,4 %	33,5 %
Lycos	45,8 %	100 %	27,4 %	44,5 %	11,5 %	28,1 %	3,1 %	43,2 %
MSN	33,9 %	23,5 %	100 %	34,4 %	11,3 %	24,1 %	1,2 %	37,0 %
Nomade	35,2 %	20,4 %	18,4 %	100 %	8,7 %	21,0 %	2,8 %	32,3 %
Open Directory	36,1 %	24,3 %	27,8 %	40,0 %	100 %	25,1 %	2,0 %	35,1 %
Voila	47,6 %	29,7 %	29,7 %	48,2 %	12,6 %	100 %	3,3 %	42,1 %
Voila PP	10,0 %	5,2 %	2,4 %	10,0 %	1,6 %	5,2 %	100 %	6,9 %
Yahoo	41,7 %	21,9 %	21,8 %	35,5 %	8,5 %	20,2 %	2,1 %	100 %

Clef de lecture : 35,2 % des URL de Nomade sont également indexées par Looksmart, tandis que 31,1 % des URL de Looksmart sont dans la base de Nomade.

Chaque annuaire a donc ses spécificités, ce que vient confirmer l'examen des taux de recouvrement entre annuaires deux à deux[1] (voir Tableau 3.22) : de manière générale, le taux de recouvrement moyen entre les différents annuaires est de 22 %, et de 24,3 % si l'on exclut le très spécifique Voila Pages Perso. Dans le détail, nous notons en premier lieu que la spécificité de l'annuaire de sites personnels Voila Pages Perso est éminemment confirmée par les très faibles taux de recouvrement avec les autres annuaires, en particulier dans le sens *VoilaPP* ➔ *autres annuaires* (au maximum 10 % des URL de Voila Pages Perso sont indexées par un autre annuaire), alors même que Voila Pages Perso est le plus petit annuaire de tous.

D'autre part, la taille des annuaires ne semble pas être le facteur déterminant de leurs recoupements : entre les trois plus grands annuaires Looksmart, Nomade, Yahoo, le taux de recouvrement deux à deux varie de 30 à 40 %, tandis que les petits annuaires ne sont pas « inclus » dans les grands. Ainsi, Open Directory, de taille modeste, partage en moyenne moins d'un tiers de ses URL avec d'autres annuaires, pourtant jusqu'à quatre fois plus gros que lui, soit autant que Looksmart, Nomade et Yahoo entre eux. Il apparaît donc que chaque annuaire indexe des sites qui lui sont spécifiques. Ceci est confirmé par l'examen, pour chaque annuaire, de la proportion d'URL qu'il est le seul à indexer (Tableau 3.23).

[1] Les annuaires étant de tailles différentes, le calcul des recoupements deux à deux entre annuaires est dissymétrique, et doit être analysé pour chaque couple d'annuaires.

Tableau 3.23. Part des sites indexés spécifiques à chaque annuaire

Annuaire	Nombre d'URL indexées	Nombre d'URL spécifiques de l'annuaire	Part des URL spécifiques
Looksmart	161 974	70 058	43,2 %
Lycos	65 866	16 241	24,7 %
MSN	76 712	30 862	40,2 %
Nomade	143 274	55 122	38,5 %
Open Directory	31 308	10 629	33,9 %
Voila	62 411	14 261	22,8 %
Voila PP	39 417	31 384	79,6 %
Yahoo	130 101	43 525	33,4 %

À l'exception de Voila Pages Perso, dont le contenu est particulier (près de 80 % d'URL spécifiques), on constate ici que Looksmart, le plus gros des annuaires, est en même temps celui dont la spécificité est la plus importante (43,2 %), résultat que nous pouvions prévoir. Moins attendu est le taux de spécificité de MSN (40,2 % d'URL spécifiques), pourtant deux fois et demie plus petit que Looksmart, et de Yahoo (33,4 %), ce dernier étant relativement peu spécifique étant donné sa taille. Il y a donc un double effet participant à la spécificité des annuaires : leur taille, qui augmente statistiquement leur chance d'indexer des sites que les autres n'ont pas, mais aussi leur positionnement éditorial, à travers le choix des sites indexés.

Des choix éditoriaux marqués

Cette notion de positionnement éditorial des annuaires se vérifie lorsque l'on examine le nombre d'URL présentées dans chaque catégorie de premier niveau des annuaires : des préférences thématiques apparaissent alors (voir Tableau 3.25 à Tableau 3.31). On voit ici nettement l'utilité rendue par l'utilisation des renvois : le calcul effectué sans suivre les renvois rend plutôt compte des choix structurels d'organisation des annuaires. Les catégories de premier niveau sont alors plutôt déséquilibrées : « Économie, Entreprise » occupe 33,4 % dans Lycos, « Espace B to B » représente 28 % de Nomade, « Villes, régions, pays » 28 % de Voila, on retrouve des logiques de classement par géolocalisation ou par domaines de métier. Le suivi des renvois rééquilibre profondément les répartitions d'URL présentées : pour Nomade, la catégorie « Mes Courses » passe en première position avec 16 % des URL présentées ; dans Voila, c'est la catégorie « Achat, vie pratique » qui domine alors avec 14 % des URL présentées. Les profils affichés sont alors beaucoup plus lisses et diversifiés ; trois groupes émergent toutefois :

- orientation « vie pratique » et services hors du Web : Looksmart (« Maison et Loisirs » à 29 %), Nomade (« Mes Courses », « Espace B to B », « Culture et loisirs » et « Société, vie pratique » totalisent 58 %), Voila (catégories « Achat, vie pratique » et « Villes, régions, pays ») ;
- classement tourné vers le monde de la sphère économique : MSN (« Entreprises » à 20 %), Open Directory (« Commerce et économie » représente 38 %), Yahoo (« Commerce et économie » à 16 %) ;
- annuaire tourné vers le monde de l'Internet : Lycos (« Informatique, Multimédia » à 14 %).

Tableau 3.24. Looksmart : répartition dans les catégories de premier niveau des URL présentées

Maison et loisirs	28,9 %
Société et politique	19,7 %
Économie et finance	13,4 %
Éducation et emploi	10,0 %
Arts et divertissements	8,5 %
Tourisme et voyages	8,1 %
Santé et beauté	3,8 %
Sports et auto	2,8 %
Shopping	2,7 %
Informatique et Internet	2,0 %

Tableau 3.25. Lycos : répartition dans les catégories de premier niveau des URL présentées

Sans renvois	
Économie, Entreprise	33,4 %
Régional	10,1 %
Art, Culture	9,5 %
Sciences, Techniques	8,6 %
Emploi, Enseignement	7,8 %
Loisirs	5,1 %
Sciences humaines	4,8 %
Sports	4,6 %
Informatique, Multimédia	4,3 %
Institutions, Société	4,1 %
Actualité, Médias	3,2 %
Féminin	1,4 %
Jeux vidéo	1,4 %
Auto-moto	0,9 %
Voyage, Tourisme	0,6 %
Célébrités	0,1 %

Avec Renvois	
Informatique, Multimédia	13,6 %
Régional	11,8 %
Sciences humaines	11,0 %
Sports	10,6 %
Voyage, Tourisme	8,8 %
Actualité, Médias	8,7 %
Sciences, Techniques	8,4 %
Économie, Entreprise	6,5 %
Féminin	5,9 %
Auto-moto	3,4 %
Institutions, Société	3,2 %
Loisirs	2,8 %
Emploi, Enseignement	2,8 %
Art, Culture	1,3 %
Jeux vidéo	0,7 %
Célébrités	0,6 %

Tableau 3.26. MSN : répartition dans les catégories de niveau 1 des URL présentées

Entreprises	20,3 %
Voyages - Tourisme	16,1 %
Vie quotidienne - Société	11,5 %
Arts - Culture - Médias	10,6 %
Loisirs - Passions	10,5 %
Savoir - Éducation	6,1 %
Sports	5,1 %
Informatique - Internet	4,2 %
Infos - Météo	2,9 %
Sciences - Techniques	2,9 %
Finances - Bourse - Patrimoine	2,7 %
Jeux - Consoles	2,0 %
Santé	1,9 %
Emploi, formation	1,7 %
Shopping	1,4 %

Tableau 3.27. Nomade : répartition dans les catégories de niveau 1 des URL présentées

Sans renvois	
Espace B to B	28,4 %
Mes Courses	14,8 %
Voyage, géographie	11,7 %
Culture et loisirs	10,4 %
Sport et détente	8,2 %
Société, Vie pratique	8,1 %
Éducation, formation	4,5 %
Nature et sciences	3,7 %
Nouvelles technologies	2,9 %
Forme et Santé	2,9 %
Actu, médias	2,4 %
Sorties, spectacles	2,0 %

Avec renvois	
Mes Courses	16,3 %
Espace B to B	16,1 %
Culture et loisirs	13,6 %
Société, Vie pratique	11,6 %
Sorties, spectacles	11,0 %
Éducation, formation	7,9 %
Voyage, géographie	5,8 %
Nouvelles technologies	5,1 %
Sport et détente	5,1 %
Actu, médias	3,6 %
Nature et sciences	2,2 %
Forme et Santé	1,8 %

Tableau 3.28. Open Directory : répartition dans les catégories de niveau 1 des URL présentées

Sans renvois	
Régional	40,0 %
Commerce et économie	17,4 %
Arts	9,0 %
Sciences	6,9 %
Informatique	6,0 %
Divertissements	5,1 %
Société	4,3 %
Santé	3,2 %
Sports	2,9 %
Formation	2,2 %
Actualité	1,2 %
Maison	1,0 %
Références	0,9 %

Avec renvois	
Commerce et économie	37,8 %
Régional	34,2 %
Société	6,1 %
Informatique	4,6 %
Divertissements	3,9 %
Sciences	3,1 %
Arts	2,9 %
Santé	2,4 %
Formation	1,9 %
Sports	1,2 %
Actualité	1,1 %
Références	0,6 %
Maison	0,3 %

Tableau 3.29. Voila : répartition dans les catégories de niveau 1 des URL présentées

Sans Renvois	
Villes, régions, pays	38,4 %
Achats, vie pratique	13,7 %
Arts, culture	7,9 %
Business, Économie	7,2 %
Loisirs, sorties	5,5 %
Informatique, internet	5,2 %
Sport, plein air	5,0 %
Tourisme, voyages	4,2 %
Enseignement	3,0 %
Administrations, politique	1,8 %
Sujets de société	1,8 %
Sciences, recherche	1,8 %
Santé, médecine	1,6 %
Actualité, médias	1,4 %
Emploi	1,4 %

Avec Renvois	
Achats, vie pratique	14,6 %
Villes, régions, pays	14,1 %
Business, Economies	13,2 %
Enseignement	11,0 %
Sciences, recherche	9,0 %
Informatique, internet	8,2 %
Tourisme, voyages	7,4 %
Arts, culture	4,3 %
Sport, plein air	4,3 %
Sujets de société	3,9 %
Santé, médecine	2,6 %
Administrations, politique	2,4 %
Emploi	2,2 %
Loisirs, sorties	1,5 %
Actualité, médias	1,1 %

Tableau 3.30. Voila PP : répartition dans les catégories de niveau 1 des URL présentées

Loisirs, tourisme	17,6 %
Art, culture	13,9 %
Inform@tique	13,4 %
Régions, pays	9,7 %
Famille, communauté	8,8 %
Job, formation	7,2 %
Sport	7,1 %
Sciences, technos	5,3 %
Société	5,0 %
Vie pratique	4,4 %
Actu, média	4,3 %
Santé, médecine	3,3 %

Tableau 3.31. Yahoo : répartition dans les catégories de niveau 1 des URL présentées

Sans renvois	
Exploration géographique	47,4 %
Commerce et Économie	21,7 %
Art et culture	8,9 %
Société	5,4 %
Sports et loisirs	5,1 %
Sciences et technologies	2,8 %
Informatique et Internet	2,0 %
Actualités et médias	2,0 %
Sciences humaines	1,6 %
Santé	1,2 %
Divertissement	0,8 %
Enseignement et Formation	0,7 %
Institutions et politique	0,2 %
Références et annuaires	0,2 %

Avec renvois	
Exploration géographique	16,5 %
Commerce et Économie	15,1 %
Société	13,0 %
Sciences humaines	11,8 %
Sports et loisirs	7,7 %
Institutions et politique	7,2 %
Sciences et technologies	6,1 %
Divertissement	5,9 %
Art et culture	5,3 %
Actualités et médias	4,2 %
Informatique et Internet	2,8 %
Enseignement et Formation	2,1 %
Santé	1,3 %
Références et annuaires	1,0 %

Pour aller plus avant dans la notion de positionnement éditorial des annuaires, nous avons examiné les différences de contenu entre annuaires pour un thème donné, par exemple l'art, l'économie ou la politique. Pour cela, nous avons qualifié le contenu des annuaires à partir des titres et des descriptifs qu'ils donnent des sites indexés sous un thème donné.

Nous avons d'abord choisi des catégories générales présentes au premier ou deuxième niveau pour les huit annuaires étudiés et *a priori* équivalentes entre annuaires. Nous avons extrait pour chaque annuaire et pour l'ensemble des sites classés sous la catégorie choisie, les titres et descriptifs associés par l'annuaire à ces sites ; le corpus ainsi constitué a été traité avec un outil d'analyse de données textuelles (le logiciel *Alceste*[1]). Cet outil nous a permis d'identifier le vocabulaire spécifique à chaque annuaire en ce qui concerne la description des sites du thème traité. Nous sommes ainsi en mesure de dégager des « profils thématiques » de chaque annuaire.

1 Voir [Reinert 1993].

Une première étude a été consacrée au thème « Art et culture », et une deuxième à la catégorie « Commerce et économie », dont nous présentons ici les résultats[1]. L'examen du vocabulaire spécifique de chaque annuaire montre une orientation très forte de Looksmart vers l'immobilier (vocabulaire spécifique : `immeuble`, `locatif`, `résidentiel`, `maison`, `banlieue`, `annonce`, …) ; Nomade affiche un profil assez généraliste, avec une orientation marquée vers l'offre de services informatiques (`solution`, `conception`, `informatique`, `internet`, `intranet`, `logiciel`, `hébergement`, …), tandis que Lycos présente une forte spécialisation dans le tourisme (`gîte`, `hôtel`, `tourisme`, `camping`, `visiter`, `restaurant`, `réservation`, …), et Voila dans l'achat en ligne et les services bancaires et financiers (`télécommerce`, `paiement sécurisé`, et `banque`, `boursier`, `financier`, `crédit`, `chèque`, …). MSN met en avant un classement géographique en privilégiant des sites nord-américains et francophones (`Amérique`, `Canada`, `canadien`, `Québec`, `Suisse`, `Bruxelles`, …). Enfin, Yahoo et Open Directory affichent tous deux un positionnement assez diversifié, qui semble refléter un classement par corps de métier.

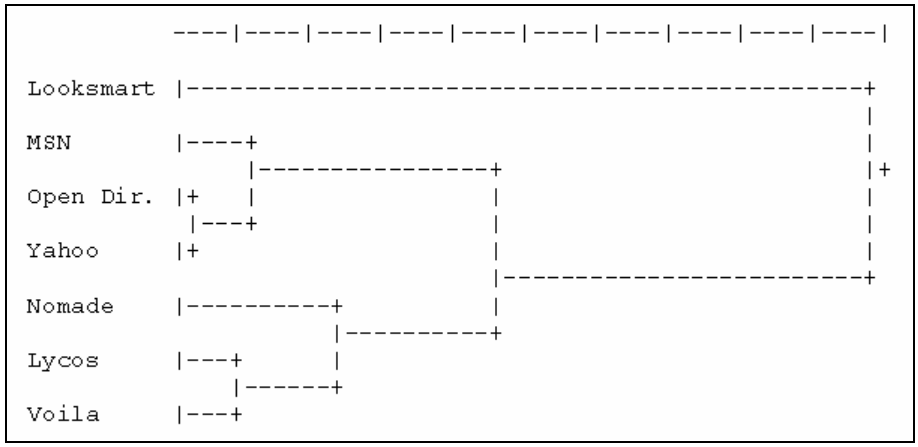


Figure 3.13. Classification des sept annuaires généralistes en 2002 sur la base des descriptifs des sites de la catégorie « Commerce et économie »

La classification des annuaires sur cette base (voir Figure 3.13) oppose le très spécifique Looksmart à l'ensemble des autres annuaires, lesquels se répartissent en deux groupes : le premier semble privilégier l'offre de services en ligne (bancaires et financiers pour Voila, touristiques pour Lycos, informatique pour les entreprises en ce qui concerne Nomade), tandis que le second paraît s'orienter vers un présentation plus large incluant l'ensemble des métiers et des activités économiques (MSN, Open Directory, Yahoo).

[1] L'annuaire Voila Pages Perso ne couvre pas le thème « Commerce et économie », c'est pour cela qu'il est absent de cette étude.

Les annuaires ont des styles différents

Chaque annuaire a une manière spécifique de présenter les sites qu'il indexe. À titre d'exemple, les descriptions du site « Bandit Mania » (www.banditmania.com), répertorié par les 8 annuaires étudiés, sont :

Annuaire	Titre	Description
Looksmart	Banditmania - Portail de la moto	Ce Repaire des motards contient plus de 2 000 pages et 1 800 photos. Dossiers, reportages, essais de motos et d'accessoires, conseils, annonces.
Lycos	Banditmania	Site non officiel de la Suzuki GSF Bandit. Caractéristiques, infos et actualité de la moto.
Nomade	Banditmania: le repère des motards	Banditmania est entièrement consacré à la moto et aux roadsters: mécanique, caractéristiques et technique, chiffres et données brutes, sons et vidéos, manuel en ligne, conseils pour le pilote, opinions, forum technique, guide moto, etc.
MSN	Bandit Mania	Bandit Mania, guide multithématique et conseils pour motards.
Open Directory	BanditMania : le site non-officiel de la Suzuki Bandit	Plusieurs centaines de pages de technique, conseils, opinions et informations illustrées par un millier de photos sur la moto et plus spécifiquement la Suzuki GSF Bandit dans toutes ses cylindrées : 250, 400, 600, 750 et 1200 cm3.
Voila	Banditmania	Webzine sur les motos - L'actualité moto (toutes marques), des dossiers, des reportages, des essais de motos, une lettre d'information gratuite et des services gratuits (petites annonces, moto puces, avis de recherche, achats groupés, etc.).
Voila Pages Persos	BanditMania : le site moto non-officiel de la Suzuki Bandit	200 pages de technique, de conseils et d'infos motos illustrées par 700 photos sur le roadster phare de Suzuki dans toutes ses cylindrées : Bandit GSF 250, 400, 600, 750 et 1200 cm3. Une large part du site est consacrée à la moto en général avec le guide du motard et les informations indispensables : assurances, pilotage, circuits, bons plans, événements, aventures, humour, adresses pour tous les motards.
Yahoo	Banditmania	Actualités, dossiers, reportages, essais et mécanique.

Les variations entre descriptions de sites d'un annuaire à l'autre sont de plusieurs ordres, et concernent en premier lieu leur longueur : MSN propose les descriptifs les plus courts, avec près de 9 mots en moyenne, tandis que ceux de Nomade et de Voila sont trois fois plus longs (voir Tableau 3.32).

Tableau 3.32. Longueur des descriptifs de sites

	Nombre moyen de mots dans le titre	Nombre moyen de mots dans le descriptif
Tous Annuaires	3,5	19,1
Looksmart	6,2	21,4
Lycos	3,8	19,9
MSN	2,6	9,3
Nomade	3,0	28,5
Open Directory	3,2	15,2
Voila	2,8	29,3
Voila PP	3,7	18,5
Yahoo	3,1	10,4

À la longueur variable des descriptifs, correspond un style particulier à chaque annuaire : le fait de proposer un résumé concis des sites indexés se traduit souvent par un style « télégraphique », où les phrases sont essentiellement nominales et la parataxe l'emporte sur la syntaxe. Ces différences sont perceptibles à travers la répartition des catégories morpho-syntaxiques utilisées dans les descriptifs de sites.

L'analyse de la répartition des catégories grammaticales majeures (verbes, adverbes, noms, adjectifs) pour chaque annuaire fait apparaître une opposition forte entre Yahoo et MSN d'un côté, et Looksmart et Nomade de l'autre (voir Tableau 3.33) : chez les premiers, noms et adjectifs sont sur-représentés, marque d'un style haché et « télégraphique » ; dans les seconds, au contraire, les descriptifs sont beaucoup plus « verbalisés », ce que traduit la présence forte de verbes et d'adverbes.

Ces observations ajoutées à celles sur la longueur des descriptifs laissent penser que si certains annuaires comme Looksmart et Yahoo sont peu loquaces, la quantité d'information qu'ils délivrent sur les sites n'est pas proportionnelle à la longueur de leurs descriptifs, car les tournures phrastiques d'ordre présentationnel (comme « Vous trouverez sur ce site » ou « Ce site vous propose ») comptent pour une bonne part dans la longueur des descriptifs de sites. De la sorte, si Looksmart ou Yahoo sont plus brefs dans leurs descriptifs que Nomade, ils n'en disent pas moins sur les sites, mais le disent différemment. C'est donc plus dans la façon de décrire que dans la précision de la description que les annuaires s'opposent, ce que traduit la répartition des personnes pronominales et verbales employées (Tableau 3.33) : nous voyons une opposition très nette, autour de l'emploi de la 2ème personne du pluriel, entre les annuaires qui présentent les sites en s'adressant directement au lecteur (Looksmart, Nomade, Voila, Voila Pages Persos) et ceux qui ne fournissent que des indications « neutres » à l'internaute (Yahoo, MSN, Open Directory). On note à cet égard, que ce sont les annuaires dont les descriptifs sont les plus longs (Nomade, Voila) qui ont le plus recours à l'emploi du « vous ».

Tableau 3.33. Distribution des principales catégories morphosyntaxiques et des personnes (pronoms et verbes) pour chaque annuaire

		Looksmart	Lycos	MSN	Nomade	Open Dir.	Voila	Voila PP	Yahoo
Catégories morpho-syntaxiques	Noms	64,8 %	72,4 %	**79,7 %**	66,4 %	73,8 %	73,9 %	70,1 %	**79,1 %**
	Verbes	**19,4 %**	10,3 %	**5,0 %**	**16,9 %**	8,4 %	9,9 %	13,2 %	**5,5 %**
	Adjectifs	12,4 %	13,7 %	14,0 %	12,4 %	15,1 %	13,3 %	12,2 %	14,2 %
	Adverbes	3,5 %	3,6 %	1,4 %	4,3 %	2,7 %	3,0 %	4,4 %	1,3 %
	Total	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Personnes (pronoms et verbes)	1re SG	0,1 %	0,3 %	3,9 %	0,3 %	0,9 %	0,2 %	9,3 %	0,9 %
	2e SG	2,4 %	4,7 %	11,0 %	0,7 %	9,6 %	2,6 %	4,0 %	27,4 %
	3e SG	50,0 %	58,3 %	76,8 %	50,9 %	60,6 %	54,9 %	46,8 %	57,0 %
	1re PL	0,1 %	0,6 %	1,1 %	0,9 %	2,9 %	0,2 %	3,9%	0,8 %
	2e PL	**34,6 %**	**23,4 %**	**2,0 %**	**40,3 %**	**15,5 %**	**33,6 %**	**29,2 %**	**3,3 %**
	3e PL	12,7 %	12,8 %	5,1 %	6,9 %	10,5 %	8,5 %	6,7 %	10,6 %
	Total	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

L'analyse morpho-syntaxique des descriptifs des sites et celle des pronoms convergent, et nous voyons deux logiques présentationnelles s'opposer : d'un côté, l'« annuaire-interlocuteur » qui entend guider l'internaute et servir d'intermédiaire entre lui et les sites (Looksmart, Voila, Voila Pages Perso, Nomade) ; de l'autre, l'« annuaire relais d'information » adoptant une posture d'intermédiation plus neutre (Lycos, Open Directory, MSN, Yahoo). C'est la position même de l'annuaire vis-à-vis de l'utilisateur qui est en jeu ici.

Ces renseignements sur les annuaires, leur taille et leur structure, leurs spécificités, leur positionnement et leur style, sont très importants pour l'utilisation que nous souhaitons en faire. Ils montrent d'une part qu'un annuaire Web est une structure mouvante, et que cette appellation recouvre des outils de classement très variés. D'autre part, les grandes différences structurelles, thématiques et stylistiques entre annuaires risquent fort de compliquer leur utilisation combinée pour décrire les parcours, et de nous obliger à n'en retenir que quelques-uns. Dans cette optique, les questions de volume et de spécialisation thématique nous aideront à sélectionner nos ressources parmi les annuaires qui décrivent bien les parcours.

> ***Synthèse.*** *L'analyse détaillée des huit annuaires étudiés montre leur grande hétérogénéité. Outre les différences de volume entre les sept annuaires généralistes, les choix structurels (multi-indexation, renvois), classificatoires, thématiques et stylistiques dénotent des stratégies différentes et une spécialisation de chacun d'eux. La disparités des descriptifs textuels invite à laisser de côté ces informations pour la caractérisation des parcours, et à se tourner vers l'exploitation de la structure en catégories.*

3.3.3 Projection des annuaires sur les parcours

Les annuaires représentent une somme d'information majeure pour décrire les contenus du Web, mais correspondent-ils aux pages visitées par les internautes, et vont-ils représenter une ressource suffisante pour l'analyse des parcours ?

Une couverture satisfaisante avec les URL visitées par les internautes

Nous avons confronté la liste des URL indexées par les annuaires à celle des URL visitées par les panels SensNet en 2001 et 2002. Pour cela, nous avons regroupé l'ensemble des URL des différents annuaires.

À cette étape, nous avons dû tout d'abord dédoublonner les URL indexées dans les différents annuaires. En effet, le mécanisme de « fichier par défaut » des serveurs Web fait qu'en l'absence de nom de fichier spécifié par l'utilisateur, la requête renvoie le contenu d'un fichier dont le nom correspond à une liste définie dans la configuration du serveur, du type `index.html` ou `index.php`. Deux adresses différentes dans les annuaires, comme `http://www.lerepairedesmotards.com/` et son équivalent `http://www.lerepairedesmotards.com/index.htm`, peuvent alors pointer vers le même contenu. Pour parer à ce problème, nous avons, lorsque ce cas semblait se présenter, comparé le résultat des deux requêtes pour voir s'il fallait regrouper ou non des URL distinctes *stricto sensu*.

Ensuite, les URL indexées ont été normalisées, et une identification particulière de la racine des sites a été pratiquée pour les pages personnelles (comme nous l'avons fait pour les URL de la base de données de trafic). Après cette étape de normalisation, nous avons comparé les URL des annuaires à celles visitées par les internautes, en gardant comme impératif que l'annuaire ne soit jamais plus spécifique que l'URL visitée. Si l'annuaire comporte une entrée « La cote de Motomag » à l'adresse `http://www.motomag.com/cot/`, on retiendra que l'annuaire décrit bien l'URL `http://www.motomag.com/cot/honda.html`, mais pas l'URL `http://www.motomag.com/` ni plus généralement toutes les URL sur ces domaine qui ne seraient pas dans le répertoire `/cot/`. Sur cette base, quatre niveaux d'appariement ont été définis, de la description générale du site par l'annuaire à celle de l'URL précisément visitée. Les niveaux d'appariement constatés sur les données SensNet 2002 confirment que les annuaires indexent majoritairement des sites et non des pages (voir Tableau 3.34), sauf dans certains cas particulier où ils pointent vers des ressources ou des documents particuliers dans des sites de grande taille (administrations, universités, portails, etc.).

Tableau 3.34. Précision de l'appariement entre annuaires 2002 et trafic SensNet 2002

Niveau d'appariement	Part des URL
URL exacte	1,8 %
Fichier sans paramètres (CGI)	1,1 %
Répertoire (y compris la racine du site)	27,8 %
Site	69,3 %

Nous avions vu précédemment que près de 63 % de l'ensemble des URL des sept annuaires généralistes ne sont indexées que par un seul annuaire ; nous pouvions donc nous attendre à ce que les annuaires aient des taux de couverture des pages visitées par les internautes très variés. Malgré cela, nous constatons que les taux de couverture des annuaires en 2002 sont assez similaires, variant entre 26 % et 32 % pour les sept annuaires généralistes (voir Tableau 3.35). Les calculs fait sur la base du temps passé sur chaque page renvoient des chiffres similaires (écarts d'un ou deux points).

Tableau 3.35. Couverture par les annuaires en 2002 des URL vues

	Trafic 2000	Trafic 2001	Trafic 2002	BibUsages
Looksmart	32,8 %	36,6 %	31,8%	33,5 %
Lycos	28,8 %	27,1 %	21,0%	25,8 %
MSN	33,2 %	34,9 %	29,7%	32,1 %
Nomade	28,9 %	32,1 %	30,2%	32,9 %
Open Directory	23,9 %	24,5 %	20,7%	21,7 %
Voila	27,4 %	33,9 %	31,7%	34,3 %
Voila PP	1,5 %	1,2 %	1,1%	1,2 %
Yahoo	30,6 %	33,7 %	27,5%	29,8 %

Nous pouvons donc supposer que, dans l'ensemble, les annuaires indexent des sites « de référence », qui concentrent beaucoup de trafic, et qu'ils sont là en adéquation avec leur mission de sélection et de conseil de sites. Ceci est confirmé par le fait que, alors qu'un site vu en 2000 par notre panel est présent en moyenne dans

8,7 sessions, ceux indexés par les annuaires sont présents en moyenne dans 12,6 sessions.

Tableau 3.36. Couverture par les annuaires en 2001 des URL vues

	Trafic 2000	Trafic 2001	Trafic 2002	BibUsages
MSN	31,6 %	31,6 %	24,1 %	20,1 %
Nomade	32,7 %	30,8 %	24,9 %	28,1 %
Open Directory	25,7 %	25,9 %	19,2 %	22,5 %
Voila	28,6 %	27,3 %	23,0 %	24,9 %
Voila PP	9,3 %	11,5 %	9,0 %	4,5 %
Yahoo	32,3 %	31,8 %	23,1 %	18,5 %

L'évolution de la couverture des annuaires en 2002 projetés sur les données 2000, 2001 et 2002, ainsi que les taux de couverture calculés sur la base des annuaires 2001 présentés au Tableau 3.36, nous montrent également un très fort effet de mise à jour des annuaires : non seulement les annuaires dans leur version 2001 couvrent mieux le trafic de l'année 2000 que celui de 2001, mais plus encore, un an plus tard, les annuaires de février 2002 couvrent moins bien le trafic 2000 que le trafic 2001, et ce malgré une augmentation moyenne de leur taille de 14 %. Les annuaires font donc un réel effort pour se mettre à jour et présenter une image fiable du Web.

Dans le même temps, les annuaires 2001 couvrent mieux le trafic 2000 que le trafic 2001 ; de manière similaire, la version 2002 des annuaires décrit mieux les pages visitées en 2001 que celles vues en 2002 : si les annuaires font un effort de mise à jour, ils suivent le trafic et l'audience mais ne la précèdent pas. Nous sommes donc pleinement motivés à employer les annuaires aspirés en 2002 plutôt qu'en 2001 pour la description des contenus visités[1].

Si l'on considère maintenant l'ensemble des URL décrites par les annuaires, la couverture globale avec les parcours est relativement importante : 48,3 % des 6,7 millions d'URL uniques visitées par le panel 2002 figurent dans les huit annuaires, qui représentent 42,5 % des 27,2 millions de pages vues. Cette couverture somme toute satisfaisante des pages visitées par les annuaires nous autorise à les utiliser pour décrire et caractériser les parcours des internautes sur le Web. À partir d'une liste d'URL « à plat », il devient possible de disposer d'informations sur les contenus visités en utilisant les descriptifs des sites proposés par les annuaires, mais également la catégorie dans laquelle se situe le site dans la structure de l'annuaire. Voici à titre d'exemple, la description par Open Directory en 2001 d'une session, effectuée le 21 décembre 2000 et comportant 21 pages visitées sur 3 sites différents :

[1] Même si l'on aurait pu souhaiter disposer d'une version 2003 des annuaires, que les contraintes temporelles ne nous ont pas permis de réaliser.

19:45:41 – 1 URL visitée sur `www.libertysurf.fr`	
	Liberty Surf : gratuité totale : 4 heures - 4 heures gratuites par mois. Fournisseur d'accès gratuit à internet sur toute la France et illimité en nombre d'heures et d'utilisateurs. Accès gratuit et portail de services.
	Régional ➔ France ➔ Commerce et économie ➔ Internet ➔ Fournisseurs d'accès ➔ *Gratuit*
19:46:06 – 10 URL visitées sur `www.boursorama.com`	
	Boursorama - Actualité des marchés, informations financières et conseils, cours des plus grandes places boursières, indices et palmarès.
	Commerce et économie ➔ Finance ➔ *Bourse*
19:51:39 – 10 URL visitées sur `www.anpe.fr`	
	ANPE - Agence Nationale Pour l'Emploi - Présentation des services de cette agence française. Consultation des offres d'emploi et informations générales sur le secteur, notamment en ce qui concerne les aides à l'embauche.
	Régional ➔ France ➔ Commerce et économie ➔ *Emploi*
19:52:49	fin de la session

Toutefois, l'emploi des annuaires ne va pas de soi : au-delà des taux de couverture, c'est l'hétérogénéité de ces ressources qui pose problème, ainsi que le type d'informations à exploiter pour décrire les parcours.

Précautions méthodologiques

Les annuaires constituent une mine d'information pour l'analyse des parcours, mais les différences qui les opposent en termes de taille, de structure, de choix organisationnels et éditoriaux et de style invitent à réfléchir sur leur mobilisation pour la description des parcours. Les différents formats de descriptifs textuels constituent un frein à leur exploitation conjointe. Pour autant, chaque annuaire pris isolément décrit moins bien que l'ensemble des huit ; n'est-il pas possible d'utiliser les éléments de structure en appareillant les catégories d'annuaires différents ?

Sur ce point, nous avons tenté de rassembler les catégories d'annuaires sur la base des sites qu'elles indexent : si un annuaire A_1 regroupe un ensemble de sites sous une catégorie donnée, jusqu'à quel point un annuaire A_2 va-t-il rapprocher ce même ensemble de sites ? En utilisant des calculs formels sur des graphes, nous avons construit des indicateurs numériques de l'*accord entre annuaires*. Si nous n'avons pas le loisir de développer ici les détails techniques de ce calcul, nous pouvons affirmer que les annuaires sont assez souvent en désaccord sur le regroupement et la classification des sites qu'ils indexent en commun : deux sites qui ont été regroupés sous la même catégorie dans un annuaire A_1 se retrouvent assez souvent classés dans des catégories disjointes et éloignées dans un annuaire A_2. Ceci s'explique par des facteurs structurels (multi-indexation, taille et finesse des catégories, etc.) mais également par des facteurs plus qualitatifs, liés aux principes de classement (co-existence des découpages géographiques et thématiques, etc.) et aux choix éditoriaux spécifiques à chaque annuaire.

Ce constat nous interdit d'utiliser conjointement les différents annuaires pour la description des parcours ; il ouvre dans le même temps la voie d'une mobilisation parallèle des différents annuaires à notre disposition, chacun apportant ses

spécificités, et les résultats constatés avec l'un confirmant ou non ceux issus d'un autre.

Les taux de couverture individuels présentés au Tableau 3.35 montrent que l'on peut décrire environ un tiers des pages vues avec chaque annuaire. En outre, si l'on ajoute à cette couverture les résultats de la catégorisation des services avec *CatService*, 790 000 d'URL supplémentaires sont caractérisées. Ainsi, 60,6 % des URL visitées sont couvertes par les annuaires et *CatService* (l'un, l'autre ou les deux ensemble), pour 55,8 % des pages vues : nous avons donc une description du contenu d'une bonne moitié des pages visitées par les panélistes sans avoir à les aspirer, ce qui représente une couverture satisfaisante.

Rien ne nous interdit, dès lors, de nous concentrer sur les sessions « bien décrites » : il nous faut pour cela estimer qualitativement les zones laissées de côté, afin de maîtriser le biais induit par cette sélection. Deux éléments font que des URL échappent à la description par les annuaires : soit elles se situent sur des noms de domaines différents de celui indexé dans l'annuaire, ce qui est le cas pour la plupart des grands sites (par exemple : `fr.news.yahoo.com` regroupe les informations du portail `fr.yahoo.com`) soit elles figurent sur des sites qui n'existent pas dans les annuaires. Pour le premier cas, le recours à *CatService* vient combler ce vide. Pour le second, dans la mesure où les annuaires indexent des sites de référence, à forte notoriété, et désireux de se faire connaître[1], on peut estimer que la moitié du trafic non décrit par les annuaires et *CatService* correspond principalement à des sites d'importance secondaire, ou qui ne désirent pas de publicité, ou bien des sites pornographiques, les grands oubliés des guides généralistes du Web. C'est donc plutôt du côté du pornographique, de l'illicite ou du confidentiel que se situe le silence des annuaires dans les données de trafic, ce dont il faudra tenir compte dans les analyses par la suite.

> ***Synthèse.*** *Chaque annuaire généraliste pris isolément permet de caractériser environ 30 % des pages vues par le panel SensNet 2002 ; utilisés conjointement, ce sont plus de quatre pages sur dix qui sont décrites. Pour autant, les différences entre les plans de classement interdisent une mobilisation combinée des huit annuaires, même si le faible recouvrement entre eux incite à maximiser la couverture avec les parcours. Pour améliorer celle-ci, on se tournera plutôt vers les descriptifs issus de* CatService*, qui, associés aux annuaires, permettent de décrire 56 % des pages vues en 2002 par le panel.*

Conclusion

La qualification des données brutes de trafic, étape indispensable à l'analyse du contenu des parcours, pose des problèmes complexes auxquels nous n'avons pas de

---

[1] Et, de plus en plus, des sites prêts à payer pour apparaître dans ces outils de recherche d'information sur le Web, avec le délaissement des soumissions de sites gratuites au profit des inscriptions payantes.

solution absolue. Au fil des différentes approches que nous avons mises en œuvre, il est apparu que les informations contenues dans l'URL elle-même, d'ordre technique, sont trop pauvres pour être exploitées de manière générale pour qualifier le trafic. Tout au plus pourra-t-on mobiliser ces informations dans des contextes locaux pour résoudre certains problèmes spécifiques (accès à des contenus sécurisés, accès à des formats atypiques de documents).

Les méthodes véritablement productives sont celles qui apportent des éléments descriptifs externes, que ceux-ci soient supervisés (catégorisation avec *CatService*), exogènes (annuaires) ou endogènes (contenu des pages visitées). Dans les deux premiers cas, on bénéficie d'une information structurée et disposant de plusieurs niveaux d'agrégation, ce qui est un atout pour l'analyse. L'approche *CatService* ne prétend pas couvrir l'ensemble du trafic et des usages, mais elle décrit très bien les principaux sites visités par les internautes, ainsi que les services utilisés : cette distinction en services rompt avec l'idée trop répandue que le Web n'est qu'un réservoir d' « informations » éparses, et ouvre la voie d'une séparation dans les traitements entre les pages orientées vers les services et l'outillage, et les pages orientées « lecture ».

Avec les annuaires, on dispose au contraire de données structurées couvrant par vocation l'ensemble des contenus Web, même si la couverture avec les parcours n'est pas complète. Si les modes d'organisations des objets du Web sont différents pour chaque annuaire, chacun d'entre eux fournit des informations éminemment exploitables pour l'analyse des parcours, en particulier dans la structuration des catégories plutôt qu'à travers les descriptions textuelles des sites.

L'aspiration de pages apporte quant à elle des informations non structurées, hétérogènes, et difficiles à traiter. Cela étant, elles ont l'avantage de donner une couverture totale des contenus visités, *modulo* les biais induits par la récupération des pages en différé des visites. Les travaux existant sur la catégorisation de sites et de pages montrent la voie d'un classement sur la base des contenus, mais ils sont à ce jour trop peu avancés pour que nous puissions les exploiter pleinement. Nous en retiendrons que l'exploitation des contenus doit tirer parti de l'ensemble des éléments propres au Web (texte, liens, structure, éléments multimédia), mais que les outils actuels ne permettent pas de le faire, tandis qu'une approche uniquement basée sur le lexique des pages s'est avérée inopérante.

In fine, la combinaison des informations issues des annuaires et de *CatService* décrit le plus efficacement les parcours sur le Web. C'est en tirant parti de la complémentarité de ces différentes informations que l'on parviendra à les exploiter pleinement : nous verrons alors comment contourner les problèmes posés par l'utilisation conjointe de ces descriptions hétérogènes. Au-delà de ces difficultés, l'enrichissement des données de trafic ouvre la voie d'une confrontation fructueuse avec des indications relatives à la temporalité et la forme des parcours, et constitue une base solide à l'analyse de la navigation Web sous le double angle des contenus et de la topologie en les replaçant dans le champ de la lecture et de l'action au sein d'un univers hypertextuel.