Thèse de doctorat de l'Université de Paris X – Nanterre

Spécialité : Sciences du langage

présentée par
Thomas BEAUVISAGE

pour obtenir le grade de Docteur de l'Université de Paris X

Sémantique des parcours des utilisateurs sur le Web

Sous la direction de François RASTIER

Soutenue en octobre 2004

Devant le jury composé de :

M. Houssem ASSADI
M. Dominique BOULLIER (rapporteur)
M. Benoît HABERT
M. Ludovic LEBART
M. François RASTIER (directeur)
M. Pierre ZWEIGENBAUM (rapporteur)

Résumé

Notre thèse a pour objectif de décrire les parcours sur le Web sur la base de données de trafic centrées-utilisateur. Nous proposons des méthodes et des outils pour enrichir de telles données de trafic, et les mettons en application pour construire une segmentation des parcours sur la base de leur forme, de leur temporalité, de leur contenu et de leur insertion dans les pratiques individuelles. Ce travail, mené au laboratoire Usages, Créativité, Ergonomie de France Télécom R&D, s'inscrit dans le projet SensNet qui vise à analyser les usages d'Internet à domicile.

La généralisation de l'accès à Internet en France entraîne une banalisation et une normalisation des pratiques du Web. Pour autant, l'activité de navigation reste mal connue : si l'analyse des *logs* des serveurs Web est maintenant bien maîtrisée, celle des traces de navigation recueillies du côté de l'internaute en situation naturelle demeure rare et complexe. Les données utilisées dans cette étude centrée-utilisateur proviennent de sondes de recueil de trafic Internet installées sur les postes des utilisateurs à domicile ; on obtient alors la liste des URL visitées par chaque internaute, qui constitue le matériau premier de l'étude. Sur cette base, nous proposons une description des parcours des internautes de page en page et de site en site centrée sur la session. Cette description intègre les informations sur les contenus visités d'une part et les territoires personnels sur le Web d'autre part, et examine leur articulation dynamique au sein des parcours.

Pour y parvenir, un premier travail consiste, après une première mise en forme de ces données brutes, à les enrichir. Sur le plan des contenus, nous proposons une méthode qui exploite les informations fournies par les annuaires du Web pour qualifier les URL visitées. Adossée à un module d'identification des services sur les portails généralistes développé dans le cadre du projet SensNet, cette description permet d'appréhender l'offre de contenus du Web dans sa diversité : informations, mais aussi services, outils, fonctionnalités. Sur le plan de la navigation, nous élaborons des indicateurs statistiques simples qui rendent compte de la forme, de la temporalité et du rythme des parcours, à l'échelle de la page et du site. En complément de cette approche *macro*, nous avons développé des outils de fouille manuelle des sessions permettant de vérifier les résultats de l'approche quantitative et de formuler des hypothèses sur les comportements des internautes. Ainsi dotés, nous disposons des outils nécessaires pour observer, au sein de données volumineuses, les liens entre forme et contenus des parcours, et mettre à jour des régularités dans les pratiques des internautes.

Nous appliquons cet outillage à trois panels : un panel représentatif de plus de 3 300 internautes en 2002, une cohorte de 600 personnes observées sur trois ans, et un panel restreint d'utilisateurs des bibliothèques numériques. Ces trois sources de données complémentaires nous amènent à établir une première typologie des sessions sur la base de leur forme et de leur temporalité : les cinq parcours-type mis à

jour s'opposent sur le plan de leur durée, de leur forme et de leur rythmique, et montrent la grande diversité des comportements.

Examinés sous l'angle des territoires personnels, ces modes prototypiques de navigation prennent sens. Au sein d'espaces Web *a priori* non bornés, les internautes dessinent des zones familières de taille restreinte autour de thématiques propres à chacun. Trois zones distinctes sont mises en évidence, auxquelles correspondent des modes d'activité et des types de contenus spécifiques : le familier, orienté vers des contenus à fort taux de renouvellement (flux d'information, services de communication), constitue le noyau dur de l'activité de navigation, et induit des parcours routiniers rapides et ciblés qui s'apparentent aux modes de consommation des média traditionnels (télévision, radio, journaux). Le territoire occasionnel délimite des zones visitées moins fréquemment, mais de manière régulière dans un contexte donné, et cible les contenus de type service ou achat : dans ce cadre, les sessions s'allongent et se complexifient, mais l'espace hypertextuel demeure connu et maîtrisé. Enfin, les parcours de découverte amènent l'internaute à mobiliser le Web comme ressource informationnelle ponctuelle de manière ciblée : dans ces sessions où la ligne brisée domine, les moteurs de recherche dessinent un espace de sites que l'utilisateur ne reverra plus pour la majorité d'entre eux.

Sur le plan méthodologique, ces résultats attestent la capacité de notre outillage à décrire et expliquer les comportements de navigation sur le Web ; ils montrent également la nécessité pour une sémantique des parcours de tenir compte des déterminations globales pour comprendre les comportements locaux, et de mener l'étude des usages sous une angle praxéologique.

Sur le plan des pratiques, on observe ainsi que le parcours Web est la résultante d'une double dynamique, celle des contenus proposés et celle de l'utilisateur, dont la confrontation induit des modalités d'activité qui dépendent autant des contenus eux-mêmes que de leur appréhension et de leur valorisation par l'utilisateur. Loin de « surfer » au gré des hyperliens, l'internaute construit, au sein d'un vaste espace hypertextuel, des zones restreintes de familiarité qui constituent l'essentiel de ses pratiques sur le Web.

Abstract

This thesis aims at describing users' paths through the Web on the basis of user-centric traffic data. We propose methods and tools to enrich traffic data, and apply them to build a segmentation of Web paths based on their shape, their temporality, their content and their place in individual practices. Our work took place in the Uses, Creativity, Ergonomics laboratory at France Telecom R&D, within a project named SensNet dealing with the analysis of domestic uses of the Web.

The generalization of Internet access in France leads to a normalization of Web practices. However, the activity of Web browsing itself remains rather unknown: while the analysis of Web servers access logs is now widely practiced, those of user-centric real-world traffic data is still rare and complex. This study relies on the analysis of data collected by probes installed on users' computers at home, which provide the time-stamped list of all the URLs visited by each Internet user. On this basis, we propose a description of Web users' paths through pages and sites centred on the session. This description integrates information on the content of pages and sites as well as on personal territories on the Web, and examines their dynamic articulation inside Web paths.

To achieve this goal, after data preparation for the analysis, we have to enrich them first. On the side of content description, we propose a method which exploits information provided by Web directories to qualify the visited URLs. Combined with a module for identifying services on generalist portals developed within the SensNet project, this description reflects the diversity of Web contents: information, but also services, tools, functionalities. On the side of browsing, we calculate robust statistical indicators which represent the form, the temporality and the rhythm of Web paths, both at page-scale and site-scale. Beside this *macro* approach, we developed tools for manually exploring sessions, that allow to verify the results of quantitative approach and to formulate hypothesis concerning Internet users' behaviour. Thus, we have the necessary tools to observe inside large datasets, links between paths' topology and content, and to highlight regularities within Web users' practices.

We apply these tools to three panels: a representative panel of more than 3.300 users in 2002, a cohort of 600 people observed during three years, and a small panel of digital libraries users. These three complementary datasets allow us to build a typology of sessions based on their topology and their temporality: the five discovered types of paths differ in terms of duration, form and rhythm, and demonstrate the great diversity of browsing behaviours.

These prototypical modes of navigation make sense when considered from the angle of personal Web territories. Within *a priori* unlimited spaces, Web users outline small zones related to specific topics. Three distinct zones are identified, which correspond to particular modes of activity and content types: the familiar

territory, oriented on regularly updated contents (information streams and communication services), forms the core of users' browsing, and implicates fast and targeted routine paths related to traditional mass media consuming modes (television, radio, newspapers). The occasional territory refers to zones which are less often visited, but regularly in a given context, and to service and e-commerce contents: in that case, Web paths are longer and more complex, whereas the hypertextual space still remains well-known and under control. Finally, in discovery paths, Internet users make use of the Web as information resource for targeted searches: in these highly non-linear sessions, search engines are often mobilized to explore Web spaces which will, for most of them, never be visited again by the user.

On the methodological side, these results attest the ability of our tools to describe and explain navigation behaviours on the Web; they also demonstrate the necessity for a semantics of Web paths to take into account global factors to understand local behaviours, and to have a praxeological approach of usage studies.

On the side of practices, we observe that a Web path results from a two dynamics: the one of the proposed contents by Web sites, and the one of the user. Their confrontation in context implicate distinct modes of activity which depend as much on the visited contents as on their reception and their valuation by the user. Far from wildly "surfing" the Internet from link to link, Web users define, within a vast hypertextual space, restricted familiar zones that constitute the core of their practices on the Web.

Remerciements

Ce travail de thèse m'aura au moins appris deux choses. La première est que la recherche est une affaire de collaborations plus que d'individualités, et qu'aucun travail sérieux ne saurait être mené sans être inscrit, de quelque manière et à quelque niveau que ce soit, au sein d'un collectif de recherche.

Je ne saurai donc manquer de remercier toutes les personnes avec lesquelles j'ai été amené à collaborer au cours de ce travail : François Rastier, en premier lieu, m'a témoigné sa confiance en acceptant de diriger cette thèse qui se nourrit abondamment de son travail et de ses conseils ; Valérie Beaudouin m'a chaleureusement accueilli au sein du laboratoire Usages, Créativité, Ergonomie de France Télécom R&D, et Houssem Assadi a encadré et guidé ce travail : tous deux m'ont témoigné une disponibilité et un soutien sans faille ; et les participants, réguliers ou occasionnels, des projets TypWeb et SensNet : la mise en commun des résultats, les réunions régulières et les discussions critiques en ont fait un projet efficace, vivant et fondamentalement collectif sans lequel notre travail n'aurait pu aboutir.

En second lieu, cette thèse m'a montré qu'un bon lecteur est une chose précieuse. Outre ceux qui ont accompagné ce travail, je tiens à remercier Dominique Boullier, Benoît Habert, Ludovic Lebart et Pierre Zweigenbaum d'avoir accepté de faire partie du jury de cette thèse et d'en être les lecteurs privilégiés. C'est avec plaisir que je soumets ce mémoire à leur jugement et leurs critiques.

Enfin, j'ai pu vérifier au cours de ces quelques années de travail que les discussions de couloir sont au moins aussi importantes que les échanges plus formels, et que ces à-côté prennent parfois la forme de l'essentiel. Merci donc à Thomas, Julien, Marie, Julia, Marc, Shark et les autres, qui ont contribué à ce travail bien plus qu'ils ne le croient.

Sommaire

Introduction

En l'espace de quelques années, l'essor du Web s'est accompagné d'une multiplication et d'une diversification de l'offre de contenus, en même temps que d'une généralisation de l'accès aux ressources. La croissance rapide de ce média a suscité dans les premiers temps les spéculations les plus diverses sur des usages encore en construction : « révolution numérique », avènement du « virtuel », nouvelle ère de l'écrit, le Net a été l'objet d'espoirs ou de rejets extrêmes. Depuis lors, de NTICs, *Nouvelles* Technologies de l'Information et de la Communication, en simples TICs, l'objet a perdu l'attrait de la nouveauté tandis que les pratiques tendaient à se banaliser et à se normaliser, et le discours spéculatif a laissé place à l'observation des pratiques réelles. Pour autant, rares sont encore les matériaux d'étude qui permettent d'en rendre compte de manière exhaustive et objective, de sorte que les usages du Web en situation sont encore mal connus et peu décrits : le parcours sur le Web, moment particulier de la rencontre entre production et réception, reste encore à découvrir.

Contexte

En matière de données, pourtant, rarement un support médiatique aura comme le Web donné la possibilité de recueillir autant de traces d'usage : c'est particulièrement le cas du côté des serveurs Web, où l'analyse des *logs*[1] est aujourd'hui monnaie courante. Les visites des internautes y sont enregistrées, comptabilisées et étudiées par les concepteurs de sites pour l'amélioration de leur offre, l'analyse de l'audience, etc. Corrélativement, la connaissance des parcours sur le Web du point de vue des sites est assez avancée aujourd'hui, et constitue un des fondements du *Web Usage Mining*, champ de recherche constitué dans le milieu des années 90 autour de l'analyse des usages du Web. Toutefois, une collection de points de vue centrés sur les sites ne saurait rendre compte des usages individuels : les sites ne connaissent pas plus leurs visiteurs que les contextes d'usage dans lesquels ils s'inscrivent ou les dynamiques personnelles. Malgré cela, les données de trafic centrées-utilisateur sont rares, difficiles à constituer, et seule une poignée de travaux ont pu disposer d'un tel matériau, pour des études de courte durée.

D'autres approches plus qualitatives ont su se placer du côté de l'utilisateur pour en observer finement les pratiques. Les sciences cognitives, dans la lignée des travaux sur les hypermédias, ont pratiqué des expériences en laboratoire afin d'étudier des comportements dans des contextes précis, notamment la recherche d'information ; les conclusions de ces travaux viennent le plus souvent alimenter un projet plus

[1] Historique de l'ensemble des requêtes adressées à un serveur ; voir Glossaire.

global de modélisation de l'utilisateur et de recherche de modèles mentaux impliqués lors de la navigation sur le Web. Malgré des tentatives pour élaborer des modèles cognitifs capables d'expliquer l'ensemble des comportements sur le Web, ces dispositifs se heurtent à un problème de généralisation des résultats à partir des expérimentations locales, et peinent à rendre compte de la diversité des situations rencontrées par les internautes.

Dans un autre champ disciplinaire, la sociologie des usages, l'ethnométhodologie et les sciences de l'information et de la communication ont cherché à décrire les pratiques en situation naturelle, par le biais de questionnaires, d'entretiens, d'observations ou d'enregistrements vidéo. Si ces approches ont su mettre à jour les implications sociales, interactionnelles et sémiotiques des TICS, elles peinent, sur la question des parcours, à atteindre une analyse à la fois fine et globale. D'une part, la méthodologie des questionnaires ou des entretiens, au-delà du statut particulier que l'on peut conférer aux déclarations et aux discours tenus dans ce cadre, dressent des descriptions à gros grain des pratiques ; d'autre part, les méthodes d'observation directe ont pour elles la richesse d'une description détaillée des modes d'activité, mais elles peinent à rendre compte de la globalité des pratiques et des situations et se heurtent à un problème de « masse critique » des données.

Une description tout à la fois globale et fine des usages de la Toile reste donc à construire, sous la forme d'une typologie des comportements de navigation tenant compte de l'offre de contenus du Web, du contexte local de l'utilisateur et de la dynamique de ses pratiques personnelles, et de la construction globale de pratiques normées par l'ensemble des internautes. Elle s'appuiera à bon droit sur des données de trafic centrées-utilisateur, qui offrent tout à la fois une perception exacte des contenus visités et une vue dans la durée de la diversité des situations d'usage. Pour cela, elle nécessite d'élaborer des méthodes et des outils capables de tenir compte de la spécificité des contenus Web et de leur mode d'appréhension, afin de rendre possible l'élaboration d'une sémantique des parcours.

Objet d'analyse et champs disciplinaires

Cette étude entend apporter une contribution à la description des usages du Web, fondée sur l'analyse de corpus de parcours effectués en situation naturelle. Nous prenons pour cela appui des données de trafic d'internautes recueillies auprès de trois panels : un panel représentatif de plus de 3 300 internautes en 2002, une cohorte de 600 personnes observées sur trois ans, et un panel restreint d'utilisateurs des bibliothèques numériques. Ce travail, mené au sein du laboratoire Usages, Créativité, Ergonomie de France Télécom R&D, s'inscrit dans le cadre plus large de projets d'étude des usages d'Internet à domicile[1].

L'objet d'étude est le parcours, c'est-à-dire la visite ordonnée et déterminée temporellement de pages et de sites Web, dans le cadre d'une unité d'action cohérente, la session. L'analyse des parcours se fait en corpus : de la même manière que la linguistique de corpus a profondément renouvelé l'approche du matériau

[1] Projets TypWeb, SensNet et BibUsages ; voir Annexe 1 pour une description complète.

textuel en s'appuyant notamment sur l'analyse de masses de documents numérisés, l'étude des parcours sur le Web s'appuie ici sur un vaste corpus de parcours au sein duquel nous souhaitons distinguer des comportements significatifs et récurrents.

Pour manipuler cet objet particulier, notre travail mobilise plusieurs disciplines :

- *sémantique textuelle* : la sémantique interprétative textuelle a montré, du palier des lexèmes à celui des textes, la construction contextuelle du sens, la détermination du local par le global et l'inscription des pratiques d'écriture et de lecture dans des genres et des situations. Les contenus du Web et leur appréhension n'échappent pas à ces déterminations sémantiques ; ils ne sauraient toutefois s'y réduire, pour deux raisons. D'une part, la dimension hypertextuelle du média tend à briser les unités textuelles en privilégiant le fragment et la recomposition d'un ensemble à partir de sources au sein du parcours ; d'autre part, les contenus Web ne peuvent être envisagés sous l'angle des textes uniquement, mais également comme un espace d'action et d'outils mis à disposition de l'internaute. Une sémantique des parcours s'appuiera donc sur une sémantique textuelle adaptée aux spécificités des contenus du Web et de leurs modes d'appréhension.
- *sociologie des usages* : l'analyse des parcours tire parti des travaux déjà menés dans le champ de la sociologie et de l'action située. D'un côté, les enjeux de l'usage des TICS sous l'angle des inégalités (thème du « fossé numérique »), de la constitution des communautés et des collectifs, et de l'impact sur les organisations forment un entour global qui guide l'interprétation des modes de navigation ; de l'autre, les approches plus praxéologiques centrées sur l'analyse de l'action en situation, notamment dans le champ de l'ethnométhodologie, apportent des descriptions situées des usages auxquelles se rattache plus particulièrement notre travail.
- *analyse de données et Web Mining* : l'analyse de données de trafic volumineuses et centrées-utilisateur nécessite de se doter des outils nécessaires à leur manipulation. Il est donc question d'informatique, à double titre : en premier lieu, le substrat technique des données et du Web lui-même doit être connu pour être manipulé ; ensuite, le caractère exploratoire de cette recherche implique de tester les limites des outils, des méthodes statistiques et des représentations existants, et d'être capable d'en produire de nouveaux pour manipuler les corpus de parcours.

Et, au-delà de l'inscription pluridisciplinaire de l'analyse de corpus de parcours, notre travail se place résolument dans le champ des sciences humaines, sous l'angle de la description des pratiques.

Notre contribution

Cette étude est exploratoire à double titre : d'une part, elle vise une description des pratiques de navigation à domicile sur des durées et des panels inédits à ce niveau de détail. D'autre part, elle s'appuie sur des données de trafic centrées-utilisateur, dont l'exploitation encore rare nécessite la mise en place d'outils et de méthodes *ad hoc*.

1. *Méthodologie et outils pour analyser les données de trafic centrées-utilisateur*
   L'analyse de traces de navigation recueillies sur les serveurs Web propose des méthodes statistiques qui ne peuvent être appliquées aux données centrées-utilisateur pour deux raisons majeures : elles laissent de côté la question des contenus (connus, lorsque la navigation est analysée sous l'angle des sites), et elles reposent sur une redondance dans les données que l'on n'observe pas sitôt qu'on se place du côté de l'utilisateur. La diversité des contenus, des modes de navigation, des formes de parcours, nécessite la mise en place de méthodes d'analyse originales qui représentent tout à la fois les contenus visités et la dynamique des parcours. Sur le plan du contenu, notre travail met en œuvre et évalue différentes stratégies pour attacher aux listes d'adresses des pages visitées (ou *URL*[1]) par les internautes des informations sur les thématiques et les fonctions des pages et des sites vus. Sur le plan de la navigation dans la session, nous élaborons des indicateurs statistiques capables de représenter la forme et la rythmique des parcours.
   Le parcours devient alors un objet complexe à analyser, hétérogène du fait de ses constituants et de ses différentes métriques. Pour l'aborder, nous mettons en place des outils de fouille manuelle des données de trafic qui permettent d'approcher au plus près la réalité des parcours et leur logique, et nous proposons une démarche statistique descriptive qui tient compte de ces particularités.

2. *Segmentation des parcours et description des pratiques de navigation*
   Les méthodes et outils que nous mettons en place pour décrire les parcours sur le Web ne sont valables que tant qu'il servent effectivement l'analyse et la caractérisation des comportements de navigation. Les corpus de parcours dont nous disposons, inédits par leur taille et leur durée d'observation, nous permettent de construire une segmentation des parcours fondée sur l'observation de régularités et la mise à jour de contextes d'usages prototypiques. Les données longitudinales exhaustives permettent également d'observer la structure et l'évolution des territoires personnels sur le Web ; confrontée aux modes prototypiques de navigation, cette vision éthologique des parcours nous amène à distinguer des modes d'appréhension type des contenus dans le contexte de l'usage.

Organisation du mémoire

Cette thèse se décompose en deux grandes étapes, qui renvoient au caractère doublement exploratoire de notre travail. La première partie s'attache à décrire les méthodes et outils que nous avons été amenés à élaborer pour analyser les données des trafic centrées-utilisateur ; la seconde mobilise cet outillage pour l'étude des pratiques de navigation, et en propose une segmentation fine sous l'angle de la forme des parcours, de leur contenu et des territoires personnels sur le Web.

---

[1] Voir Glossaire.

Nous posons dans le Chapitre 1 un cadre d'analyse et de réflexion pour appréhender les parcours sur le Web, et positionnons notre travail par rapport aux études déjà menées sur cet objet. Après une présentation du type de données dont on dispose, le Chapitre 2 expose les différentes étapes de leur mise en forme : il s'agit de les nettoyer des scories qu'elles contiennent, et d'identifier, au sein des listes horodatées d'URL, des sessions, des pages, des sites. Le Chapitre 3 est consacré aux différentes stratégies mises en œuvre pour attacher aux données de trafic des informations de contenu : exploitation des URL brutes, aspiration de pages, catégorisation semi-automatique des URL, mobilisation des annuaires du Web. Cette caractérisation des contenus est complétée sur le plan de la dynamique des parcours par l'élaboration d'outils de fouille manuelle et d'indicateurs statistiques représentant la forme et la rythmique des sessions, qui sont exposés au Chapitre 4.

La seconde partie montre l'exploitation de cet outillage pour la description des usages du Web et la segmentation des parcours en situation. Le Chapitre 5 décrit tout d'abord la composition des trois panels sur lesquels se base cette étude ; il propose ensuite deux explorations des données sur la base de leur forme d'une part et de leur contenu de l'autre, qui permettent d'en dresser le profil et le comportement statistique. Ces éléments servent de base à une classification des sessions en cinq groupes à partir des indicateurs topologiques, sur lesquels nous projetons les contenus visités pour apercevoir le lien fort qui unit ces deux composantes. Le Chapitre 6 confronte ces profils-type de parcours avec des éléments de contexte relatifs à l'utilisateur : en s'appuyant sur la structure et la dynamique des territoires personnels sur le Web, nous mettons à jour trois modes d'appréhension prototypiques du Web qui impliquent tout à la fois les types de contenus et la dynamique de l'usage. L'examen spécifique des modes d'accès et de manipulation des bibliothèques électroniques et des fonds numérisés permet d'approfondir ce problème, et montre la prévalence de l'usage sur la nature propre des documents lorsque ceux-ci sont immergés dans le contexte du Web.

I

Appréhender la navigation sur le Web : questions, méthodes, données, outils

Qu'est-ce qu'un parcours sur le Web ? Comment appréhender cet objet de recherche articulant les logiques de production des contenus et celles de leur usage en situation ? Quelles données et quels outils va-t-on mettre en œuvre pour en construire des représentations fidèles ? C'est à cet ensemble de questions que répond cette première partie. Dans un premier temps, nous proposons une approche praxéologique des parcours fondée sur la description et l'interprétation des situations de navigation et des régularités construites par l'usage, en s'appuyant sur des données de trafic centrées-utilisateur. Après avoir décrit le format et les contraintes de ce matériau brut, nous exposons les méthodes que nous avons élaborées pour l'enrichir et le manipuler : représentation des contenus visités, indicateurs topologiques et temporels des parcours, outils de visualisation. Ce n'est qu'ainsi outillés que nous pouvons envisager ensuite une analyse des parcours sur le Web tenant compte de leur forme, de leur contenu et de leur inscription dans les pratiques individuelles.

Chapitre 1

Appréhender les parcours sur le Web

Le travail que nous présentons ici se propose d'analyser les parcours sur le Web sur la base de données de trafic centrées-utilisateur. Deux questions sont soulevées : d'une part, la définition de ce qu'est un parcours et en quoi cet objet s'articule dans l'activité de navigation en général ; d'autre part, à quels questionnements sur cet objet les données de trafic permettent-elles de répondre, et quelles méthodologies cela implique-t-il.

1.1 Le parcours comme objet d'analyse

Nous entendons poser les bases méthodologiques et empiriques d'une sémantique des parcours ; nous estimons qu'il y a là un champ de recherche à part entière dont la spécificité ne se construit pas uniquement sur la particularité sémiotique de son objet, mais également sur les modes d'interaction entre l'utilisateur et les contenus qu'il appréhende.

1.1.1 Le parcours au centre de l'activité de navigation

La diffusion en milieux domestique et professionnel des outils informatiques en général et de l'accès à Internet en particulier s'accompagne d'une banalisation de ces outils et de l'expérience de leur usage. Socialement attesté et délimité, le fait d'« aller sur le Web » est à ranger au même rang que « prendre un café » ou « passer un coup de téléphone » : si la dénomination n'en réduit pas la complexité ni la diversité en termes de pratiques et de situations, elle désigne une activité identifiée et bornée dans le temps qui légitime son étude en tant que telle.

Comprendre les parcours en situation

L'appréhension de cette activité particulière d' « aller sur Internet[1] » intéresse plusieurs champs disciplinaires, qui ne l'interrogent pas de la même manière, n'y valorisent pas les mêmes éléments – lorsque, travaillant sur le même aspect, ils n'y portent pas des vues divergentes du fait de prémices diamétralement opposées.

Puisqu'il s'agit d'activité, situons tout d'abord son cadre : on se limitera à l'accès à Internet par un terminal « classique » (informatique personnelle, de type PC), c'est-à-dire en excluant les accès *via* la téléphonie mobile et les assistants personnels. L'évolution des terminaux eux-mêmes et de leur interopérabilité bouleversera peut-être cette division, en autorisant des modes d'accès nomades (dans la rue, dans le train, etc.), semi-nomades (on pense par exemple aux bornes WiFi dans le cadre de l'accès à un réseau d'entreprise) ou mobiles (au sein de l'espace domestique). Posons qu'aujourd'hui encore, pour aller sur Internet, il faut un ordinateur, c'est-à-dire un terminal relativement volumineux composé d'un écran, d'un clavier et d'un dispositif de pointage – souris ou autres sur ordinateurs portables.

Cette question est d'importance dans le courant de l'action située : en appréhendant la navigation comme pratique incorporée, on ne peut manquer de s'attacher à décrire non seulement le contenu de l'écran, mais également le couplage de l'individu avec les dispositifs techniques de médiation (le clavier, la souris, etc.) et son entour, en termes de perturbation, de sollicitation ou de co-construction de l'activité[2]. Notre approche laissera de côté ces aspects, même si nous partageons avec les tenants de l'action située le souci de placer les pratiques en contexte. En centrant notre analyse non pas sur la pratique incorporée, mais sur la pratique à l'écran, on laissera de côté la question du couplage de l'individu avec l'outil technique, mais ce faisant, on met l'accent sur ce qui conduit l'activité, la guide et l'organise. Nous ne cherchons pas à minimiser les déterminations qu'induisent sur le cours d'action les sollicitations auxquelles l'individu est soumis lors d'une activité de navigation ; mais gageons qu'elle est au moins autant visible à l'écran que devant, et que c'est au sein même du passage de page en page et de site en site, dans le choix des contenus visités, dans la longueur des séquences de visionnage, leur rythmique, leur agencement interne, que se trouve l'essentiel d'un parcours sur le Web.

On traitera ici de la navigation sur le Web uniquement ; reconnaissons qu'il s'agit d'une réduction de l'utilisation d'Internet, qui permet également de faire de la messagerie asynchrone ou instantanée, des jeux, du téléchargement de fichier, etc. Nous laissons en particulier de côté les éléments d'entrelacement entre ces différents supports de l'accès aux ressources disponibles sur le réseau : un utilisateur peut suivre un lien à partir d'un mail, puis répondre par mail à l'expéditeur pour lui donner son avis sur le site en question ; dans le cadre d'une séance de *chat*, le Web peut être mobilisé comme ressource externe et support de conversation de manière

---

[1] « Internet » recouvre l'ensemble des applications, ressources et outils disponibles *via* le réseau : Web, messagerie, *peer-to-peer*, chat, etc. Le Web se restreint à l'accès à des sites par le biais d'un navigateur (voir Glossaire).

[2] Voir notamment les observations rapportées dans [Relieu & Olszewska 2004].

ponctuelle[1] ; dans une pratique de *peer-to-peer*, la recherche de fichiers à télécharger sur les réseaux d'échange peut se faire *via* une interface Web ; etc. Ces éléments ne doivent pas nous interdire de restreindre l'étude au Web : considérons que les autres outils Internet sont à mettre sur le même rang, pour ce qui est de l'analyse des comportements de navigation sur la Toile, que les éléments externes à Internet – programmes et documents présents sur l'ordinateur, matériau à la disposition de l'internaute en situation de navigation, interactions avec d'autres personnes en coprésence ou par téléphone.

Nous nous concentrerons donc sur l'activité de navigation « dans l'écran » et posons que, au titre d'activité, elle prend sens et se construit *en contexte*. En observant les parcours de page en page et de site en site, on se place à la croisée de chemins très divers, de situations variées, et dans des types de pratiques très différentes. De la même manière qu'une étude de la lecture doit prendre en compte les situations sociales dans lesquelles celle-ci se produit autant que ses supports et ses contenus, l'analyse des parcours se situe dans une diversité de pratiques sociales qui en modifient profondément le sens. Notre travail s'attachera à décrire des régularités et des modalités particulières de la pratique du Web, et les spécificités de ces modalités en regard des situations. Dans cette perspective, nous posons que l'appréhension des contenus est contextuelle, et ce à double titre : d'une part, deux personnes n'appréhenderont pas de la même manière une même page ou un même site, et d'autre part, un même contenu pourra être appréhendé différemment par un utilisateur donné dans deux contextes différents.

En ce sens, la sémantique des parcours que nous proposons se situe dans la perspective de la Sémantique Interprétative textuelle de F. Rastier[2], en reprenant au palier de l'appréhension des contenus du Web les éléments que la Sémantique Interprétative a mis à jour du côté des textes : détermination du local par le global, inscription des pratiques dans des genres et des situations, construction contextuelle du sens. Elle ne peut toutefois s'y réduire, pour deux raisons : d'une part, la dimension hypertextuelle du média tend à briser les unités textuelles en privilégiant le fragment et la recomposition d'un ensemble à partir de sources au sein du parcours. Au palier méso-sémantique, le déploiement des isotopies s'en trouve bouleversé, et ne peut être étudié qu'à travers l'infinité de corpus produits par la pratique : le site, qui pourrait correspondre au livre dans la mesure où il correspond à une unité éditoriale assumant une thématique et une topique spécifiques, n'est plus une unité d'analyse systématiquement pertinente en termes de réception et donc de plan interprétatif. D'autre part, les contenus Web ne peuvent être envisagés sous l'angle des textes uniquement : il ne s'agit pas uniquement du caractère multimédia des contenus (images, bandeaux, menus interactifs, etc.), qu'une sémantique textuelle est à même de prendre en compte, mais du fait que le Web propose, outre des contenu à « lire », de l'outillage. Celui-ci se compose d'outils de recherche, de communication (WebMail, forums, WebChat, etc.), de jeux, d'achat en ligne, etc. qui

---

[1] Voir par exemple [Beaudouin & Velkovska 1999].

[2] Voir notamment [Rastier 1987] et [Rastier 1989].

valent en tant que support d'activité et induisent des séquences d'action bien distinctes des contenus qu'ils véhiculent[1]. Ce n'est donc pas, une sémantique textuelle, mais une théorie de l'action qui permettra d'appréhender les parcours sur la Toile ; on posera ainsi, en suivant F. Rastier dans le projet d'une Sémiotique des cultures, que :

> Les théories de l'action ont privilégié l'axe de la représentation, et notamment le rapport entre les représentations et la motricité, c'est-à-dire les deux niveaux périphériques de la zone identitaire. Le niveau sémiotique est resté peu questionné en tant que tel. Si l'on tient compte maintenant de l'axe de l'interprétation, il faut prendre en considération les trois disciplines qui s'y articulent, d'ailleurs non sans de notables différences de statut.
> (i) Au niveau des présentations, la phénoménologie et la psychologie rivalisent pour décrire le flux de conscience comme activité.
> (ii) Au niveau sémiotique, l'herméneutique matérielle, entendue comme organon de la sémiotique des cultures, décrit les cours d'action sémiotiques. Discipline subordonnée, la sémantique interprétative prend spécifiquement pour objet le sens des textes.
> (iii) Au niveau physique, la praxéologie comprend une kinésique, mais aussi une technologie qui inclut les techniques du corps. Ainsi se dessinerait un regroupement de disciplines considérées comme désuètes ou marginales : mais l'ergonomie, la technologie, la sémiotique, l'herméneutique, la phénoménologie, toutes conçues comme des disciplines de la raison pratique, pourraient y trouver le lieu de rencontres nécessaires. Un de leurs premiers enjeux pourrait être de redéfinir la notion de pratique.[2]

Une sémantique des parcours développera donc une approche se réclamant d'une *herméneutique matérielle qui se concentre sur la dynamique production / réception* et sur sa manifestation et son déroulement temporel et ordonné dans le cadre du parcours. Partant, nous faisons l'hypothèse que *forme et contenu des parcours sont liés*, et que c'est au sein de cette dynamique que se mettent en place des structures actantielles où l'on cherchera à trouver des invariants et des situations prototypiques, à travers l'observation des pratiques effectives et de leur diversité.

Les différents paliers de l'analyse

Un parcours sur le Web s'apparente à un parcours de lecture et d'action dans le cadre d'une navigation hypertextuelle. Il se présente, pour un utilisateur donné, comme un cheminement régi par une série de contraintes internes (le projet général de l'utilisateur, ses compétences, les contenus visités) et externes (le dispositif technique hypertextuel qui sous-tend la navigation, les contenus proposés, leur organisation et leur présentation par les producteurs, leur accessibilité) au sein du

---

[1] On rejoint d'ailleurs ici F. Rastier, qui propose dans [Rastier 2003] une classification des « choses » en trois types : « les outils (en comprenant par là aussi les outils de communication comme les médias et les commandes, informatiques par exemple) ; les signes (linguistiques ou non : mots, symboles, chiffres, etc.) ; enfin les œuvres, qui sont issues d'une combinaison de signes ».

[2] [Rastier 2001b], pp. 212-213.

Web. Le terme de « parcours » inclut dans son étymologie des éléments qui recouvrent à notre sens les différents aspects des parcours sur Internet : 1) courir sans s'arrêter, courir en toute hâte ; 2) parcourir, traverser ; 3) parcourir du regard, lire, voir ; 4) parcourir par la parole, passer rapidement sur (un sujet), passer rapidement en revue, glisser sur, effleurer. Dynamique, transversalité, lecture multimédia et interaction sont des éléments spécifiques et fondamentaux des parcours sur le Web.

On peut décomposer l'analyse d'un parcours en cinq échelles distinctes, bien que celles-ci soient étroitement entremêlées (nous reviendrons par la suite sur ces interactions), avec au sein de chacune d'elle une spécificité de la confrontation entre l'internaute et le matériau multimédia à sa disposition.

1. *niveau micro* : une suite de pages vues.
   La sémantique des parcours prend comme palier inférieur de l'analyse de la page Web en tant qu'unité ergonomique et navigationnelle élémentaire. À ce niveau d'analyse, on s'attache à décrire le contenu de la page, que ce soit par des ressources externes ou par l'analyse de son contenu textuel et de ses propriétés (texte statique / dynamique, requête, page simple / complexe, utilisation de scripts côté client, etc.). Il est à noter ici que l'on se place, en terme de « pages vues », du point de vue de l'utilisateur sur le plan ergonomique, ce qui implique une reconstitution de la page à partir des différents éléments qui peuvent la composer, dans la mesure où une page telle qu'elle est vue peut être le résultat d'une série de requêtes. La page Web apparaît, de ce point de vue, comme un assemblage dynamique plus que comme un objet figé et clos sur lui-même.

2. *niveau mini* : une visite au sein d'un site
   Le niveau *mini* s'attache à décrire le parcours d'un utilisateur au sein d'un site donné, et, dans ce cadre, la rencontre dynamique entre l'ensemble que forme le site en termes de discours et d'unicité éditoriale et le chemin qu'y suit un utilisateur, la manière dont il appréhende, sélectionne son contenu et participe à son élaboration.

3. *niveau méso* : un parcours cohérent de lecture et d'action.
   À ce niveau d'analyse, on s'intéresse à l'articulation entre forme et contenu du parcours au sein de la temporalité de la session. Ici, une sémantique des parcours s'attachera d'une part à analyser et représenter ce qui relève de la « topologie » du parcours, à savoir les phénomènes de revisite, de détour, de retour en arrière, etc. ; d'autre part, elle étudiera l'articulation des contenus des différentes pages et sites visités au cours de la session, ce qui l'amènera à s'interroger sur les principes de cohérence alors à l'œuvre (notions d'activité intentionnelle, de projet, de surf, etc.).

4. *niveau macro* : un ou plusieurs cours d'action au sein d'une session.
   Dans le cadre borné dans le temps que constitue la session, on s'attachera à décrire ici le nombre, l'articulation, l'enchaînement ou l'entrelacement des différentes séquences homogènes de navigation. À ce niveau d'analyse, on cherchera à voir quels peuvent être les enchaînements typiques entre les

différents cours d'action (par exemple : portail de FAI[1] – WebMail – recherche), et les formes qu'ils peuvent prendre en fonction des contenus visités.

5. *niveau méga* : un parcours d'utilisateur parmi d'autres.
   La méga-sémantique des parcours replace le parcours dans le cadre de projets et de contextes d'usage particuliers qu'elle s'attache à décrire. Elle vise à découvrir et analyser des invariants au sein des différentes sessions envisagées au niveau *macro*, et examine les corrélations que ces constantes entretiennent avec des éléments extérieurs au parcours, tels que l'expérience de l'utilisateur, sa connaissance du « thème » du parcours, le fait qu'il ait déjà visité tel ou tel site auparavant, etc. À cette échelle, la sémantique des parcours observe l'articulation de l'ensemble des sessions entre elles pour un utilisateur donné, et, symétriquement, les différences et les similitudes entre sessions d'utilisateurs différents.

Ce que nous pouvons résumer dans le tableau suivant :

Tableau 1.1 - Sémantique des parcours Web : grille analytique

	niveau d'analyse du support de l'action	*niveau d'analyse de l'action*	*éléments assemblés*	*objet décrit*
micro	page	appréhension de l'interface	composition des requêtes formant les pages	contenu en termes thématique et fonctionnel
mini	site	navigation à l'intérieur d'un site	pages visitées sur le site	contenus proposés / accédés
méso	assemblage de pages sur un ou plusieurs sites	chaîne opératoire mobilisant pages et sites	une/plusieurs pages, sur un/plusieurs sites	routine de navigation
macro	session	séquence d'activité Web bornée dans le temps	groupes de pages regroupées en sites	organisation séquentielle des routines
méga	utilisateur	activité inscrite dans les pratiques, routines	les sessions d'un utilisateur	pratiques de l'utilisateur

À travers cet appareil analytique allant de la page à l'utilisateur, une sémantique des parcours dispose d'un cadre de travail pour l'analyse des usages du Web qui permet de prendre en compte l'ensemble des phénomènes en jeu du côté de la production des contenus comme de leur réception.

[1] Fournisseur d'Accès à Internet.

1.1.2 Un champ d'études encore nouveau

L'analyse de parcours sur le Web demeure un champ de recherche assez peu exploré ; elle hérite certes des travaux menés auparavant sur les hypertextes et les interactions homme-machine, mais dans le cadre du Web, elle est bien souvent réduite à l'observation de la navigation sur un site en particulier, ou à des pratiques ciblées « en laboratoire », et bien peu d'études ont pu travailler sur des données d'usage à grande échelle, et sur des utilisateurs observés en situation naturelle.

Comportement d'utilisateur

Les questions que nous soulevons ici ont peu été traitées jusqu'alors, ou sous un angle qui ne nous satisfait pas entièrement. D'un côté, les sciences cognitives ont mis en avant, au sein d'un paradigme sujet/objet, un modèle de l'activité humaine comme « système de traitement d'information » ; appliqué à la navigation, cette approche se retrouve dans les nombreux travaux issus du champ de la Recherche d'Information et de l'Intelligence Artificielle. L'approche informatique commune des parcours réduit ainsi les contenus Web à des informations, ou au mieux à un espace documentaire, dont le principal défaut est de n'être ni structuré, ni hiérarchisé.

Ces postulats se retrouvent dans la plupart de travaux sur les hypertextes, antérieurs au Web, qui portent alors principalement sur la modélisation de l'utilisateur à travers l'étude de ses parcours dans un système hypermédia donné ; les applications sont alors tournées vers les recommandations de conception et surtout la mise en place d'hypermédias adaptatifs (*adaptive hypermedia*). Nous renvoyons à la lecture de [Brusilovsky 1996] pour un panorama très complet des problématiques, des méthodes et des applications relatives aux hypermédias adaptatifs avant l'émergence du Web, complété en 2001 dans [Brusilovsky 2001] pour les études centrées sur le Web. Dans leur ensemble, les travaux décrits ne visent pas tant la description des pratiques que, à travers une modélisation du comportement, une meilleure conception des systèmes hypertextuels, en particulier dans le champ des sciences de l'éducation (application à des encyclopédies, des méthodes d'apprentissage multimédia) et de la recherche d'information, proche de l'ingénierie documentaire.

Avec l'apparition du Web, toute une série de travaux a suivi cette voie en conservant les paradigmes issus des études sur les hypermédias ; ces recherches se situent dans le champ des sciences cognitives et s'orientent vers la modélisation de l'utilisateur en situation de navigation sur le Web. Nous renvoyons à la lecture de [Modjeska 1997] pour un panorama certes un peu daté des travaux effectués dans ce domaine, mais qui rend bien compte des problématiques soulevées par cette approche, qui fait la part belle aux perceptions, aux « structures cognitives » et aux « modèles mentaux » de l'utilisateur. Dans la plupart des cas, il s'agit d'études centrées-utilisateur sur des échantillons restreints, parfois tournées vers l'« usabilité » d'un site en particulier et le problème de la « désorientation » des utilisateurs, mais le plus souvent orientées vers la recherche d'information. Ce paradigme, directement hérité de l'ingénierie documentaire, domine encore la recherche sur la navigation Web, où les contenus sont assimilés à des documents contenant des « molécules informationnelles », avec en arrière-plan une vision orientée « exécution de tâche » et résolution de problème. On trouve ainsi un certain nombre de travaux sur les

stratégies des utilisateurs en recherche d'informations (Choo sur les *knowledge workers*[1], Jansen sur les usages du moteur de recherche Excite[2]) ou encore sur l'usage de certains types de sites particuliers[3].

Ces travaux ne sont pas dénués d'intérêt, car ils permettent d'isoler, dans un contexte particulier, des questions précises sur les comportements : soulignons ici une étude particulièrement intéressante, *Web Search Behavior of Internet Experts and Newbies* menée par Hölscher et Strube en 2000 ([Hölscher & Strube 2000]), qui montre, dans un contexte de recherche d'information, l'importance de la double expertise des utilisateurs en termes de maniement des outils de recherche et de navigation sur le Web, mais aussi de connaissance du domaine sur lequel porte la recherche. Pour autant, l'approche est ici réductrice, car elle isole l'utilisateur de son cadre habituel d'activité, avec toutes les variations et perturbations que celui-ci peut comporter, et elle lui impose des activités qui ne sont peut-être pas du tout représentatives de ses pratiques.

L'étude menée par Byrne en 1999 tente de répondre à ces problèmes en proposant une approche globale de l'usage s'appuyant sur un dispositif de recueil de données original dans le champ de l'analyse de « tâches » des utilisateurs. En 1999, Byrne *et alii* proposent une *taskonomy* de l'usage du Web[4], qui rend compte de l'analyse à l'aide de la vidéo de l'activité Web de dix personnes pendant une journée, et a pour but de comprendre les tâches engagées par l'utilisateur quand il navigue au quotidien. Les participants, des utilisateurs expérimentés du Web, sont soumis à un double enregistrement vidéo, pointé sur eux et sur leur écran, qu'ils mettent en marche lorsqu'ils naviguent. Ils sont en outre invités à commenter oralement leurs actions pour faciliter le travail de dépouillement des données par la suite. L'analyse des vidéos décompose la navigation en tâches à deux niveaux de codage ; au premier niveau, les actions de base comptent huit catégories : *use information*, *locate information*, *provide information*, *find on page*, *navigate*, *configure browser*, *manage window* et *react to environment* ; chacune de ces catégories se décompose ensuite en sous-catégories plus fines. Le résultat majeur de l'étude est l'imbrication des tâches entre elles : une tâche peut générer n'importe quel autre type de tâche, et la plus fréquente tâche, *Use Information*, génère d'autres tâches du type *Locate Information*, *Navigate* et *Find On Page*. En outre, l'étude fournit des observations avancées sur certaines tâches :

- dans la tâche *Use Information*, décomposée en *reading*, *print*, *duplicate*, *view*, *listen*, et *download*, la sous-tâche la plus fréquente est *reading*, ce qui replace la lecture au centre de la navigation et de l'appréhension des contenus ;
- pour la tâche *Locate Information*, le moteur de recherche s'impose comme le point de départ privilégié ;

---

[1] Voir [Choo *et al.* 1999] et [Choo *et al.* 2000].

[2] Voir [Jansen *et al.* 1998a] et [Jansen *et al.* 1998b].

[3] Par exemple, [Jones *et al.* 1998] sur les bibliothèques électroniques.

[4] Voir [Byrne *et al.* 1999a] et [Byrne *et al.* 1999b].

- pour *Find On Page*, le sous-type le plus fréquent est de loin *related*, par rapport à *image*, *interesting*, *string* et *tagged* ; mais en durée, les cinq sont équilibrés.

Enfin, les auteurs observent que dans la manière d'accéder aux pages, le lien hypertexte fait plus de la moitié des requêtes, tandis que les actions de type *back* et *autre* se partagent le reste.

L'entreprise taxinomique de Byrne est d'autant plus intéressante qu'elle se fonde sur des observations de la vie « de tous les jours », et qu'elle tente d'embrasser la diversité des pratiques. Toutefois, nous ferons à cette étude la même critique qu'à l'approche cognitiviste des parcours sur le Web : contenus proposés, modes d'accès et activité de navigation sont enfermés dans le paradigme de la recherche d'information et, par extension, les contenus du Web sont valorisés sous cet angle unique. En arrière-plan, se dessinent les approches mentalistes issues des sciences cognitives qui réduisent les parcours sur le Web à la réalisation d'un projet par un sujet à l'aide de l'outil technique que constitue le Web. Ce faisant, l'approche cognitive conclut à des équivalences entre motifs de navigation, tâche et motivation de l'utilisateur qui sont réductrices dès lors que l'on examine la diversité de l'offre de contenus sur le Web autant que les usages qui en sont faits.

Dès lors, nous laisserons volontiers de côté ces approches orientées modélisation pour notre analyse, et y opposons une approche descriptive pragmatique qui s'attache à replacer les modes de navigation dans le cadre de pratique avérées, et à prendre en compte la singularité des situations, des contenus et des individus. On cherchera certes des invariants dans l'ensemble des parcours observés, mais sans les relier à de quelconques modèles mentaux ou psychologiques. Pour cela, nous empruntons à Leroi-Gourhan[1] la notion de « chaîne opératoire », définie comme un processus de travail qui mène d'une matière première à un objet fini. Se décomposant en une série d'étapes, la chaîne opératoire intègre un projet, un savoir-faire, un geste, une matière première, un outil ; elle s'articule et s'imbrique avec d'autres chaînes, qu'elle peut croiser et influencer. Appliqué plus spécifiquement aux parcours sur le Web, ce concept de chaîne opératoire permet de rendre compte à la fois des aspects techniques liés au maniement de l'outil informatique en général et du Web en particulier, de l'importance des connaissances et du savoir-faire de l'utilisateur, et de l'implication de ces deux éléments dans des « projets » qui sous-tendent la navigation. Par opposition à la tâche, le projet engage et construit les savoir-faire, peut être décomposé en plusieurs chaînes opératoires, et ne cesse de s'élaborer, de se redéfinir et de s'alimenter au fur et à mesure qu'il se réalise. L'ambition d'une sémantique des parcours sera dès lors de mettre à jour des chaînes opératoires récurrentes, et d'examiner les structures, les savoir-faire, l'outillage et le déploiement temporel et rythmique des mouvements et de la gestuelle navigationnels.

Web Mining et données de trafic

Avec le développement du Web tant du côté de l'offre de contenus que de l'accès, s'est constitué depuis la fin des années 90 un champ de recherche autour du « Web

---

[1] [Leroi-Gourhan 1943] et [Leroi-Gourhan 1964].

Mining ». Plutôt orienté vers l'analyse de données, les méthodes statistiques et les aspects applicatifs, le Web Mining se divise en trois domaines : Web Content Mining pour l'analyse des contenus, Web Structure Mining pour l'étude globale de l'organisation hypertextuelle de la Toile, et Web Usages Mining sur le plan des usages[1]. Dans ce dernier, qui a fait des données de trafic son matériau privilégié, on distingue communément les approches centrées-serveur (*site-centric*), qui traitent de données recueillies sur un site particulier, et centrées-utilisateur (*user-centric*) qui se basent sur des informations collectées du côté de l'utilisateur.

Le travail que nous présentons ici, basé sur l'analyse de traces de navigation collectées sur les postes des internautes, se place résolument dans cette dernière approche. Bien qu'elle soit la seule approche appropriée pour l'analyse des usages, elle est difficile à mettre en œuvre, et demeure peu pratiquée : très peu d'études ont comme matériau des données de navigation enregistrées du côté de l'utilisateur en situation d'usage réelle – nous en comptons au total quatre.

La première d'entre elles est celle de Catledge et Pitkow en 1995, *Characterizing browsing strategies in the World-Wide-Web*[2]. Travail fondateur et unique, cet article présente l'analyse de données recueillies au niveau du navigateur M*osaic* pendant trois semaines auprès de cent sept utilisateurs. Celles-ci contiennent non seulement les URL visitées par l'utilisateur et la date, mais aussi l'ensemble des actions sur le navigateur ; l'ouverture d'une page est ainsi décomposée en :

- *Selection of hyperlink in document*
- *Go back one document*
- *Open file via a URL*
- *Go to document via Hotlist*
- *Go forward one document*
- *Open local file*
- *Go to the Home document*
- *Go to document via Window History*

L'étude montre un certain nombre de résultats intéressants :

1. Découpage en sessions : le temps moyen entre deux actions sur le navigateur est de 9,3 minutes. Partant, le temps d'inactivité retenu pour définir la fin d'une session est de 25,5 minutes.
2. Séparation par protocole : 80 % de la navigation se fait par HTTP, dont 4 % générés par CGI (contenu dynamique).
3. Méthode d'interaction : 52 % des actions de navigation sont faites par le suivi de liens dans les pages, et 41 % par des *Back*. Les raccourcis clavier ne sont presque jamais employés.
4. Séquences répétées : une corrélation linéaire est observée entre le nombre moyen de pages vues sur un site dans une visite et le nombre de visites du

---

[1] Pour une vue générale et synthétique du champ du Web Mining, on trouvera dans [Kosala & Blockeel 2000] et dans [Srivastava *et al.* 2003] des informations claires et une sélection d'articles.

[2] [Catledge & Pitkow 1995].

site : beaucoup de sites sont vus sur une faible longueur, peu sur un longueur importante.

5. À l'intérieur d'un site, les auteurs constatent une stratégie *spoke and hub*, c'est-à-dire une très forte utilisation du *Back* pour des séries d'avant-arrière autour d'une page-pivot. Ceci suggère que cette forme de navigation est indépendante du nombre de liens proposés dans une page.
6. Autre méthode de navigation souvent observée : les pages personnelles comme une sorte d'index vers des pages intéressantes.

Ces résultats commencent à être un peu anciens en regard de l'évolution du Web et de la diffusion de son accès ; cela étant, les conclusions et la méthode n'en sont pas moins intéressantes, en particulier dans la capacité à lier les pages vues à des modes d'interaction sur le navigateur.

À la même époque, Crovella et Bestavros publient *Characteristics of WWW client-based traces* en 1995 avec Cunha[1], et *Self-similarity in World Wide Web traffic - Evidence and possible cause* en 1996[2]. Les deux articles présentent l'analyse de traces de navigation côté-client sur près d'un million de requêtes ; comme dans [Catledge & Pitkow 1995], le navigateur *Mosaic* sert de support au recueil de données : 37 ordinateurs partagés (stations de travail Sun) sont équipées du dispositif de recueil dans des salles d'une université d'informatique. Entre autres résultats intéressants, nous notons le constat que le trafic répond à une loi de puissance (*power-law distribution*) : une faible part des documents concentre la majorité des requêtes, tandis qu'un grand nombre de pages ne sont vues que très rarement. Cette loi est également observée en ce qui concerne la taille des fichiers, ce qui intéresse particulièrement les deux auteurs dont la perspective est d'améliorer les techniques de *cache* afin d'augmenter la rapidité d'accès aux pages Web.

La troisième étude, présentée en 1997 par L. Tauscher et S. Greenberg[3] se penche sur la « revisite » de pages Web : l'objectif est d'élaborer des modèles de revisite, et d'en tirer des conclusions pour la conception des systèmes d'historique des navigateurs. Les données analysées sont constituées par les traces de navigation de 23 utilisateurs observés pendant six semaines. L'analyse montre qu'en moyenne 58 % des requêtes pour un utilisateur donné pointent vers une page qu'il a déjà visitée, en même temps que le « vocabulaire » des URL ne cesse de croître avec le temps. Des entretiens ont été menés par la suite avec les utilisateurs, qui montrent que les causes de la visite de nouvelles pages sont essentiellement : 1) le besoin de nouvelles informations ; 2) le désir d'explorer un site en particulier ; 3) la page est recommandée par un collègue, et 4) la page a été trouvée en cherchant autre chose. L'étude croise également ces données avec un enregistrement des actions sur le navigateur similaire à celui pratiqué dans [Catledge & Pitkow 1995].

---

[1] [Cunha *et al.* 1995].

[2] [Crovella & Bestavros 1996].

[3] Voir [Tauscher & Greenberg 1997a] et [Tauscher & Greenberg 1997b].

Enfin, en 2000, Cockburn et McKenzie présentent *What do Web users do ? An empirical analysis of Web use*[1]. La méthode de recueil de données consiste à utiliser les fichiers *history.dat* et *bookmark.html* du navigateur Netscape pour 70 utilisateurs (membres de la faculté) sur 4 mois (d'octobre 1999 à janvier 2000) ; pour chaque page, sont notés, outre l'URL, les dates de premier et de dernier accès et le nombre de visites, avec une collecte chaque jour. Dans les données récupérées, n'apparaissent que les URL explicitement demandées par l'utilisateur, les paramètres des CGI se trouvent tronqués, et les différentes pages des *frameset* sont regroupées sous une seule entrée. Cockburn et McKenzie font plusieurs constats :

- croissance assez régulière du vocabulaire (les pages) dans le temps, de manière globale ainsi que pour chaque utilisateur.
- forte corrélation entre le nombre de visites et le vocabulaire. En moyenne, quatre visites pour une page (« for each new URL added to the overall vocabulary, four pages are revisited »).
- taux de revisite : sur l'ensemble du panel, le taux de revisite est de 81 %.
- pour chaque utilisateur, peu de pages sont visitées très régulièrement (en moyenne, 24 % du vocabulaire de chaque individu) : on retrouve un comportement de type « loi de Zipf ». Souvent, les utilisateurs ont des raccourcis vers les deux pages les plus vues (*home page* ou *bookmark*).
- la majorité des pages sont vues une seconde ou moins : « browsing is a *rapidly* interactive activity ».
- *bookmarks* : ils sont nombreux chez tous les sujets. Dans le temps, le nombre d'ajouts est supérieur au nombre de suppressions, de sorte que la taille des *bookmarks* ne cesse d'augmenter.
- les utilisateurs, bien qu'appartenant au même département de l'université, voient des espaces très différents sur le Web : 91 % des pages visitées hors du site de l'université n'ont été vues que par un seul utilisateur.

Les auteurs concluent sur les implications de ces résultats pour la conception de navigateurs, ce qui ne nous intéresse pas directement ici, mais leurs résultats empiriques sont tout à faire réutilisables pour l'analyse des parcours.

Ces études sont très précieuses, en premier lieu parce qu'elles proposent des pistes pour l'exploitation de données de trafic, ce qui nous intéresse particulièrement étant donné notre matériau, mais également parce qu'elles traitent de la navigation du côté de l'utilisateur et en situation « réelle ». Le faible nombre de travaux « centrés-utilisateur » tient à la rareté des données de ce type, et ces quatre travaux fournissent d'importants résultats en termes statistiques et méthodologiques. Ils esquissent une image de la navigation ouvrant l'utilisateur à un nombre toujours renouvelé de pages, au sein d'une pratique complexe où les fonctionnalités des navigateurs autant que les contenus des pages entrent en ligne de compte. Deux reproches peuvent toutefois leur être faits : d'une part, les données concernent toujours des catégories particulières d'utilisateurs (étudiants en informatique le plus souvent), et pour une durée relativement limitée (quatre mois au maximum). D'autre part, le contenu des pages n'est jamais abordé : il semblerait particulièrement intéressant de savoir quelles corrélations peuvent exister entre stratégies de navigation, visite de nouveaux sites et

[1] [Cockburn & McKenzie 2000], repris dans [McKenzie & Cockburn 2001].

thème ou service proposés par les pages accédées. Il entrera donc dans les objectifs d'une sémantique des parcours d'avoir une approche dynamique tenant dans un même temps les logiques de production et de réception.

1.2 Au croisement de deux dynamiques

Le travail que nous présentons ici se base principalement sur l'analyse de traces de navigation recueillies du côté de l'utilisateur. Dès lors, à l'enjeu descriptif des modes de parcours, s'ajoute celui de l'exploitation des traces d'usage : pour remplir les objectifs d'une sémantique des parcours, nous devons à la fois être capables de décrire finement leur double contexte, celui des contenus comme celui de l'internaute, mais également donner sens à ce mouvement en intégrant ces descriptions au sein de la dynamique du parcours.

La distinction méthodologique en paliers que nous avons proposée ainsi que leur ordre de présentation ne doivent en rien masquer les interactions fortes qui existent entre les cinq échelles d'analyse : de la même manière que la Sémantique Interprétative pose la détermination du local (thématique, dialectique, dialogique, tactique) par le global (genres, discours, pratiques), la sémantique des parcours que nous proposons postule la double primauté du projet et du savoir-faire de l'utilisateur (niveaux *macro* et *méga*) sur les contenus visités et leur articulation (niveaux *micro*, *mini* et *méso*). Entre ces cinq paliers d'analyse, se jouent des influences réciproques qu'une sémantique des parcours se doit d'explorer et d'expliciter. À titre d'exemple, on peut citer :

- *méga vers micro* : les centres d'intérêt de l'utilisateur définissent la thématique et la fonction des pages visitées. L'expertise ergonomique de l'internaute, sa connaissance de la sémiotique des pages Web modifie son appréhension des pages. Certaines études[1] ont ainsi montré chez les primo-accédants, la confusion entre adresses Web et adresses de messagerie, ou que les bannières publicitaires ne sont pas identifiées en tant que telles, ce qui induit des incompréhensions et des détours au sein du parcours.
- *macro vers micro* : l'appréhension du contenu d'une page est elle-même fonction de la position de la page dans la session, et de la façon dont elle s'inscrit dans le projet en cours : la lecture d'un contenu donné met en jeu un processus interprétatif qui dépend de l'utilisateur et de la chaîne opératoire dans laquelle il se place. Par exemple, un site comme celui de la Fnac peut être appréhendé dans sa fonction première de catalogue de vente en ligne, mais aussi comme répertoire d'opinions d'internautes sur un produit.
- *méso vers méga* : les contenus et services proposés par un site peuvent amener l'utilisateur à des pratiques régulières. Par exemple, un site fournissant le programme télévisé va plaire à l'utilisateur, qui y reviendra fréquemment.

---

[1] Voir par exemple [Relieu & Olszewska 2004] ou [Cotte 2002].

- *méso vers mini* : l'ensemble du contenu du site a une incidence directe sur le niveau *mini* (un site ne peut pas proposer ce qu'il n'a pas). Les chemins des visiteurs au sein d'un site sont susceptibles d'entraîner, à court ou moyen terme, une réorganisation de ce site ; ceci est particulièrement le cas pour les sites marchands qui souhaitent améliorer leur navigabilité.
- *méso vers macro* : la propension des sites à pointer vers d'autres sites, ainsi que le type de sites vers lesquels ils renvoient, ont une influence sur l'ensemble des sites visités dans la session : moteurs de recherche, sites personnels avec page de liens et site commerciaux ont à cet égard des influences très contrastées.
- *micro vers méso* : le contenu peut renvoyer à des activités hors Web, que ce soient encore des activités sur Internet (mail, *chat*, etc.) ou dans de tout autres domaines et sur de tout autres supports (le Web comme ressource d'informations). Un site peut proposer à l'utilisateur des informations sur la navigation, les outils du Web, ou des services d'aide à la navigation qui auront une influence sur son comportement général.
- *micro et mini vers méso et macro* : le type de pages et de sites visités contraignent à telle ou telle forme de navigation (par exemple : l'utilisation du WebMail passe par une authentification de l'utilisateur). Le parcours se trouve encadré dans une forme de schéma actantiel défini par le site qui guide son parcours. Les liens présents sur une page définissent des possibilités de navigation hypertextuelle. Certaines pages contiennent des scripts qui opèrent des actions d'ordre navigationnel : redirection, ouverture automatique d'une ou plusieurs fenêtres, etc. La nécessité pour l'utilisateur de disposer de *plugins* (briques logicielles additionnelles) au sein du navigateur, ou même d'un navigateur particulier, peut lui interdire l'accès à certaines pages ou certains sites.

Il ne suffit pas, bien évidemment, d'énumérer ces influences réciproques entre différents paliers, il importe de les quantifier, de les ordonner, et d'évaluer les cas où tel élément prend le pas sur tel autre et de quelle manière. Notre travail vise à décrire ces interactions, ainsi qu'à évaluer dans quelle mesure leur connaissance peut permettre d'élaborer un système d'analyse des parcours sur le Web en vue de l'étude des usages d'Internet. Dans la mesure où l'on traitera de données volumineuses pour y chercher des phénomènes récurrents, on cherchera à se doter de représentations synthétiques aux différents niveaux d'analyse des parcours, de la page à l'utilisateur.

1.2.1 Décrire les contenus

À l'échelle de la page, une sémantique des parcours doit prendre en compte la nature spécifique des contenus Web. Dans la littérature, ceux-ci sont très généralement considérés en termes d'informations ou de connaissances : c'est de cette manière que Vennevar Bush imaginait en 1945 le système de nature hypertextuelle qu'il baptisa Memex : « Un memex est un appareil dans lequel une personne stocke tous ses livres, ses archives et sa correspondance, et qui est mécanisé de façon à pouvoir être consulté de manière très rapide et très flexible. Il s'agit d'un

supplément agrandi et intime de sa mémoire. »[1]. Aujourd'hui encore, on retrouve l'assimilation des contenus Web à de simples molécules informationnelles au sein du projet du Web Sémantique, que son initiateur Berners-Lee définissait ainsi en 2001 : « Le Web Sémantique est une extension du Web actuel, où l'information a un sens bien déterminé, permettant une meilleure coopération entre les machines et les individus »[2].

Cette conception traverse la majorité de la littérature sur l'hypertexte, on la retrouve par exemple chez Balpe en 1996 :

> Mais c'est plus en termes de finalités que l'hypertexte se comprend […]. En effet, on ne peut oublier que l'hypertexte permet avant tout de mettre les capacités de calcul et de présentation d'un ordinateur au service de l'information structurée ou non, en réalisant des associations entre des éléments de nature différente, associations conduites par l'intelligence ou l'intuition de l'utilisateur.[3]

On la retrouve également dans l'approche documentaire du Web, ce qu'illustre notamment l'ensemble des interventions présentées dans le colloque en ligne Text-e organisé par la BPI en 2001, en particulier dans [Chartier 2003] ou dans [Broadbent & Cara 2003] ; elle se retrouve également dans l'approche pourtant orientée vers l'ethnométhodologie de [Ghitalla *et al.* 2003], où le Web est envisagé en termes d' « architecture documentaire numérique » (terme repris à Broadbent et Cara), et la dimension interactive des pages centrée sur la manipulation des interfaces[4].

Cette définition des contenus du Web en termes d'informations et de connaissances nous paraît doublement réductrice. Tout d'abord, si le Web, en tant que système hypertextuel, s'apparente dans certains cas à une collection de textes ou de documents multimédia, cela n'implique pas que l'on puisse assimiler leur contenu à la mise en forme d'informations ou de connaissances : le Web n'est pas simplement un objet hybride entre encyclopédie déstructurée et annuaire. Plus encore, le paradigme du document nous paraît insuffisant pour décrire la diversité des contenus accessibles sur le Web : outils de communication (WebMail, WebChat, etc.), espaces coopératifs, jeux en ligne, sites communautaires, tous ces éléments invitent à

---

[1] « A memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory. » ([Bush 1946]).

[2] « The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation » ([Berners-Lee *et al.* 2001]).

[3] [Balpe *et al.* 1996], p.18.

[4] Par exemple : « On pourrait croire, parfois que le document à l'écran revisite à sa façon les deux procédés essentiels qui ont façonné l'histoire des supports d'écriture : le *scrolling* épouse le principe du déroulement du parchemin, le tourne pages électronique rappelle l'agencement des pages d'un livre. L'écran lui-même peut épouser les contours d'une page autonome, la « page écran » faisant alors coïncider espace d'affichage et géographie du document. Mais ni l'écran, ni même parfois l'espace de la « page » comme ensemble signifiant, ne sont la mesure de l'activité. C'est d'abord de la maîtrise du fenêtrage que dépend cette dernière […]. » ([Ghitalla *et al.* 2003], p. 164).

envisager également la Toile comme un fournisseur de services et plus généralement d'outillage. L'attention particulière des travaux sur les écrits numériques portée aux interfaces Web, aux règles complexes de leur composition, aux différentes zones et leur statut sémiotique, à la valorisation des textes insérés dans ce contexte, semble faire presque oublier le « travail des serveurs » et, pourrait-on dire, ce que « le clic renvoie ». C'est dans cette perspective qu'on regrettera que dans les réflexions de Souchier ou Jeanneret sur les « écrits d'écran »[1], l'attention soit tellement portée sur la sémiotique des pages Web qu'elle en néglige d'entrer pleinement dans les éléments interactifs, et ne s'attache qu'à ce qui relève du texte et de la lecture ; certes, il est toujours question de lecture quels que soient les contenus, mais ce palier ne réduit ni n'explique la navigation elle-même. À titre d'exemple, lorsqu'un internaute joue aux échecs sur le site Yahoo, il lit la page, la position des pièces, l'horloge, les classements, mais avant tout, il joue aux échecs. L'appréhension des contenus Web se base donc sur la lecture, mais ne s'y réduit pas toujours, et une sémantique des parcours doit se doter d'une représentation des contenus qui tienne pleinement compte des types de contenus, des types d'activités qu'ils suggèrent et autorisent et des interactions avec les fournisseurs de contenus.

Dans cette optique, on se penchera avec attention sur les études visant à distinguer, selon l'approche retenue, des genres ou des types de pages et de sites Web à partir de corpus. Au sein de l'approche générique, on trouve en particulier les études présentées dans [Karlgren & Cutting 1994] et [Karlgren *et al.* 1998], [Dimitrova *et al.* 2002], [Rehm 2002], [Roussinov *et al.* 2001] et [Dillon & Gushrowski 2000]. Si les divisions génériques varient selon les auteurs, tous partagent le fait de travailler sur des corpus de pages et de chercher à caractériser leurs genres sur la base de leur contenu textuel mais aussi de traits structurels et présentationnels ; les traits retenus donnent ainsi une place importante 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. Si l'on peut discuter les choix faits pour fonder les distinctions génériques au sein de la production Web, ces études ont l'avantage de prendre en compte la spécificité des contenus dans leur dimension hypertextuelle et fonctionnelle (par exemple : Karlgren distingue un genre « FAQ »).

L'approche typologique propose une démarche plus inductive, tout en conservant les aspects fonctionnels des classes constituées : dans *The connectivity sonar*[2], 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-là[3]. Dans une perspective

---

[1] Voir notamment [Jeanneret & Souchier 1999] et [Souchier 2000].

[2] [Amitay *et al.* 2003].

[3] « 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

différente, et plus large en ce qui concerne les traits retenus pour décrire les sites et les pages, les projets TypWeb et SensNet suivent une démarche similaire dans la description des contenus. L'objectif est de parvenir « 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. »[1]. Le projet s'appuie sur l'extraction de l'ensemble des éléments textuels, structurels et formels des pages et des sites, pour leur analyse à l'aide de traitements matriciels. 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. Il a ainsi mis en avant une distinction entre pages « à contenu » et pages « d'orientation », qui s'avère précieuse pour l'appréciation de la dynamique des parcours.

En somme, rejetant le paradigme informationnel ou documentaire des contenus Web, et privilégiant une approche issue de la linguistique de corpus et adaptée aux spécificités du Web, nous appuierons une analyse des contenus au sein des parcours sur la Sémantique Interprétative telle qu'elle a été définie par François Rastier[2]. En tant que textes, les pages Web s'inscrivent dans des pratiques d'écriture, des discours, des genres, et incluent en leur sein les dimensions thématique, dialectique, dialogique et tactique. Cela étant, comme nous l'avons déjà dit, nous estimons que cela n'est pas suffisant dans le cadre de la publication et des contenus Web, et qu'il est nécessaire de considérer également le Web en termes d'outillage. Pour répondre à ces objectifs, on tentera de se doter de représentations à partir de sources endogènes, par la constitution de corpus notamment, et exogènes adaptées, à l'aide des annuaires du Web.

En outre, pour compléter cette approche, l'analyse des parcours gagnerait à inclure dans ses représentations une vision globale du Web en termes d'interconnexion : nous retenons notamment la distinction faite dans [Broder *et al.* 2000], sur la base des liens hypertexte entre pages, entre un fort noyau très interconnecté, une zone qui mène à ce noyau mais à laquelle on accède difficilement, et à l'inverse une zone pointée par le noyau central mais dont il est difficile de sortir. De telles observations sont particulièrement intéressantes : dans une navigation de site en site, l'impossibilité de « sortir » d'un site ou au contraire l'impossibilité d'aller sur un site car aucun lien n'y mène sont des facteurs importants. Ramenées aux cheminements des internautes de site en site et aux pratiques observées dans la durée, ces éléments peuvent permettre d'expliquer la régularité des comportements et la construction de territoires personnels et de routines sur le Web.

---

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. »

[1] [Beaudouin *et al.* 2001].

[2] Voir [Rastier 1987] et [Rastier 1989].

1.2.2 Dynamique des parcours et des individus

La séquentialité est un aspect essentiel de la sémantique des parcours : de la même manière qu'un texte ne peut être considéré, sinon au prix d'une réduction importante de ses composants sémantiques, à un « sac de mots » ou de phrases, un parcours sur le Web ne saurait être réduit à une collection de pages. L'ancrage temporel des parcours conforte cette position ; mais plus encore, le parcours s'apparente à une série d'actions (de navigation) dont l'ensemble ordonné seul peut donner le sens. Cette question est perceptible dans bien des travaux : les réflexions autour de la notion de tâche, que l'on retrouve en particulier dans les travaux de Byrne *et alii* ([Byrne *et al.* 1999a] et [Byrne *et al.* 1999b]) et la *taskonomy* qu'ils proposent, les travaux sur les comportements d'utilisateurs en recherche d'informations ([Choo *et al.* 1999]), les études sur la revisite de pages ([Tauscher & Greenberg 1997a] et [Cockburn & McKenzie 2000]) ou les recherches plus générales sur la navigation Web ([Catledge & Pitkow 1995], [Huberman *et al.* 1998]). Cela étant, l'analyse du parcours comme suite ordonnée de pages est dans presque tous les cas réduite au nombre de pages vues sur un site, ou de sites différents visités.

Pour répondre à ce problème, on se tournera volontiers vers les études centrées-serveur, qui font au contraire la part belle à la recherche de motifs de navigation au sein des parcours sur un site donné. Une littérature relativement abondante traite de l'analyse des parcours d'utilisateurs d'un point de vue *site-centric*, sur la base de l'analyse des *logs* des serveurs Web. Un tel engouement s'explique par les enjeux économiques sous-jacents à ces recherches : les sites à vocation commerciale souhaitent disposer de données les plus précises possibles sur leur fréquentation, afin de savoir quelles pages sont les plus visitées, comment les utilisateurs y arrivent, et comment les faire « rester » plus longtemps sur le site. Nous renvoyons à la lecture des articles de Masand en 2000[1] et de Srivastave en 2003[2] pour un panorama de ces recherches. Si le point de vue centré-serveur ne correspond pas aux objectifs d'une sémantique des parcours centrée sur les pratiques d'utilisateurs, certaines méthodes employées pour l'analyse des visites de sites particuliers peuvent être mobilisées.

Dans ce cadre, trois études retiennent particulièrement notre attention : en premier lieu, Borges et Levene proposent en 1998[3] une modélisation des parcours sur un site sous forme de Grammaire Probabiliste Hypertextuelle (HPG : *Hypertext Probabilistic Grammar*) qui permet de fouiller des bases de parcours et d'identifier des séquences récurrentes suivies par les internautes. À la même époque, et poursuivant les mêmes objectifs, Srivastava *et alii*[4] présentent le logiciel *WebMiner*, qui applique des techniques de *data mining* à l'étude des usages du Web ; en complément, le système *WebSIFT* doit permettre d'identifier les motifs d'usages les plus intéressants *via* l'analyse du contenu et de la structure d'un site. Enfin,

---

[1] [Masand & Spiliopoulou 2000].

[2] [Kosala & Blockeel 2000].

[3] Voir [Borges & Levene 1998], ainsi que [Borges & Levene 1999] et [Borges & Levene 2000].

[4] [Cooley *et al.* 1997], [Cooley *et al.* 1999a] et [Mobasher *et al.* 2000].

Spiliopoulou, Faulstich et Winkler présentent en 1999 un outil, *Web Usage Miner*[1], capable d'agréger sous forme d'arbre les différentes navigations suivies au sein d'un site. Un langage proche du SQL, *MINT*, permet de fouiller, sous forme de requête, dans la base des parcours effectués sur un site et de connaître la probabilité qu'un utilisateur voie une page étant donné les autres pages vues avant ou après. De manière générale, ces outils et méthodes centrés-serveur tiennent souvent pour acquis que le contenu des pages est connu, et appliquent des méthodes d'analyse statistique sur des séries de symboles représentant les pages. Cela étant, ces approches sont intéressantes en ce qu'elles traitent réellement de l'aspect séquentiel et parfois même temporel des parcours, et permettent d'identifier des motifs de navigation pertinents. Ce qui leur manque avant tout, c'est une connaissance et un suivi des utilisateurs ; certains proposent des méthodes pour tenter de les deviner, comme [Murray & Durrell 1999] qui fait correspondre informations socio-démographiques et centres d'intérêt. Dans [Chevalier *et al.* 2003], les auteurs proposent d'identifier, au sein des visiteurs d'un site donné, des profils socio-démographiques distincts et des modes de navigation correspondants. Toutefois, le positionnement structurel de ces travaux du côté des sites les rendent inaptes à dépasser des corrélations locales entre certains types de sites et certaines variables relatives à l'utilisateur, et leur interdit d'embrasser la diversité des pratiques. Une sémantique des parcours ne saurait se satisfaire des résultats obtenus dans ce cadre en termes d'usages, et appelle une méthodologie centrée sur l'internaute qui prenne en compte sa dynamique d'usage dans l'analyse.

Les expériences menées en psychologie cognitive et tournées vers la modélisation du comportement des utilisateurs rentrent dans ce cadre, mais ne nous satisfont pas pour les raisons que nous avons déjà évoquées, et c'est bien plutôt vers la sociologie des usages que nous chercherons des éléments de description globaux et de contextualisation des parcours. Ainsi que le rappelle Jouët dans son *Retour critique sur la sociologie des usages*, dans tous les travaux de ce champ de recherche, « l'usage est analysé comme un construit social. […] La sociologie des usages, à l'opposé de la problématique de la traduction, n'étudie pas tant l'amont que l'aval, c'est-à-dire l'usage resitué dans l'action sociale. La construction de l'usage ne se réduit dès lors pas aux seules formes d'utilisation prescrites par la technique qui font certes partie de l'usage, mais s'étend aux multiples processus d'intermédiations qui se jouent pour lui donner sa qualité d'usage social. »[2] L'auteur distingue quatre problématiques principales, quoique souvent entremêlées au sein des études de cas, qui traversent le champ de la sociologie des usages :

- la généalogie des usages met en parallèle l'évolution des outils techniques et leur insertion dans les pratiques et les équipements existants ;
- la question de l'appropriation s'attache à décrire la construction des usages par les individus, les situations interactionnelles induites par l'objet

[1] [Spiliopoulou *et al.* 1999].

[2] [Jouët 2000], p. 499.

technique, et la dimension de construction de l'identité personnelle et sociale induite par les TIC ;
- le questionnement sur le lien social vise l'élaboration ou la modification des liens interpersonnels et des collectifs à l'aide des TIC ;
- enfin, la question des rapports sociaux prête attention au fait que les TIC s'insèrent dans des rapports sociaux et que, en tant qu'objets symboliques, ils constituent des enjeux de pouvoir.

Que retiendrons-nous de ces éléments pour l'analyse des parcours ? La majorité des études menées dans ce champ adoptent des méthodologies plus qualitatives que quantitatives : entretiens, observations, enregistrements vidéo, carnets de correspondants, etc. Comme le note Jouët, « si seule l'approche qualitative peut tenter de dégager la signification des actes de communication au niveau individuel et le sens social des usages auprès de groupes sociaux spécifiques, la démarche quantitative se révèle riche pour donner à l'usage une dimension plus macrosociale »[1]. En travaillant sur des données de trafic, on ne se trouve totalement ni dans l'une, ni dans l'autre des deux perspectives : la finesse de ce type de matériau autorise des analyses très précises sur les modes de navigation, bien qu'elle les désincarne, tandis que la technicité du matériau autorise des analyses globales et des croisements statistiques pour faire émerger des phénomènes récurrents à grande échelle.

Ne perdant pas de vue qu'elle s'intéresse principalement aux parcours « dans » l'écran et à l'activité de navigation, une sémantique des parcours cherchera ainsi à trouver des éléments d'explication des comportements en les rattachant à un utilisateur, mais dans une perspective toutefois endogène : les caractéristiques des parcours Web d'un individu sont sans doute corrélées statistiquement avec des variables socio-démographiques, mais c'est dans le contexte plus précis de la vie de l'utilisateur sur le Web, de ses parcours passés et de son corpus de sites et de pages que l'on peut faire surgir une signification de la navigation perçue comme activité située.

En nous centrant sur l'activité de navigation envisagée comme chaîne opératoire et réalisation d'un projet, nous nous heurtons, avec les données de trafic dont nous disposons, au problème de la reconstruction *a posteriori* d'une intentionnalité de l'utilisateur. Avec ce type de matériau, cette dimension – problématique en elle-même – échappe totalement à notre regard, et la question du projet qui sous-tend le parcours demeure indécidable. Pour répondre à ce type de questions, il parait bien plus approprié de mener des entretiens avec les utilisateurs, de les observer directement ou de leur présenter des parcours qu'ils ont faits en leur demandant de les commenter et d'en expliciter la logique. Figure énigmatique, l'utilisateur sera pour nous à la fois fuyant, car inaccessible, et omniprésent, car principe d'existence des parcours, de sorte que nous ne le considérerons pas tant sous l'angle de ce qu'il est (son âge, sa profession, etc.) que de ce qu'il fait (sa navigation).

C'est alors sur la notion de *territoires personnels* que peut s'appuyer l'appréhension du sens des parcours. Les données de trafic dont nous disposons, enregistrées sur une longue période, permettent de rapporter la visite d'un site au

---

[1] [Jouët 2000], p. 514.

corpus individuel de sessions et d'espaces Web de l'individu, et de la valoriser en regard des visites précédentes. Nous l'avons dit précédemment, une page peut être appréhendée de manière différente par un même individu dans deux contextes distincts ; corrélativement, la primauté du global sur le local nous invite à penser que l'appréhension d'une page ou d'un site est fortement déterminé par l'ensemble des visites précédentes du site ou de la page. Au fil du temps, de l'expérience à l'usage, la pratique individuelle délimite au sein de la Toile un territoire où se distinguent le routinier, l'habituel et l'exceptionnel, et où se dessinent des modes d'activité, des comportements et des temporalités distincts. Sur la base d'un corpus volumineux de traces d'activités en situation naturelle, l'analyse de la structure des ces territoires, de leur contenu et des formes de parcours qui y sont liés alimente, en quelque sorte, une éthologie de la navigation sur le Web.

Conclusion

Le Web n'est pas seulement un lieu de lecture mais également d'activité, et la navigation n'est pas tant un parcours de lecture qu'un régime d'action particulier pouvant supporter différents types d'activités : lecture, écriture, jeu, communication, etc. Le parcours sur le Web s'apparente alors à la rencontre dynamique entre un utilisateur et des contenus, à un moment et dans une chaîne opératoire donnés. Les enjeux d'une analyse des parcours sur le Web ne se déterminent alors plus en termes d'accès à l'information, mais amènent à examiner dans quel régime d'action se situe l'utilisateur à un moment donné en fonction de ses routines, ses motivations (ce qui vient de l'utilisateur, cause exogène) et de ce qui est disponible sur le Web (nature du contenu, cause endogène). La navigation Web s'apparente ainsi à un objet complexe dont l'analyse nécessite un appareillage méthodologique spécifique à même de prendre en compte les dynamiques qui s'y jouent ; le travail que nous présentons ici entend, sur la base de données de trafic centrées-utilisateur, explorer des pistes pour y parvenir.

L'ambition générale d'une sémantique des parcours étant une description la plus fine possible des parcours sur le Web, celle-ci affronte un double problème : elle doit parvenir à remplir la tâche qu'elle s'est fixée aux différents paliers de l'analyse, ce qui implique de parvenir à caractériser les pages tout autant que les panélistes avec la même acuité. Elle doit également être capable de décrire les interactions qui existent entre les différentes échelles et d'en tenir compte dans un système descriptif et analytique complet. L'appareil méthodologique que nous avons exposé montre assez clairement la place primordiale que doit tenir l'activité de navigation proprement dite, rattachée à un utilisateur donné, dans la sémantique des parcours. Quand bien même l'analyse de la production a un rôle à y jouer, c'est bien autour de l'internaute que se noue l'essentiel des éléments du parcours, et dans cette perspective, les données de trafic centrées-utilisateur dont nous disposons sont tout particulièrement adaptées à l'analyse des parcours.

Pour cela, nous ferons bien évidemment appel aux travaux déjà menés du côté tant de l'analyse de la page que de celle des stratégies et des modes de navigation ; nous tentons surtout de mobiliser ces recherches jusqu'ici disjointes autour de l'analyse des parcours. Nous avons recours pour y parvenir à une série de champs

disciplinaires, au premier rang desquels la linguistique informatique et la sémantique interprétative, associées aux outils de la statistique descriptive. Et, par-delà l'appel à tel ou tel outillage analytique susceptible d'éclairer un point précis de nos recherches, nous plaçons résolument nos travaux dans le cadre des sciences humaines et, corrélativement, d'une praxéologie.

Chapitre 2

Préparation et fouille des données

Nous travaillons sur des données de trafic centrées-utilisateur : nous avons vu les avantages méthodologiques que présente cette approche, mais pas encore abordé les problèmes techniques que posent leur recueil et leur traitement. L'objet de ce chapitre est de faire le point sur ces questions, en présentant les dispositifs logiciels de recueil de trafic sur des postes d'utilisateurs, le format des données recueillies ainsi que les étapes de pré-traitement de ces données indispensables à leur analyse.

2.1 Données de trafic « centrées-utilisateur »

Si le recueil de données de trafic centrées-serveur est maintenant relativement standardisé et validé dans ses méthodes[1], la collecte d'informations au niveau des postes d'utilisateurs est encore un champ d'activité en devenir, du fait des différents choix technologiques possibles et du degré de finesse des informations que l'on souhaite recueillir.

2.1.1 Technologies de recueil de données

Continuité ergonomique, ruptures techniques

Avant d'envisager les solutions techniques à mettre en œuvre pour recueillir des traces d'activité sur Internet et plus généralement sur la machine, quelques questions de méthode se posent. En développant des outils de recueil de trafic, on est obligé

[1] Après une période d'effervescence au moment de la « bulle Internet » et du fort développement des sites Web, des instances de normalisation et de certification (en France, l'association Diffusion Contrôle joue ce rôle) sont apparues afin de garantir et de rendre comparables les chiffres d'audience des sites. Les mesures sont ainsi faites par des instituts spécialisés (eStat, Xiti, etc.) dont les méthodologies sont contrôlées et qui tiennent le rôle de tiers de confiance.

d'avoir un angle d'approche technique de l'activité sur Internet. Quels protocoles sont utilisés, par quelles applications passent-ils, quel est le format des données envoyées et comment les intercepter ? La notion de protocole est ici fondamentale : pour dire les choses simplement, un protocole réseau définit, dans une architecture client-serveur, la façon dont le client doit formuler ses requêtes au serveur et dont celui-ci doit répondre[1]. L'élément important ici est que chaque protocole est lié à une famille d'interfaces et définit un champ d'interaction possibles avec les serveurs : par exemple, HTTP pour le Web, en utilisant comme logiciels clients la famille des navigateurs.

Dans l'absolu, une infinité de protocoles est possible, et chacun peut s'inventer un protocole pour faire communiquer des machines entre elles. Dans les faits, on distingue deux types de protocoles : les protocoles « standards », dont les spécifications sont publiques et strictement encadrées par des consortiums internationaux, et les protocoles « propriétaires », le plus souvent attachés à un éditeur de logiciels pour un type d'application particulier, et dont les spécifications sont tenues secrètes. Les principaux protocoles standards d'Internet sont les suivants :

- HTTP (Hyper Text Transfer Protocol) : protocole du Web, accompagné de sa version « sécurisée » HTTPS, il fonctionne sur un modèle informatique de communication de type client-serveur. Il permet au client d'envoyer à un serveur HTTP, autrement appelé serveur Web, une requête accompagnée ou non de paramètres ; le serveur lui renverra un contenu qui peut être soit un fichier, soit le résultat de l'exécution d'un programme.
- FTP (File Transfer Protocol) : également basé sur un modèle client-serveur, ce protocole a des possibilités plus limitées que HTTP, puisqu'il est dédié au transfert de fichiers, c'est-à-dire qu'il est impossible au client de donner des paramètres à sa requête ou de faire exécuter un programme sur le serveur. Le client se connectant à un serveur FTP ne peut que transférer des fichiers du serveur distant vers sa machine ou en sens inverse.
- POP3 et SMTP : utilisés pour la messagerie électronique, ils définissent les paramètres de la communication avec des serveurs de messagerie, respectivement pour la réception et l'envoi de messages. Ils nécessitent l'emploi de logiciels de messagerie spécifiques, tels Microsoft Outlook Express ou Netscape Messenger.
- NNTP : protocole employé pour l'accès aux « newsgroups » (ou « forums »), il est la plupart du temps pris en charge par les logiciels de messagerie.
- messagerie instantanée : plusieurs protocoles, propriétaires ou non, coexistent. On citera en particulier ICQ, MSN Messenger et le Messager de Yahoo.

---

[1] Pour plus de détails sur le fonctionnement technique du protocole HTTP voir Annexe 2, « Requêtes Web : mille-feuille technique ».

On a donc une triple correspondance entre protocoles, modes d'interaction et interfaces logicielles. Pour chaque protocole, un dispositif de recueil de trafic doit connaître les modalités techniques d'échange entre client et serveur, et produit des données dont l'interprétation ne sera pas la même selon le cas. Ainsi, la notion de durée qui opère pour l'analyse de l'activité sur le Web est inadaptée dans le cas de la messagerie, où l'action sur le réseau est un envoi ou une réception, qui met de côté la temporalité de l'écriture du message.

Cette séparation technique s'oppose peu ou prou au point de vue de l'utilisateur. Qu'il perçoive ou non les différences techniques sous-jacentes à l'utilisation de différents logiciels pour différents modes d'activité sur Internet, tous ces éléments sont pour lui en totale continuité. Ils coexistent sur l'ordinateur, et renvoient les uns aux autres au niveau des contenus proposés (un site propose de le contacter *via* la messagerie, une intervention dans un forum renvoie à une page Web, des logiciels de *peer-to-peer* ont recours à des pages Web pour leur mise à jour, etc.). Cette interpénétration forte des différents services offerts sur Internet implique que les dispositifs de recueil de données adoptent une démarche complète et large, qui s'oppose presque naturellement aux impératifs techniques auxquels ils sont confrontés dans l'analyse des protocoles.

La restriction du traçage à tel ou tel protocole s'impose donc comme une limitation incompatible avec un point de vue utilisateur global. Plus encore, pour l'exemple du Web, le traçage des actions menées *via* le protocole HTTP ne suffisent pas toujours pour connaître toute l'activité Web de l'utilisateur. Les navigateurs assurent, entre autres fonctions, le dialogue entre la machine de l'utilisateur et le serveur Web. S'ils sont originellement destinés à exploiter le protocole HTTP, d'autres protocoles sont « supportés » par les navigateurs modernes, en particulier FTP, et d'autre part ils ont la capacité de lancer l'exécution de programmes qui :

- utilisent des protocoles dits propriétaires (RealMedia ou AOL, par exemple) qui entrent pleinement dans la composition des pages. Une vidéo au format RealMedia peut ainsi apparaître dans une page Web, mais être lue par le lecteur RealPlayer lequel en télécharge le contenu en utilisant un protocole qu'il est le seul à connaître et maîtriser.
- savent traiter le type de fichier renvoyé par le serveur Web. Un exemple courant en est l'affichage des fichiers au format PDF, qu'un système de *plugin* permet de visualiser à l'intérieur de la fenêtre du navigateur ; il en est de même pour les *applets* Java, les documents Microsoft Word, le format multimédia Flash, etc.

Dans l'autre sens, le Web propose des interfaces pour accéder aux autres services (WebMail, WebChat, forums, etc.) qui font du HTTP le support de services qui originellement ne lui étaient pas attribués.

En somme, derrière les belles interfaces utilisateurs et l'interopérabilité croissante des outils et des services sur Internet, on découvre un univers de dispositifs techniques dont les interactions sont complexes. Un système de recueil de trafic se doit de prendre en compte cette complexité et cette discontinuité tout en ayant en vue la continuité de cet ensemble pour l'utilisateur.

Plusieurs choix technologiques possibles

Face à ces impératifs, plusieurs stratégies ont été élaborées pour recueillir des données d'usage centrées-utilisateur en « conditions naturelles ». Elles correspondent à différents positionnements du dispositif dans la chaîne technique de traitement de l'activité Internet ou Web.

La première approche est externe, et consiste à procéder à des enregistrements vidéo de l'utilisateur et de son écran. C'est la méthode retenue dans [Byrne *et al.* 1999a] : durant dix jours, les participants à l'expérimentation ont été invités à déclencher la caméra dès qu'ils naviguaient sur le Web, ainsi qu'à commenter leurs actions afin que l'interprétation des enregistrement soit facilitée. Ce dispositif, assez intrusif et peu aisé à mettre en œuvre, se centre plus sur les actions de l'utilisateur sur l'interface (ouverture de page, enregistrement, impression, etc.) et les tâches effectuées, codées en une « taskonomy » à huit entrées[1]. Il ne permet pas à proprement parler de recueillir des données de trafic, mais permet d'observer finement le comportement de l'utilisateur face à l'IHM, et permet d'obtenir des données très fines de ce point de vue.

Pour recueillir des données de trafic proprement dites, c'est-à-dire des enregistrements horodatés d'actions techniques typées, il est nécessaire d'avoir recours à des composants logiciels. Cette solution se décline génériquement en autant de positions du dispositif dans la chaîne de traitement des requêtes envoyées d'un poste vers des serveurs distants.

Premier cas, une sonde peut être intégrée au logiciel client lui-même, par exemple un navigateur. La première étude d'usages du Web centrée-utilisateur, [Catledge & Pitkow 1995], utilise ce procédé en modifiant le navigateur XMosaic, de même que [Cunha *et al.* 1995] et [Tauscher & Greenberg 1997a]. D'autres solutions, telles que celle mise en œuvre par la société WebGalaxis, consistent à développer des composants logiciels qui s'intègrent au navigateur. Dans tous les cas, il s'agit d'enregistrer les actions de l'utilisateur sur l'interface. Les données recueillies perdent en couverture ce qu'elles gagnent en précision : seul le logiciel pour lequel a été développé le composant est tracé, mais le niveau de traçage est très fin. Pour le cas des navigateurs, il est possible de savoir si une page a été ouverte à partir d'un lien, d'une entrée dans les Favoris ou tapée par l'utilisateur, si la page est imprimée, si l'ascenseur est utilisé, si plusieurs fenêtres sont ouvertes en même temps, etc.

Notons qu'une version plus légère de cette stratégie centrée sur le logiciel utilisée consiste à recourir aux fonctionnalités déjà existantes d'enregistrement de données. Pour le cas des clients de *chat*, par exemple, il est la plupart du temps possible d'enregistrer une trace des échanges ; pour les navigateurs, [Cockburn & McKenzie 2000] mettent à contribution le système d'historique de Netscape Navigator.

---

[1] Ces entrées sont : « use information », « locate information », « provide information », « find on page », « navigate », « configure browser », « manage window » et « react to environment ». Voir [Byrne *et al.* 1999b].

À l'autre bout de la chaîne, les outils de métrologie des réseaux permettent de se positionner à des points intermédiaires entre le poste de l'utilisateur et les serveurs de contenus et de services : serveurs proxy, routeurs, DSLAM, répartiteurs, etc[1]. Les données sont regroupées par poste client (par adresse IP de machine). Les sondes, non intrusives, examinent les en-têtes des paques IP et peuvent ainsi mesurer le volume échangé par protocole (par numéro de port, plus précisément). Il est ainsi possible d'avoir des informations précises et horodatées sur les types de protocoles utilisés et les volumétries engagées : Web, messagerie classique, *peer-to-peer*, etc.

Enfin, certains dispositifs ont une position intermédiaire entre ces deux types de solutions. Il s'agit de positionner la sonde sur le poste de l'utilisateur, au niveau de la couche réseau : l'ensemble des communications entre la machine et l'extérieur peut être tracée, en même temps que l'on peut identifier des utilisateurs particuliers et non seulement l'usage d'un poste en général. La sonde doit ensuite, pour chaque protocole, mettre en œuvre des modules logiciels spécifiques afin de repérer, d'analyser et d'extraire les informations qui y sont liées : pour la messagerie sortante par exemple (protocole SMTP), la sonde doit être capable de reconnaître les champs spécifiant les destinataires, les pièces jointes, etc. Bien évidemment, ceci est plus aisé lorsque les protocoles sont documentés, comme c'est le cas pour la majorité des protocoles utilisés sur le Net (HTTP, POP, SMTP, NNTP) ; sans cela, l'analyse s'avère plus délicate et nécessite de faire de la rétroconception pour décrypter les modes de communication client-serveur, par exemple pour le protocole Exchange utilisé par Microsoft Outlook.

Deux dispositifs de ce type nous fournissent les données sur lesquelles nous travaillons ici, dont nous détaillons ci-dessous le fonctionnement et les données recueillies. Avant cela, notons que les trois grands types de méthodes pour recueillir des données trafic que nous venons d'exposer ne nous apparaissent pas comme exclusives les unes des autres : chaque dispositif apporte des informations particulières et un niveau de détail ou de couverture propre, et c'est plus dans la complémentarité ou la sélection de problématiques d'usages particulières qu'il faut envisager leur déploiement.

Technologies de recueil de données

Les données de trafic sur lesquelles nous travaillons sont issues de deux projets de recherche différents qui utilisent deux sondes distinctes. Si nous détaillons au Chapitre 5 les panels et les durées d'observations de ces projets, précisons d'ores et déjà nos sources. D'un côté, nous disposons de données fournies par la société de

---

[1] La métrologie des réseaux s'est développée initialement autour des problématiques de performance et d'architecture des réseaux. L'exploitation des traces de trafic pour l'analyse des usages est une préoccupation plus récente, c'est pourquoi nous ne présentons ici que succinctement cette discipline, et renvoyons à la lecture de [Owezarski 2001] pour une présentation générale du domaine.

mesure d'audience NetValue dans le cadre des partenariats TypWeb et SensNet[1], utilisant la technologie NetMeter développée par NetValue ; de l'autre, nous avons des traces issues du projet BibUsages, recueillies à l'aide de la sonde Audinet développée par France Télécom R&D[2].

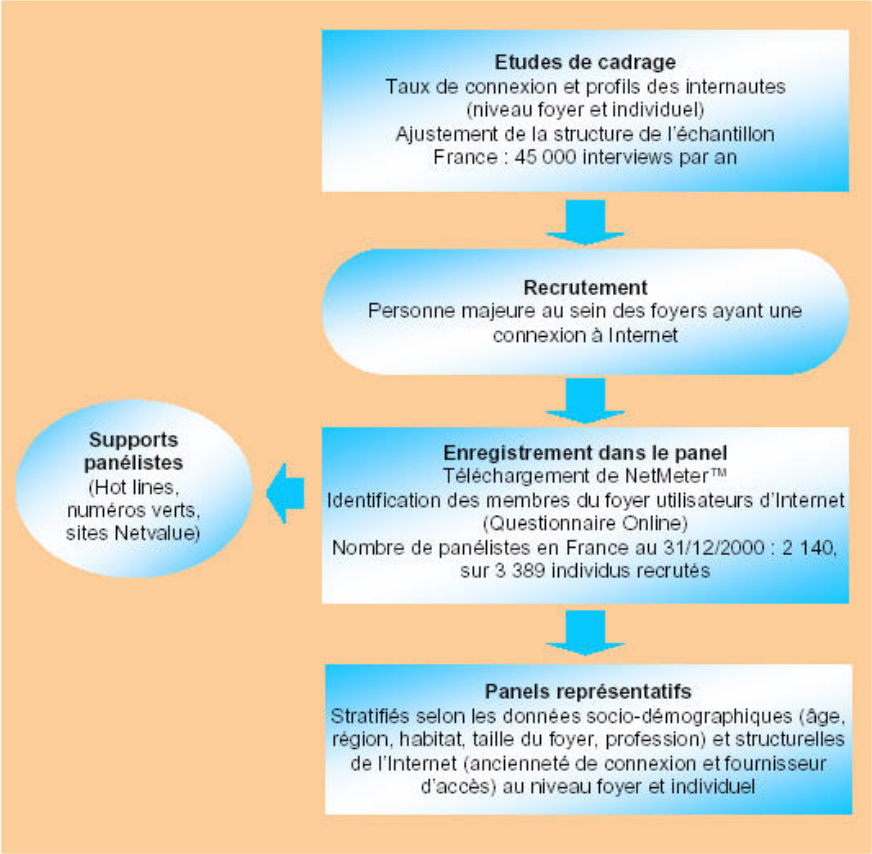


Figure 2.1. Constitution et le suivi du panel NetValue (source : NetValue, 2002)

[1] Le projet TypWeb est un partenariat entre NetValue, France Télécom R&D, Wanadoo S.A. et HEC, mené en 2000-2001. Son objectif était d'exploiter de manière approfondie les données de trafic du panel France de NetValue sur l'année 2000. Le projet SensNet (2002-2004) prolonge le projet TypWeb, en intégrant des données de trafic de 2001 et 2002 pour la France, et de 2002 pour l'Espagne et le Royaume-Uni. Les partenaires sont : NetValue, France Télécom R&D, le LIMSI et Paris III. Voir en Annexe 1, « Projets » pour une description complète de ces deux projets.

[2] Le projet BibUsages est un projet RNRT mené en 2002 par France Télécom R&D et la Bibliothèque Nationale de France, et portant sur l'étude des usages des bibliothèques électroniques. Voir en Annexe 1 pour une description approfondie du projet.

Le dispositif de recueil de trafic du panel NetValue repose sur la technologie NetMeter, développée par la société NetValue. La constitution et le suivi du panel NetValue sont décrits dans la Figure 2.1 ci-dessus. Le suivi de l'activité Internet est réalisé en temps réel grâce au logiciel NetMeter, implanté sur l'ordinateur de chaque panéliste.

L'analyse des informations au niveau individuel est faite grâce à l'identification des différents utilisateurs du foyer. NetMeter est compatible avec les dernières versions des systèmes d'exploitation et fonctionne en tâche de fond sur l'ordinateur du panéliste. Il démarre automatiquement et enregistre en permanence les utilisations d'Internet ; par ailleurs, la sonde prend très peu de place et ne gêne pas le fonctionnement des applications habituelles. Régulièrement, voire quotidiennement, les données enregistrées par NetMeter sont envoyées automatiquement via la connexion Internet vers un serveur dédié de NetValue sans que cette transmission perturbe l'utilisateur. Ces données sont ensuite validées et chargées dans une base de données.

Les données du panel BibUsages sont recueillies à l'aide de la technologie Audinet, développée par Laurent Rabret à France Télécom R&D - DAC. Audinet est composé de logiciels clients et serveurs, dont la Figure 2.2 résume l'architecture.

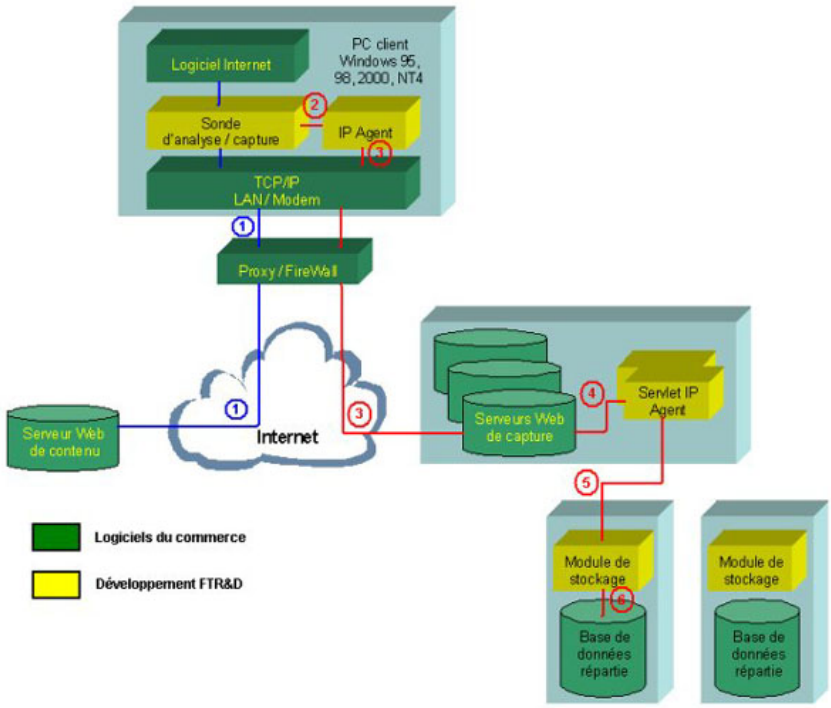


Figure 2.2. Architecture logicielle d'Audinet (source : L. Rabret, FTR&D)

Les logiciels clients sont installés sur la machine des internautes ; durant l'installation, la sonde d'Audinet est insérée dans le cœur du système d'exploitation Windows (Windows 95, 98, 2000 , NT 4 et XP sont supportés). Elle devient

automatiquement active lorsqu'un logiciel client (« Internet Explorer » ou « Netscape Navigator » par exemple) accède à Internet. L'ensemble des échanges réseau est alors analysé, puis le résultat de l'analyse transmis vers un serveur de collecte via le protocole HTTP ou HTTPS du Web. La sonde qui capture les flux Internet a été optimisée pour avoir un impact minimal sur les logiciels utilisés par les clients. Les informations recueillies sont ensuite envoyées vers une application locale qui, à son tour, transmet les données vers le serveur de collecte. Ces données ne sont transmises que lorsque le trafic réseau est faible, afin que le client ne constate aucune dégradation de performance du réseau après avoir installé Audinet.

Les composants du serveur de collecte Audinet ont été développés en Java. Des serveurs Web du commerce gèrent les connexions avec les clients afin de rapatrier les données, les transmettent aux composants spécifiques Audinet, lesquels inscrivent les informations de trafic dans une base de données. L'horodatage des actions des utilisateurs est réalisé par le serveur de collecte, les heures des postes clients étant la plupart du temps peu fiables et peu homogènes.

Pour les technologies NetMeter comme Audinet, l'analyse et le recueil de données se fait au niveau de la couche réseau, c'est-à-dire entre les différentes applications accédant au réseau vers des postes distants. Nous pouvons ainsi connaître les applications réseau employées par les utilisateurs, et pour certaines d'entre elles (navigateurs, logiciels de messagerie, etc), l'analyse des flux transitant par le réseau est faite de manière plus poussée de sorte que les informations telles que l'URL demandée sur le Web, l'adresse du destinataire d'un mail, le nom d'un newsgroup sont identifiés et ces informations envoyées aux serveurs de collecte.

Synthèse. *Les données de trafic centrées-utilisateur reposent sur la collecte de traces d'échanges entre le poste de l'utilisateur et les serveurs distants. Elles reposent sur l'installation d'un dispositif technique spécifique qui enregistre tout ou partie de la communication sur le réseau.*

2.1.2 Format des données

Informations recueillies

In fine, les données recueillies, contiennent pour chaque protocole la trace de chaque requête, c'est-à-dire l'heure exacte de l'action et les informations propres au protocole utilisé, et ce pour chaque utilisateur.

En outre, nous disposons du nom des exécutables accédant au réseau par TCP/IP : *iexplore.exe* pour Internet Explorer, *msimn.exe* pour Outlook Express, etc. Ceci permet de détecter l'utilisation d'applications comme RealPlayer ou Kazaa, et d'avoir des données élémentaires pour les protocoles qui ne sont pas analysés dans le détail. Par exemple, l'exécutable *cs.exe* correspond au fait de jouer en réseau à Counter Strike ; même sans analyser dans le détail le contenu des échanges dans le cadre du protocole utilisé par le jeu, on peut savoir si l'utilisateur y joue, quand et sur quelles durées.

En ce qui concerne les protocoles analysés, chaque protocole renvoie des informations qui lui sont propres. Pour les protocoles non Web, nous avons les informations suivantes :

- POP (messagerie entrante) :
  - adresse de l'expéditeur ;
  - adresse des destinataires directs ;
  - adresse des destinataires en copie ;
  - date de réception sur le serveur de messagerie ;
  - date de réception par le client ;
  - le sujet du message ;
  - la taille totale du message ;
  - le nombre de fichiers joints ;
  - les noms des fichiers joints.
- SMTP (messagerie sortante) :
  - adresse de l'expéditeur ;
  - adresse des destinataires directs ;
  - adresse des destinataires en copie ;
  - adresse des destinataires en copie cachée ;
  - date d'envoi par le client ;
  - le sujet du message ;
  - la taille totale du message ;
  - le nombre de fichiers joints ;
  - les noms des fichiers joints.
- NNTP (forums) :
  - nom du groupe de discussion ;
  - adresse de l'expéditeur ;
  - le sujet du message ;
  - le type de message ;
  - la taille totale du message ;
  - date de réception par le client.

En ce qui concerne le trafic Web (protocole HTTP), pour NetMeter comme pour Audinet, toutes les requêtes effectuées par le navigateur ne sont pas enregistrées : des filtres écartent les fichiers de type image (formats GIF, JPEG, PNG, etc.) lorsque ceux-ci sont intégrés dans une page, comme c'est le cas pour l'immense majorité des pages Web. Pour le trafic Web, les deux sondes recueillent les informations suivantes :

- date d'envoi de la requête : la précision est à la seconde. On pourrait souhaiter disposer d'une précision plus fine, de l'ordre du centième de seconde, car la rapidité et la superposition des requêtes font fréquemment se chevaucher plusieurs actions en une même seconde. Audinet permet de connaître également les dates de réception du premier et du dernier paquet IP de données (début et fin de chargement du fichier), ce qui pallie un peu ce désagrément.
- URL demandée : notons que nous n'avons pas ici d'information sur les arguments passés aux requêtes de type POST.

date	URL	Referer
10/10/2002 07:14:10	http://www.free.fr/	NULL
10/10/2002 07:14:12	http://chaines.free.fr/script/thema.js	http://www.free.fr/
10/10/2002 07:14:13	http://www.free.fr/free.css	http://www.free.fr/
10/10/2002 07:14:13	http://img.free.fr/img/mymail.pl	http://www.free.fr/
10/10/2002 07:14:14	http://ad.fr.doubleclick.net/adj/ts.free.fr/portail/accueil;dcopt=ist;kw=x;sz=468x60;c	http://www.free.fr/
10/10/2002 07:14:20	http://www.caramail.com/	NULL
10/10/2002 07:14:21	http://www.caramail.lycos.fr/	NULL
10/10/2002 07:14:21	http://www44.caramail.lycos.fr/general.html	http://www.caramail.lycos.fr/
10/10/2002 07:14:23	http://www44.caramail.lycos.fr/Bin/Utils/styleSheet.css	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:26	http://ads-fr.spray.net/js.ng/btype=1&country=fr&kw=NULL&adpos=1&affiliate=fr	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:29	http://ads-fr.spray.net/js.ng/btype=36&country=fr&kw=NULL&adpos=1&affiliate=fr	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:31	http://ads-fr.spray.net/js.ng/btype=36&country=fr&kw=NULL&adpos=2&affiliate=fr	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:31	http://ads-fr.spray.net/js.ng/btype=36&country=fr&kw=NULL&adpos=4&affiliate=fr	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:32	http://ads-fr.spray.net/js.ng/btype=36&country=fr&kw=NULL&adpos=3&affiliate=fr	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:48	http://js.cybermonitor.com/lycos.js	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:49	http://s0b.bluestreak.com/lx.e?ir&s=108677&n=2002.10.10.5.14.28.0	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:49	http://www44.caramail.lycos.fr/cgi-bin/baltop	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:50	http://stat3.cybermonitor.com/lycos_v?R=homepage_caramail1&S=total;homepa	http://www44.caramail.lycos.fr/general.html
10/10/2002 07:14:50	http://ads-fr.spray.net/js.ng/country=fr&kw=NULL&btype=1&adpos=1&affiliate=fr	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:14:51	http://stat3.cybermonitor.com/lycos_v?R=homepage_caramail1&S=total;homepa	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:14:53	http://ads-fr.spray.net/js.ng/country=fr&kw=NULL&btype=36&adpos=1&affiliate=fr	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:14:54	http://adfarm.mediaplex.com/ad/bn/709-4893-3826-21?mpt=2002.10.10.5.14.49	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:14:57	http://ads-fr.spray.net/js.ng/country=fr&kw=NULL&btype=36&adpos=2&affiliate=fr	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:15:02	http://ads-fr.spray.net/js.ng/country=fr&kw=NULL&btype=36&adpos=3&affiliate=fr	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:15:12	http://ads-fr.spray.net/js.ng/country=fr&kw=NULL&btype=36&adpos=4&affiliate=fr	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:15:14	http://www44.caramail.lycos.fr/cgi-bin/baltop	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:15:15	http://www44.caramail.lycos.fr/cgi-bin/NbParFolder	http://www44.caramail.lycos.fr/cgi-bin/baltop
10/10/2002 07:15:20	http://www44.caramail.lycos.fr/cgi-bin/Folder	http://www44.caramail.lycos.fr/cgi-bin/baltop

Figure 2.3. Extrait de données de trafic Web

- Referer : les navigateurs renseignent, dans une requête HTTP, un champ nommé « Referer », qui a pour valeur l'URL de la page d'où provient la requête (lors du suivi d'un lien, de l'envoi d'un formulaire), ou est vide dans les autres cas (par exemple lorsque l'utilisateur entre manuellement l'URL ou en sélectionne une dans ses favoris).
- code retour HTTP : les serveurs Web renvoient, dans l'en-tête HTTP des réponses aux requêtes, un « code retour » qui indique comment la requête a été traitée. Par exemple : 200 si la requête est traitée correctement, 404 si la ressource demandée n'existe pas, etc.
- taille : Audinet fournit la taille en octets de chaque fichier téléchargé par l'utilisateur.

La Figure 2.3 p. 52 fournit un exemple d'extrait de données de navigation : il s'agit de la visite de trois pages par un utilisateur en octobre 2002 (pour des raisons de lecture, nous n'avons affiché que la date, l'URL et le Referer). Dans cet exemple, du point de vue de l'utilisateur, seules trois pages sont demandées : la page d'accueil de Free (`www.free.fr`), la page d'accueil du site de WebMail Caramail, et l'affichage du contenu de la messagerie sur ce site après authentification.

On perçoit ici la distorsion entre la perception de l'utilisateur et les données recueillies, qui comptent, malgré le filtrage des images, vingt-huit requêtes pour ces trois pages : bandeaux publicitaires, compteurs, feuilles de style entrent dans la composition de la page vue et génèrent autant de requêtes de la part du navigateur.

Formalisme pour le stockage

On ne s'étonnera pas, après cet exemple, que les données recueillies soient volumineuses. Ce problème est loin d'être anecdotique, et nécessite une véritable réflexion sur les formalismes de stockage des données de trafic.

Les systèmes de gestion de bases de données (SGBD) fournissent indiscutablement un mode de stockage adapté à ce type de données. Pour autant, il importe de trouver un équilibre entre redondance des données et performance des requêtes SQL qui les exploiteront. Dans ce cadre, la plateforme de traitement de données de trafic développée à France Télécom R&D[1] propose un formalisme adapté à ces impératifs.

Pour les données de trafic Web, sur lesquelles porte notre travail, deux tables distinctes en rendent compte :

- table d'URL : elle regroupe l'ensemble des URL distinctes visitées pour un panel donné, et contient les champs suivants :
  - pag_id : identifiant unique de page,
  - url : l'adresse de la page,

  ainsi que l'ensemble des éléments relatifs à l'URL qui sont calculés par la suite (voir ci-dessous, ainsi que le Chapitre 3, « De l'URL au contenu »).

[1] Plateforme développée dans le cadre des projets TypWeb et SensNet.

- table de navigation WEB : elle contient les éléments horodatés et personnalisés de navigation, et elle est structurée autour des champs suivants :
  - pan_id : identifiant de panéliste,
  - date : la date de l'action, précise à la seconde,
  - pag_id : l'identifiant de l'URL demandée dans la table des URL,
  - referer : l'identifiant du Referer de la requête dans la table des URL,

  ainsi que la duplication d'informations relatives aux URL, issues de la table d'URL, recopiées ici pour des raisons de performance.

C'est sur ces données de base et dans ce formalisme que travaillent l'ensemble des traitements décrits par la suite. Les développements logiciels, qui occupent une part non négligeable de notre travail, sont adaptés à ce type de données et s'intègrent ainsi à la plateforme et aux outils développés à France Télécom R&D pour l'analyse de données de trafic.

> ***Synthèse.*** *En ce qui concerne l'accès au Web, les sondes de recueil de trafic analysent l'ensemble des requêtes HTTP envoyées par l'internaute. Elles fournissent une liste horodatée exhaustive des URL demandées par un utilisateur donné.*

2.2 Formatage des données pour l'analyse de trafic

Les données de trafic Web, même transposées pour correspondre au schéma de base de données décrit ci-dessus, sont encore dans une forme très « brute », et nécessitent une série de traitements pour être effectivement exploitables. Nous traitons dans cette partie des éléments de formatage des données : identification des sessions, des sites et des pages « vues », où nous verrons que ces étapes posent déjà, avant même d'envisager l'analyse des usages du Web, une série de problèmes incontournables. Il ne s'agit pas encore ici de parler d'enrichissement des données, auquel sont consacrés le Chapitre 3 et le Chapitre 4, mais d'une phase de pré-traitement dont la validité conditionne les travaux ultérieurs.

2.2.1 Identifier les sessions

L'identification de sessions correspond à la nécessité de repérer, au sein des données de trafic, des plages d'activité cohérentes de l'utilisateur. Ce repérage a déjà fait l'objet de nombreux travaux du côté de l'analyse des *logs* de serveurs Web[1], mais le point de vue et la complexité des données de trafic centrées-utilisateur posent des problèmes spécifiques qui nécessitent de mettre en place des stratégies *ad hoc*.

---

[1] Voir par exemple [Cooley *et al.* 1999a].

Sessions multi-protocoles

Nous l'avons dit, les discontinuités techniques observées dans la séparation des différents protocoles au sein des données de trafic s'opposent à la continuité et à la complémentarité des outils et des interfaces du point de vue de l'utilisateur. Dans la pratique, on peut observer des entrelacements très forts entre les différents outils, comme en témoigne l'exemple présenté à la Figure 2.4 ci-dessous.

Dans cet extrait de données globales de trafic, le panéliste s'est reconnecté à 18h32, après une interruption d'une heure, et a navigué sur le Web, ouvert son outil de Messagerie Instantanée (IM), reçu un mail, refait de l'IM, puis du Web, puis de l'IM, envoyé un message, fait du Web et enfin fait de l'IM. L'adoption du point de vue utilisateur impose de tenir compte de l'ensemble de cette activité pour identifier les sessions. Ainsi, dans le cadre du projet TypWeb, une méthodologie spécifique a été mise en place, qui intègre l'ensemble des protocoles mobilisés pour le repérage des sessions Internet. Ceci modifie, souvent significativement, la durée mesurée des sessions, ainsi que leur nombre, et a une influence sur les mesures d'utilisation de services au cours d'une session.

pan_id	date	type	proto	duree
18829	2000-06-24 12:31:45	Web	http	8
18829	2000-06-24 12:31:53	Web	http	12
18829	2000-06-24 12:32:25	Autre	Messager	4
18829	2000-06-24 12:33:20	Autre	Messager	1925
.....				
18829	2000-06-24 12:55:31	Autre	Messager	563
18829	2000-06-24 13:02:52	Autre	Messager	10
18829	2000-06-24 13:03:57	Autre	Messager	6
18829	2000-06-24 14:42:58	Mail	sendmail	0
18829	2000-06-24 14:43:12	Web	http	10
18829	2000-06-24 14:43:22	Web	http	12
18829	2000-06-24 14:43:56	Autre	Messager	4
18829	2000-06-24 17:32:05	Web	http	24
18829	2000-06-24 17:32:29	Web	http	283
18829	2000-06-24 17:32:46	Autre	Messager	3
18829	2000-06-24 17:33:24	Autre	Messager	105
18829	2000-06-24 18:32:33	Web	http	7
18829	2000-06-24 18:32:58	Web	http	4
18829	2000-06-24 18:33:27	Autre	Messager	45
18829	2000-06-24 18:36:09	Mail	recvmail	0
18829	2000-06-24 18:38:51	Autre	Messager	607
18829	2000-06-24 18:39:24	Autre	Messager	6
.....				
18829	2000-06-24 18:48:40	Autre	Messager	4
18829	2000-06-24 18:48:49	Autre	Messager	5
18829	2000-06-24 18:49:26	Web	http	5
18829	2000-06-24 18:49:31	Web	http	13
18829	2000-06-24 18:50:04	Autre	Messager	4
18829	2000-06-24 19:06:11	Mail	sendmail	0
18829	2000-06-24 19:07:24	Web	http	6
18829	2000-06-24 19:07:30	Web	http	11
18829	2000-06-24 19:08:01	Autre	Messager	4

Figure 2.4. Exemple de session Internet multiprotocoles

Ces éléments ont nécessité de se pencher également sur la définition de la limite des sessions. Dans les données, l'utilisateur ne donne pas d'indication pour dire quand il commence et finit d' « utiliser Internet » : on observe des traces d'activité entrecoupées de périodes plus ou moins longues d'inactivité (aucune trace). L'enjeu

est de définir quelle période d'inactivité on retient pour déclarer qu'une session est terminée et que les traces suivantes appartiennent à une autre session.

Dans le cas où l'on a des données de trafic durant une soirée, et les suivantes le lendemain, la chose est assez simple et intuitive : les deux plages d'activité Internet sont bien distinctes, et correspondent à deux sessions différentes. Mais cette différence est parfois plus ténue : que dire de quelqu'un qui suspend son activité Web pendant 45 minutes, et reprend, en quelque sorte, là où il s'était arrêté ? Derrière ce questionnement, se profilent deux problèmes : le premier, technique, tient au fait qu'on ne suit pas l'utilisateur « à domicile », et que l'on ne sait pas pourquoi ni comment ont lieu les interruptions d'activité Internet. Le second est d'ordre théorique : dans quelle mesure une interruption, quelle que soit sa durée, signifie-t-elle la fin ou la suspension d'une activité ? Quels sont les éléments de continuité entre deux moments d'une même activité séparés de plusieurs minutes ou plusieurs heures ? Ces questions dépassent le cadre de notre travail présent, et ne sont de toute façon pas décidables à l'aide des données dont nous disposons, nous n'irons pas plus avant sur cette question ; gardons toutefois à l'esprit que, une fois de plus, les données de trafic introduisent du discontinu là où il peut y avoir une continuité pour l'utilisateur, cette fois au niveau de la temporalité des activités construites.

Il n'en subsiste pas moins la nécessité d'identifier une durée d'inactivité Internet pour borner les sessions : dans le cadre du projet TypWeb, plusieurs limites ont été éprouvées. Mis à part les valeurs extrêmes (plus de 24 heures), l'écart moyen entre les événements (page Web consultée, mail reçu ou envoyé, etc.), étant de 12 minutes, trois hypothèses ont été testées : attribution de la fin de session au bout de 15, 30 et 45 minutes d'inactivité. Les différences de paramètres testées (durée moyenne d'une session, nombre de sessions par mois, etc.) se stabilisant entre les hypothèses de 30 et 45 min, par rapport aux 15 et 30 minutes, la limite retenue est finalement celle de 30 minutes d'inactivité comme seuil d'attribution d'une nouvelle session, ce que représente la Figure 2.5.

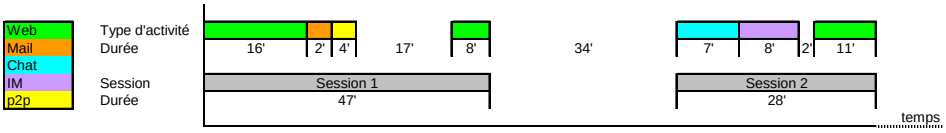


Figure 2.5. Identification des sessions sur la base de 30 minutes d'inactivité

Limitations

Ainsi, une session Internet devient, dans nos données, une période homogène d'activité sur Internet. Elle ne postule pas pour autant une homogénéité pour l'utilisateur : plusieurs cours d'action distincts peuvent coexister au sein d'une même session, et un même cours d'action peut être distribué dans des sessions distinctes et non consécutives. Néanmoins, la session n'est pas totalement disjointe de l'activité de l'utilisateur, puisqu'elle rend compte du fait que l'utilisateur, physiquement présent devant son ordinateur, est impliqué dans des processus de communication, de recherche d'information, de loisir, etc. sur le réseau, avec des individus, des sites, des services distants. Cette distanciation des ressources, renforcée par le processus de connexion, fait de l'activité « on line » un état objectif assumé par l'utilisateur, et

valide la session comme unité de temps et d'action fidèle à l'activité effective perçue par l'internaute.

Quelques limitations, toutefois, sont inhérentes au dispositif technique de recueil de données : les sondes utilisées ici ne tracent que les activités qui impliquent une communication réseau. Deux biais sont induits, le premier du côté de l'inclusion (de l'ordre de la précision) et le second de l'exclusion (de l'ordre du rappel).

En matière de précision, on a postulé jusqu'alors que la trace impliquait l'activité de l'internaute. Ceci n'est pas toujours vrai, en particulier avec les connexions haut débit qui invitent à laisser la communication systématiquement ouverte, pour faire du *peer-to-peer* notamment. Dans ces conditions, certaines actions peuvent se faire automatiquement sur le poste en l'absence de l'utilisateur : rafraîchissement automatique d'une page Web, réception périodique de nouveaux messages, téléchargement, etc. On peut ainsi observer dans nos données certaines sessions de plus de 24 heures ; fort heureusement, ces éléments sont très rares, et le problème reste mineur. Il sera toutefois à prendre en compte à l'avenir, avec la diffusion des accès Internet haut débit illimités et l'élargissement des offres de contenus et de services sur Internet incitant à une connexion active en permanence.

Réciproquement, certaines traces ne donnent qu'une vision partielle de l'activité. Nous pensons en particulier à la messagerie : nous avons trace des moments d'envoi et de réception des messages, mais pas du tout de l'écriture des messages, alors même que c'est cette activité qui est au centre du processus et dont il serait intéressant d'analyser la dimension temporelle. En ce qui concerne le Web, le phénomène est plus ténu, mais nous en savons pas non plus ce qui relève de la consultation hors-ligne, que ce soit dans le cache du navigateur ou que cela résulte d'une stratégie de l'utilisateur de récupérer au plus vite les documents pour les trier une fois la connexion fermée.

Malgré ces quelques bémols, les données de trafic dont nous disposons sont globalement fiables, riches et complètes. Ne travaillant, dans le cadre de cette étude, que sur l'analyse du trafic Web, nous laissons de côté les problèmes liés à la messagerie et aux autres protocoles, et évitons les écueils et les difficultés qui y sont liés. Nous conservons néanmoins une approche globale de l'activité Internet dans la mesure où les sessions que nous utilisons sont calculées avec l'ensemble des données de trafic disponibles.

> ***Synthèse.*** *Les sessions représentent des unités cohérentes d'activité sur Internet. Afin de tenir compte de l'entrelacement et de la complémentarité des différents outils Internet (Web, mail, etc.), leur identification se base sur l'ensemble des traces de trafic enregistrées. Une période d'inactivité de plus de 30 minutes marque la séparation entre deux sessions.*

2.2.2 Traitement des URL

Les données de trafic Web nécessitent, pour être pleinement exploitées, que les URL soient décomposées en différents éléments ; cette étape permet ensuite d'identifier les sites sur lesquels ont été visitées les pages, moyennant cependant quelques traitements spécifiques.

Découpage des URL

Une URL (Uniform Resource Location) est un format de nommage universel pour désigner une ressource sur Internet. Derrière les formes canoniques les plus connues et courantes, on observe une certaine diversité, comme en témoignent quelques exemples d'URL rencontrées dans les données de trafic dont nous disposons :

- `http://www.sncf.fr`
- `http://fr.search.yahoo.com/search/fr?o=1&zw=1&p=escargots+bourgogne&d=y&za=and&h=c&g=0&n=20`
- `http://194.51.10.18:8080/enoviewer/servlet/GetGeoData?rset=WNOAA2000&ps=3000.0&pq=50.62500000000001,16.875`
- `https://www.lbmicro.com:443/cgi-bin/emcgi?session=eyRCVCC0`
- `ftp://ftp.schneeberger.fr/schneeberger/Pub/dc2/dc2nc20a.txt`
- `ftp://mp3@007mp3.dyndns.org:21/%3D%3DFULL%20ALBUMS%20002%3D%3D/Tom%20Jones%20-%20Reload/`
- `aol://aol.prop/4344:3873.d1_res.35914016.591441539`

De manière générale, une URL se présente comme une chaîne de caractères répondant à un format particulier, que l'on peut décomposer ainsi :

```
[protocole]://[utilisateur]@[serveur]:[port]/
        [répertoire]/[fichier]?[arguments]#[ancre]
```

Plus précisément, ces éléments sont :

- *Protocole* : pour le Web, il s'agit, en ce qui concerne les protocoles publics, de HTTP, HTTPS, FTP et GOPHER. On trouve également des protocoles propriétaires, en particulier AOL (voir en Annexe 2 pour plus de détails sur les protocoles).
- *Utilisateur* : certains serveurs, le plus souvent les serveurs FTP, nécessitent une authentification des clients (spécification d'un nom d'utilisateur et d'un mot de passe) par ce biais. Le nom d'utilisateur peut être intégré à l'URL sous la forme `nomUtilisateur@`, mais dans la pratique, il est très rarement renseigné.
- *Serveur et port* : l'adresse de la machine distante, et optionnellement le nom du port utilisé (par défaut, le port 80 est utilisé si aucun n'est spécifié). Dans le cadre du protocole TCP/IP, cette adresse est écrite sous forme de quatre numéros allant de `0` à `255` (quatre fois 8 bits) ; on la note donc sous la forme `xxx.xxx.xxx.xxx` où chaque `xxx` représente un entier de `0` à `255`. Cela étant, grâce au système DNS (Domain Name System), une machine distante peut être désignée sous une forme plus intelligible, le nom de domaine. On appelle ainsi nom de domaine le nom à deux composantes, dont la première est un nom correspondant au nom de l'organisation ou de l'entreprise, le second à la classification de domaine (*.fr*, *.com*, etc.). Chaque machine d'un domaine est appelée hôte. Le nom d'hôte qui lui est attribué doit être unique dans le domaine considéré (le serveur Web d'un domaine porte généralement le nom `www`).
- *Répertoire* : le chemin sur le serveur vers le répertoire contenant le fichier visé. Il peut contenir des codes représentant des caractères non

alphanumériques, accentués, ou de la ponctuation, par l'intermédiaire des règles de leur valeur Unicode notée en base hexadécimale, par exemple :

Caractère	Notation
[espace]	`%20`
é	`%E9`
à	`%E0`
É	`%C9`

- *Fichier* : le nom du fichier demandé par l'utilisateur, que ce soit un fichier « statique » ou qu'il appelle l'exécution d'une ou plusieurs procédures sur le serveur qui compose le résultat renvoyé à l'utilisateur. Trois choses sont à noter à cet endroit : d'une part, les serveurs Web ont un mécanisme de fichier appelé par défaut. Si aucun nom de fichier n'est spécifié dans l'URL, ils puisent dans une liste de noms spécifiés dans la configuration du serveur, examinent si un fichier correspond à ce nom dans le répertoire demandé, et renvoient vers le fichier correspondant s'il existe. Dans la pratique, le nom `index.html` et ses corrélats (`index.htm`, `welcome.html`, `bienvenue.html`, `index.php`, etc.) sont la valeur par défaut des serveurs, mais rien n'empêche l'administrateur d'un serveur de spécifier le nom qu'il désire, ou aucun.
  Si ce mécanisme n'est pas activé dans la configuration d'un serveur, ou qu'aucun fichier dont le nom correspond à ceux fixés par défaut n'est trouvé, le serveur renvoie, si la configuration l'y autorise, la liste des fichiers contenus dans le répertoire. En cas d'interdiction, le serveur renvoie une erreur de type 403, « Access Forbidden ».
- *Argument* : lors d'une requête adressée à un serveur Web, la présence d'arguments permet à l'utilisateur de passer des paramètres aux programmes exécutés sur le serveur. La série d'argument est de la forme '`variable=valeur`', les différentes affectations étant séparées par le caractère '`&`'.
- *Ancre* : en contexte Web également, l'ancre est interprétée par le navigateur. Elle permet, lorsque la page nécessite l'utilisation de l'ascenseur pour être vue en entier, de positionner le début de l'affichage au niveau de l'ancre et non au début de la page.

Ces éléments peuvent fournir en eux-mêmes des informations intéressantes sur les contenus accédés sur Internet, dans la mesure où ils sont fortement corrélés à certains services (voir 3.1, « Les URL, porteuses d'informations » p. 71 pour une exploration de cette piste).

Dans la plateforme d'analyse de trafic développée dans le cadre du projet SensNet, ces éléments sont extraits et renseignent différents champs de la table d'URL dans la base de données :

- *proto* : le protocole utilisé, tel que décrit ci-dessus.
- *site* : contient la partie située entre les '//' et le premier '/' suivant, soit l'agrégation du nom d'utilisateur (très rarement renseigné), du nom de serveur et du port (peu fréquent). Exemples : `www.wanadoo.fr`, `www.lbmicro.com:443`.

- *path* : le chemin vers la ressource sollicitée.
- *file* : le nom de fichier ou de script appelé.
- *query* : l'ensemble des arguments passés au script appelé. Notons que le logiciel ne traite pas les arguments dont le passage répond à une syntaxe autre que celle utilisant le '?' (les ';' dans les pages jsp, les ',' dans certains moteurs de templates, etc.).
- *ref* : le nom de l'ancre spécifiée.

Ce sont ces différents éléments que l'on extrait lors du premier formatage des données de trafic relatives aux URL.

Qu'est-ce qu'un site ?

Ce premier découpage, quoiqu'indispensable, n'est pas encore parfait. Pour l'analyse des parcours sur le Web, nous avons besoin de savoir assez précisément quels sont les différents sites visités dans les sessions, et le champ *site* ici renseigné ne suffit pas toujours à répondre à cet impératif.

En effet, la notion de site, pour intuitive qu'elle soit, se révèle être problématique pour l'analyse. La définition technique qui associerait un site à un DNS (la forme « textuelle » d'une adresse IP), est certes valable dans la majorité des cas, mais rencontre quatre écueils :

- domaines et sous-domaines : certains sites de taille importante se répartissent sur plusieurs DNS, comme par exemple TF1, qui détient toutes les adresses en `tf1.fr`, dont `www.tf1.fr`, `mobiles.tf1.fr`, etc. De la même manière, le LIP6 a un site Web, `www.lip6.fr`, mais aussi un site FTP, `ftp.lip6.fr`, les deux étant intimement liés. Dernier exemple, le site CPAN (Comprehensive Perl Archive Network), accessible sur `www.cpan.org`, propose un service de recherche de modules sur `search.cpan.org`. Il importe de se demander dans quels cas il faut dissocier (`www.tf1.fr` parle de la chaîne de télévision, `mobiles.tf1.fr` est spécialisé sur la téléphonie, et les deux ont des entrées différenciées dans les annuaires du Web) et dans quels autres il faut regrouper (le site CPAN et son module de recherche). L'unité de compte du site peut alors ne pas être le DNS, mais une partie du DNS seulement.
- Problème de réduction : à l'inverse, le DNS peut être beaucoup trop général, c'est particulièrement le cas des sites personnels chez certains hébergeurs, comme Wanadoo, qui pour chaque site personnel fournissent une adresse du type `perso.wanadoo.fr` + `/nomDuSite`, par exemple `http://perso.wanadoo.fr/moto.histo/`. En se basant sur le DNS, on regrouperait l'ensemble des sites personnels de Wanadoo et les services qui y sont attachés dans une même entité.
- Alias : un même site peut avoir plusieurs DNS, par exemple `www.yahoo.fr` et `fr.yahoo.com` qui correspondent à la même adresse IP, `217.12.3.11`, le premier redirigeant systématiquement vers le second.
- sites répartis : certains sites, sous une contrainte de place, se répartissent sur plusieurs « endroits ». C'est un cas observé sur des sites personnels ou semi-

personnels : ainsi, l'auteur du site *Les MP3 de Bibix*[1], accessible sur www.mp3debibix.fr.st qui propose un grand nombre de fichier son et vidéo en téléchargement, a été contraint d'ouvrir des comptes auprès de plusieurs hébergeurs (un chez Multimania, deux chez Free) et de répartir ses volumineux fichiers chez l'un et chez l'autre, ce qui reste transparent pour l'utilisateur. Autre exemple, le site *Les trucs à la con de Nico*[2], accessible par www.trucalacon.net ainsi que par www.trucsalacon.com, propose un nombre important de programmes à télécharger et les stocke sur un compte chez Free (trucsalacon.free.fr) et un autre chez l'hébergeur Worldnet (http://home.worldnet.fr/~nicg/trucalacon/).

Le problème de la définition technique précise de ce qu'est un site est loin d'être anecdotique : il intéresse de près les sites commerciaux soucieux de mesurer leur audience, ce dont dépend leurs tarifs publicitaires – voire leur cotation en bourse. Dans ce cadre, les sociétés de mesure d'audience et les sites font appel en France à un organisme tiers, Diffusion Contrôle[3], qui certifie les méthodologies de mesure de fréquentation des sites. Pour cela, Diffusion Contrôle s'est penché sur la définition des sites, et distingue trois niveaux :

- le *site*, où le site recoupe le *host* hébergeant les ressources ;
- le *portail*, qui regroupe les différents sites d'un même domaine, par exemple les sites de tf1.fr ;
- Le *groupe*, dont les activités peuvent être réparties sur plusieurs sites complètement différents. Ainsi, Caramail fait partie du groupe Lycos depuis que celui-ci l'a racheté, mais l'adresse www.caramail.com reste valide.

Dans la plateforme de traitements SensNet, le module *CatService* répond partiellement à la nécessité de faire ces distinctions, en regroupant et en catégorisant les différentes pages vues sur les grands portails généralistes ou les sites de médias (voir 3.1.3, « Catégorisation semi-automatique avec *CatService* » pour une description complète du fonctionnement de l'outil), mais il est bien difficile de le faire de manière systématique, et il est apparu nécessaire de mettre en place une chaîne de traitement permettant de redéfinir ce qu'est un site à partir d'une URL en tenant compte de l'ensemble de ces contraintes.

Identifier les « sites éditoriaux »

Face à ces problèmes, nous avançons la notion de « site éditorial » : nous considérons un site comme un espace de publication relevant d'une seule entité éditoriale, que ce soit un individu, un organisme, une entreprise. La définition est ici

---

[1] Observé en mars 2002.

[2] Observé en mars 2002.

[3] Voir http://www.diffusion-controle.com/, en particulier le « Bureau Internet et Multimédia » pour les éléments relatifs aux médias électroniques (http://www.diffusion-controle.com/fr/procedures/bim/bim_fr_0.php).

moins capitalistique qu'auctoriale[1], et dans ce cadre, les sites personnels doivent impérativement être distingués de celui de l'hébergeur. Le traitement que nous réalisons pour identifier le « site éditorial » auquel appartient une URL est donc basé sur un traitement différencié entre les sites personnels, bien souvent hébergés par un fournisseur d'accès (Wanadoo, Free, etc.), et les autres sites.

Dans le premier cas, nous avons réservé un traitement précis et poussé au problème de l'agrégation fausse de différents sites, comme c'est le cas sur certains sites personnels. Nous l'avons dit, l'hébergeur Wanadoo place l'ensemble de ses sites personnels sous le *host* `perso.wanadoo.fr` ; réduire le site au nom de domaine amènerait ainsi à assimiler l'ensemble des sites personnels de Wanadoo à une seule et même entité – agrégat problématique, dont l'audience forte autant que la diversité des contenus parasiteraient fortement les analyses.

Pour ce faire, nous avons développé un programme capable de traiter le problème de la réduction. Il s'agit, dans ce composant logiciel, de proposer un découpage des URL qui tiennent compte de la notion de site éditorial pour les pages personnelles, et renvoie pour chaque URL, un « site éditorial » et un « chemin éditorial » répondant à ces définitions. Deux champs sont concernés dans la table des URL visitées, *editorial_host* et *editorial_path*, renseignés de la façon suivante. Après identification des URL relevant des hébergeurs de pages personnelles, des règles particulières de découpage sont appliquées en fonction de chaque hébergeur, sur la base d'expressions régulières *ad hoc*.

Pour cela, nous avons dressé une liste aussi exhaustive que possible des hébergeurs de pages personnelles[2], que nous avons classés en fonction de la syntaxe des adresses des sites personnels qu'ils abritent. Trois groupes d'hébergeurs ont été identifiés sur cette base :

- DNS spécifique à chaque site personnel : c'est par exemple le cas pour Free, dont les adresses de sites personnels sont de la forme `[nom-du-site].free.fr`. Dans ce cas, le champ *editorial_host* équivaut au champ *site*.
- l'adresse du site est rattaché à un DNS générique suivi d'un nom de répertoire particulier à chaque site, comme sur Wanadoo : `perso.wanadoo.fr/[nom-du-site]`.
- DNS générique suivi d'un nombre complexe et variable de répertoires dont le nommage revient à l'hébergeur, suivi du nom du répertoire contenant le site personnel. C'est le cas des sites hébergés par Geocities et AuFeminin.

Pour les pages ne relevant pas de la catégorie des sites personnels, le problème est, à l'inverse, de l'ordre de la scission. Pour retrouver une équivalence entre le site et l'autorité éditoriale, on s'efforce dans ce cas de se rapprocher du nom de domaine tel

---

[1] Nous nous éloignons ici de la définition proposée par Diffusion Contrôle, pour qui « Un Site Web correspond à une entité éditoriale disponible sur l'Internet, et placée sous la responsabilité d'un Editeur », et qui privilégient ainsi plutôt la notion d'éditeur que d'auteur, avec les questions juridiques qui se profilent en arrière-plan.

[2] Nous en avons dénombré plus d'une quarantaine en 2003.

qu'il peut être acheté et déposé par un individu, une société ou un organisme. Pour cela, on part du domaine de premier niveau (*Top Level Domain*, ou TLD) et on inclut le nom qui précède, par exemple : `www.koodpo.com` est transformé en `koodpo.com`.

Au sein des TLD, il existe une distinction entre les TLD génériques du type `.com`, `.org`, etc., qui correspondent – en principe – à une classification plutôt thématique, et les TLD par pays (`.fr`, `.be`, `.uk`, etc.). Dans les deux cas, il est possible de détenir un nom de domaine immédiatement sous l'arborescence du TLD, du type `monsite.com` ou `monsite.fr`, mais certains sous-domaines sont réservés et il est nécessaire de descendre d'un pas dans l'arborescence pour accéder au site éditorial. Ce repérage n'est pas aisé : pour les TLD génériques, les sous-domaines réservés sont assez bien renseignés, mais pour les TLD par pays, chaque État dispose d'une autorité de gestion indépendante qui est libre de ses choix et ne les documente pas toujours. Nous avons donc identifié les sous-domaines réservés suivants à partir des données et de la documentation lorsque celle-ci est disponible :

TLD	Sous-domaines réservés identifiés
`.fr` et `.re`	`tm.fr`, `st.fr`, `asso.fr`, `com.fr` (*idem* pour `.re`)
`.uk`	`co.uk`, `me.uk`, `org.uk`, `ltd.uk`, `plc.uk`, `net.uk`, `sch.uk`, `ac.uk`
`.ca`	`ab.ca`, `bc.ca`, `mb.ca`, `nb.ca`, `nf.ca`, `ns.ca`, `nt.ca`, `on.ca`, `pe.ca`, `qc.ca`, `sk.ca`, `gc.ca`
`.com`	`br.com`, `cn.com`, `de.com`, `eu.com`, `gb.com`, `hu.com`, `no.com`, `qc.com`, `ru.com`, `sa.com`, `se.com`, `uk.com`, `us.com`, `uy.com`, `za.com`
`.net`	`gb.net`, `se.net`, `uk.net`
`.st` et `.fm`	`fr.st`
`.be`	`ac.be`
`.jp`	`ad.jp`, `ac.jp`, `co.jp`, `go.jp`, `or.jp`, `ne.jp`, `gr.jp`, `ed.jp`, `lg.jp`, `geo.jp`
`.tw`	`com.tw`, `edu.tw`
`.ru`	`com.ru`, `net.ru`, `org.ru`, `pp.ru`, `by.ru`
`.to`	`go.to`

En complément, nous avons tenu à distinguer les différents ministères au sein des sites gouvernementaux français en `.gouv.fr` : `finances.gouv.fr`, `interieur.gouv.fr`, etc.

Cette méthode permet de regrouper de manière très efficace les différentes pages vues sur un même portail, en particulier dans le cas des portails généralistes. Le Tableau 2.1 en donne un exemple pour les sites du Ministère des finances, du Crédit Lyonnais et d'Ebay UK ; dans le cas de Voila, ce sont 199 domaines distincts qui sont regroupés sous l'entité `voila.fr`, 87 dans le cas d'`aol.fr`.

Tableau 2.1. Exemples de regroupement en sites éditoriaux

Site éditorial calculé	Domaines regroupés (données de trafic : France 2002)

`finances.gouv.fr`	`alize.finances.gouv.fr`
	`alize2.finances.gouv.fr`
	`concours.douane.finances.gouv.fr`
	`lekiosque.finances.gouv.fr`
	`tarif.douane.finances.gouv.fr`
	`www.deb.douane.finances.gouv.fr`
	`www.dpa.finances.gouv.fr`
	`www.finances.gouv.fr`
	`www.icp.finances.gouv.fr`

	www.telepaiement.cp.finances.gouv.fr www2.finances.gouv.fr www3.finances.gouv.fr www4.finances.gouv.fr
creditlyonnais.fr	abcl.creditlyonnais.fr ABCLnet.creditlyonnais.fr access.creditlyonnais.fr e.creditlyonnais.fr gro.creditlyonnais.fr interactif.creditlyonnais.fr sherlocks.creditlyonnais.fr www.abclnet.creditlyonnais.fr www.access.creditlyonnais.fr www.creditlyonnais.fr www.e.creditlyonnais.fr www.finance.creditlyonnais.fr www.interactif.creditlyonnais.fr www.particuliers.creditlyonnais.fr www.professionnels.creditlyonnais.fr
ebay.co.uk	cgi.ebay.co.uk cgi1.ebay.co.uk cgi2.ebay.co.uk cgi3.ebay.co.uk cgi6.ebay.co.uk cq-search.ebay.co.uk ebay.co.uk listings.ebay.co.uk pages.ebay.co.uk search.ebay.co.uk search.stores.ebay.co.uk www.ebay.co.uk www.stores.ebay.co.uk

Cette redéfinition du site se révèle très utile au niveau méso-analytique de notre travail : si le problème des alias n'est ici pas traité, nous estimons que nous pouvons, avec les données ainsi obtenues, travailler de manière fiable sur les parcours à l'intérieur d'un site ou, à un autre niveau d'analyse, sur les différents sites visités et leur agencement. C'est donc au résultat de ce calcul que nous ferons référence par la suite lorsque nous parlerons de site.

> ***Synthèse.*** *Le rattachement de chaque* URL *à un site donné ne peut se satisfaire d'une assimilation au serveur désigné dans l'*URL*. Un module spéficique permet de reconnaître le* site éditorial*, qui fait correspondre un site à un auteur ou une entité qui a la responsabilité de son contenu.*

2.2.3 Recomposer les pages

Si nous avons, au terme des traitements décrits jusqu'ici, traité le problème de la définition des sites et, ce faisant, proposé un nouveau découpage des URL plus représentatif des entités de production que ne l'était un simple découpage « technique », le niveau intermédiaire que représente la page pose encore problème. En effet, une page Web, unité ergonomique objective pour l'utilisateur, est le fruit d'une composition d'éléments hétérogènes qu'il importe de regrouper au sein des données de trafic.

La page Web, unité ergonomique inaccessible

Une page Web est le résultat de l'assemblage d'éléments hétérogènes, dont la production est assurée par un ou plusieurs serveurs Web, et la composition par le navigateur. Nous avons déjà aperçu ce phénomène dans l'extrait de données de trafic proposé ci-dessus (Figure 2.3. Extrait de données de trafic Web, p. 52).

Du côté du navigateur, le format de base du Web, HTML, contient les instructions nécessaires à la collecte des différents composants ainsi qu'à la mise en forme de l'ensemble ; pour chaque image, par exemple, le navigateur lance une requête auprès d'un serveur Web pour récupérer le fichier et l'intègre ensuite à la page. L'implication de ce dispositif est double : en premier lieu, la page Web apparaît comme une construction d'éléments dont la source n'est pas forcément unique, et dont la nature peut être variée (images, sons, texte, etc.), bref, le contenu en est éminemment polysémiotique. D'autre part, le navigateur est un élément très actif de la navigation : d'une action de l'utilisateur il génère une série de requêtes, et exécute des instructions contenues dans les pages ou les en-têtes HTTP, qui vont de l'inclusion d'éléments dans une page à la redirection automatique, la mise en place de cookies ou l'ouverture d'une ou plusieurs fenêtres. La Figure 2.6 décrit schématiquement les différentes opérations effectuées par le navigateur auprès des serveurs Web et de l'utilisateur pour une requête Web classique.

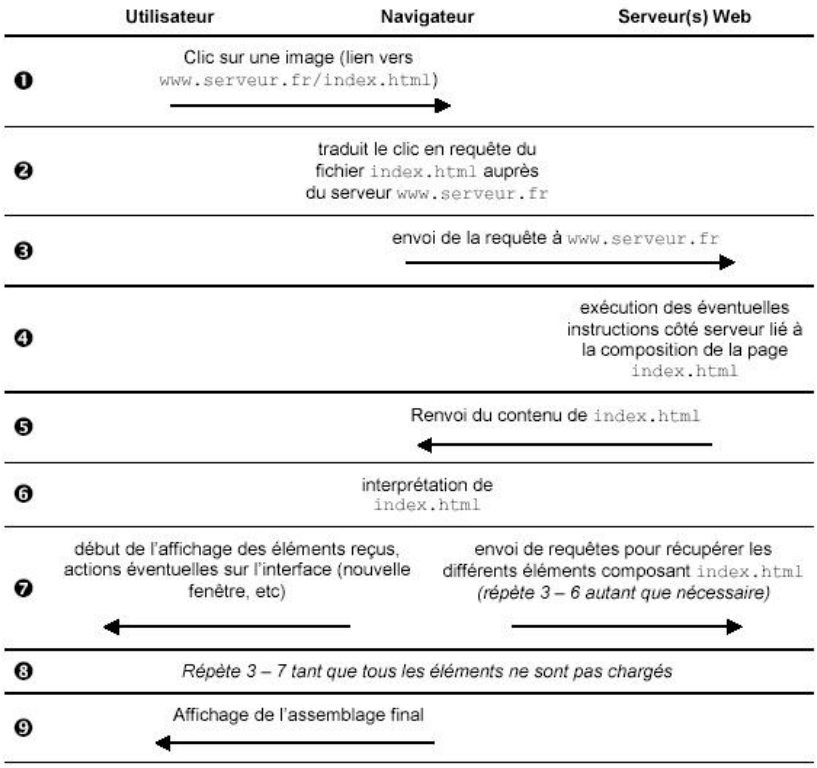


Figure 2.6. Requête d'une page auprès d'un serveur Web

Nous avons vu que les navigateurs sont également capables de faire appel à des programmes tiers, par un système de *plugin*, pour l'exécution de certaines tâches ou l'affichage de certains documents (vidéo en RealMedia, multimédia en Flash, etc.). De ce fait, les modalités d'interactions de l'utilisateur avec le navigateur ainsi que les types de contenus accessibles par le Web se trouvent démultipliées, et la notion de page doit être revisitée.

Le dispositif de recueil de données est, de ce point de vue, assez imparfait puisqu'il trace l'ensemble des requêtes produites par chaque composant de la page sans distinguer quelle requête source est à l'origine des autres. En outre, les actions relatives aux interfaces « animées » nous échappent : c'est bien sûr le cas pour certains formats comme les animations flash, mais également pour les pages HTML dont l'emploi de Javascript permet d'afficher et de masquer certains éléments. À titre d'exemple, le bandeau de navigation de Wanadoo peut prendre sept états différents en fonction du placement du pointeur sur certaines zones (voir Figure 2.7), sans que cela apparaisse dans les données, aucune requête n'étant générée à cette occasion en dehors des fichiers GIF impliqués dans la mise en page.

Figure 2.7. Les sept états possibles du bandeau de navigation de Wanadoo (nov. 2003)

Les sondes orientées trafic plutôt qu'interfaces perdent donc de vue la page, et ne conservent des actions sur les interfaces que ce qui est générateur de communication avec des serveurs distants.

Solutions partielles

On se voit, face à ce problème, contraint de tenter de reconstruire la page à partir de ses composants, ou *a minima* d'écarter le « bruit » dans les données pour se rapprocher le plus possible d'une correspondance entre requête et page vue. Plusieurs types de problèmes techniques sont soulevés, dont les solutions ne sont souvent que partielles.

En premier lieu, il est de bon aloi de supprimer des données les requêtes pointant explicitement vers des images ou des fichiers dont on est sûr qu'ils ne constituent pas des sources de pages. Ce filtrage est aisé, puisqu'il consiste à repérer les extensions de fichiers correspondant à des formats de fichiers précis, dont le nombre est réduit : 'jpg' ou 'jpeg' pour le format JPEG, 'gif' pour le format GIF, 'css' pour les feuilles de style, 'js' pour les javascripts externes, etc. Ce dispositif est d'ailleurs intégré dans les sondes, qui ne transmettent pas les données relatives à la plupart de ces fichiers.

Toutefois, certaines images peuvent être le résultat de requêtes effectuées à destination de scripts, avec le passage de paramètres particuliers, impossibles à repérer comme fichiers images sur la base de l'URL. C'est presque systématiquement le cas des bandeaux publicitaires, qui pointent le plus souvent vers des serveurs externes spécialisés dans la fourniture de ce type de service : *doubleclic*, *adserver*, etc. C'est pour éviter ce problème, et s'approcher autant que faire se peut d'une équivalence entre requêtes enregistrées et unités ergonomiques perçues, que l'on filtre dans les données les pages correspondant à des bannières publicitaires et des serveurs de comptage destinés à la mesure d'audience (filtrage sur la base d'une liste de domaines dédiés). Nous écartons également les requêtes envoyées par les barres d'outil intégrées au navigateur, du type Google Toolbar[1], qui envoient automatiquement des requêtes, pour renseigner par exemple, dans le cas de Google, le *PageRank* de la page visitée par l'utilisateur. Ce faisant, on écarte un volume conséquent de requêtes : dans les données dont nous disposons, ces requêtes parasites représentent 8 à 9 % du trafic des internautes.

Un autre biais technique vient parasiter nos données, celui des *frames* (ou « cadres ») Ce mécanisme permet d'intégrer dans une même fenêtre de navigateur plusieurs pages distinctes, qui forment une unité ergonomique pour l'utilisateur. Ainsi, la page d'accueil de Wanadoo en 2002 était composée de quatre pages HTML différentes (voir Figure 2.8) appelées par une seule page vide de tout contenu qui en détermine l'assemblage, appelée *frameset*. Au total, ce sont cinq requêtes pour une seule page, d'autant plus difficile à déceler qu'elles pointent toutes vers des fichiers HTML, donc potentiellement de véritables pages autonomes.

---

[1] Voir `toolbar.google.com`.

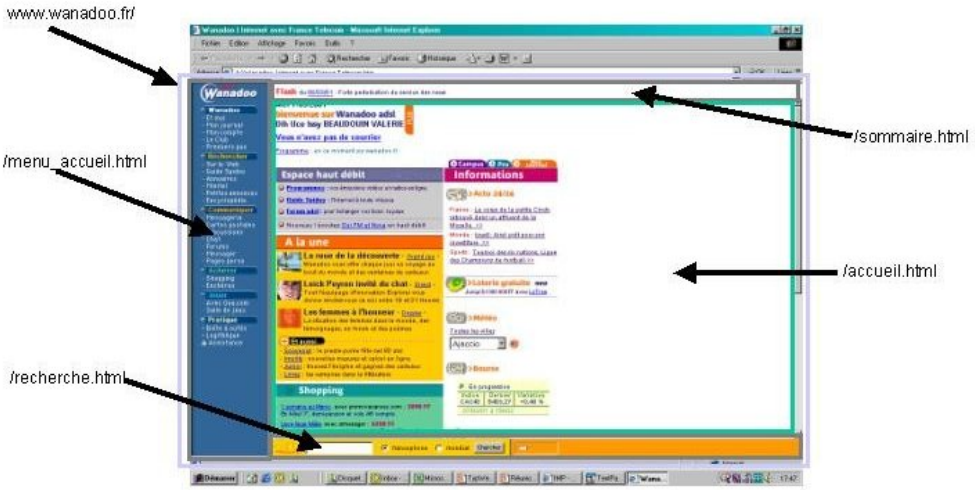


Figure 2.8. Les quatre frames *composant la page d'accueil de Wanadoo (2002)*

Ce problème est à ce jour quasiment impossible à régler de manière satisfaisante sur la base de données de trafic : ni l'exploitation du champ *Referer* dans les en-têtes HTTP, qui renseigne l'URL à l'origine des requêtes, ni la détection de rafales de requêtes ne permettent de distinguer systématiquement si les requêtes correspondent à des pages ou à des composants de pages.

Dans ces conditions, la page en tant qu'unité micro-analytique nous échappe souvent ; nous sommes réduits à postuler qu'une requête équivaut, après un certain nettoyage, « plus ou moins » à une page, sans pouvoir évaluer précisément les biais induits par cette approximation. Toutefois, ce biais invite fortement à rester très prudent dès qu'il s'agit de compter des « pages » là où l'on ne compte que des requêtes, les chiffres obtenus dépendant fortement des choix de conception des développeurs Web, ce qui les rend difficilement comparables. On sera bien plutôt tenté de s'appuyer sur les durées de visite, ainsi que sur des échelles d'analyse plus agrégées au niveau du site.

> ***Synthèse.*** *Une page, en tant qu'unité ergonomique élémentaire perçue par l'utilisateur, peut correspondre à plusieurs URL dans les données de trafic, chaque URL désignant un des composants de la page. Le filtrage de certains types de fichiers dans les données résout partiellement ce biais, mais le problème des* frames *reste entier.*

Conclusion

Au terme de ce chapitre, on constate que le choix de tel ou tel dispositif de recueil de données d'usages d'Internet conditionne pour beaucoup les interrogations que l'on va pouvoir soumettre par la suite aux données recueillies. Pour le cas de données de trafic, celles sur lesquelles nous travaillons, on se trouve à un niveau intermédiaire qui donne un panorama complet de l'ensemble des usages liés à des accès au réseau (Web, messagerie électronique, jeu, *peer-to-peer*, messagerie instantanée, etc.) tout en

les reliant à un utilisateur particulier. Ce point de vue induit à la fois un rapprochement et un éloignement par rapport à celui de l'utilisateur : d'un côté, il permet d'appréhender l'activité Internet dans son ensemble, et de percevoir les éléments de continuité entre les différents outils utilisés par l'internaute. Revers de la médaille, on se trouve en même temps éloigné des interfaces, et l'on ne trace que la communication avec l'extérieur, parfois périphérique à l'action elle-même. Dans le cas du trafic Web, cette distorsion nous empêche de connaître précisément l'interaction avec les pages proposées, leur mise en forme, l'utilisation des fonctionnalités ergonomiques (ascenseur, impression, sauvegarde, multifenêtrage, etc.), et nous oblige à manipuler une approximation de la page à partir de la requête.

Néanmoins, ces données ont l'avantage, pour le Web, de couvrir l'ensemble des navigateurs disponibles, et de rendre compte de la totalité du trafic effectué avec précision et exhaustivité. À terme, ce sont des données précieuses et éminemment exploitables que l'on obtient. Le passage de données brutes à des données formatées pour l'analyse attire d'ores et déjà l'attention sur une série de problèmes liés aux dispositifs techniques qui sous-tendent la publication de contenus sur le Web. Des traitements adaptés, en particulier pour l'identification de « sites éditoriaux », sont nécessaires dès cette étape. Au terme de ce travail de mise en forme, on dispose d'une base de données de trafic prête à l'emploi, que l'on va tenter d'enrichir à l'aide de descriptions relatives à la forme et au contenu des parcours afin de répondre aux questions que nous nous posons sur la navigation.

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.

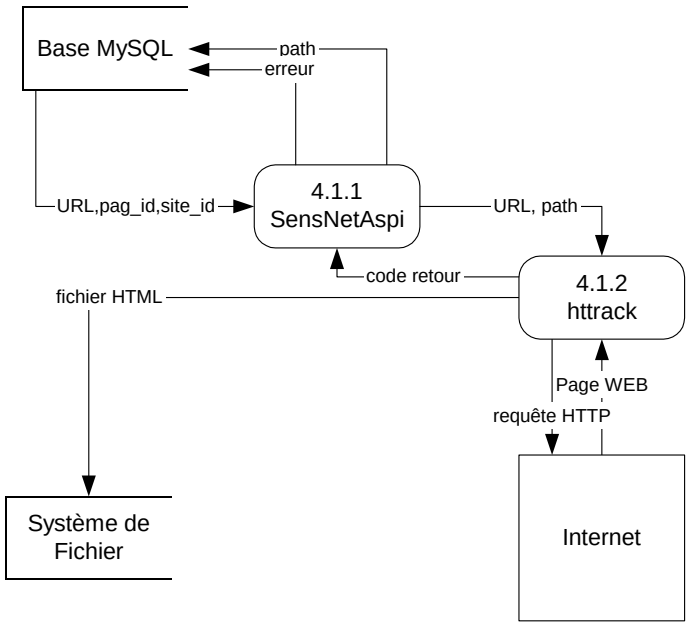


Figure 3.2. Fonctionnement du module SensNetAspi

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.

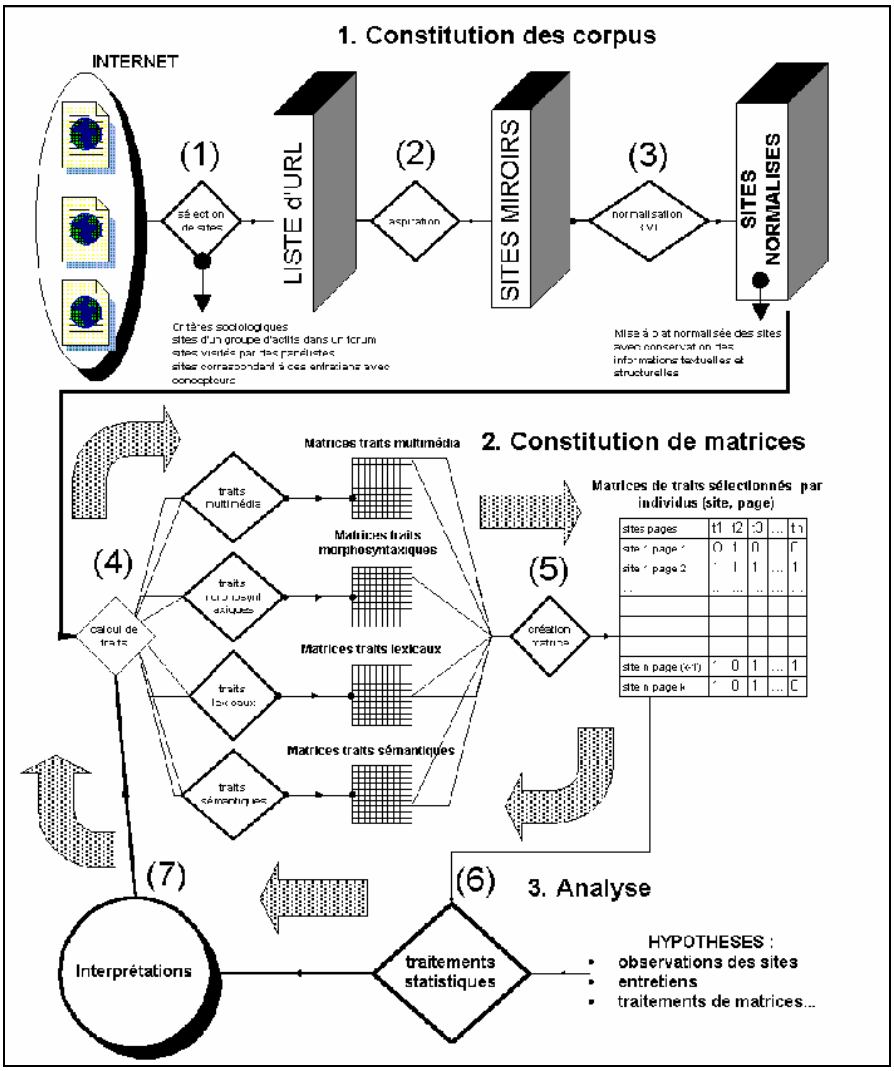


Figure 3.3. Système de catégorisation de sites et de pages dans TypWeb (présenté dans [Beaudouin et al. 2001])

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 :

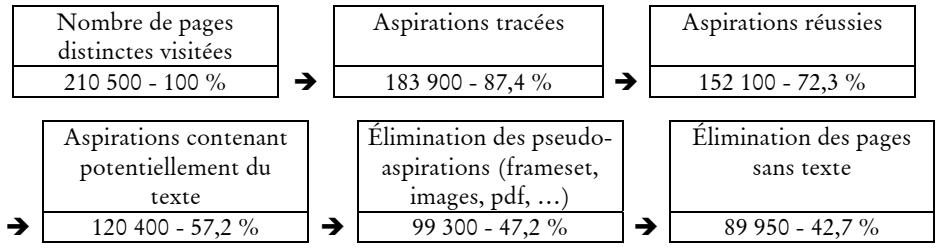
```
REFRESH(0 sec): http://sunearth.gsfc.nasa.gov/evsecef.htm
Click here...
```

  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.

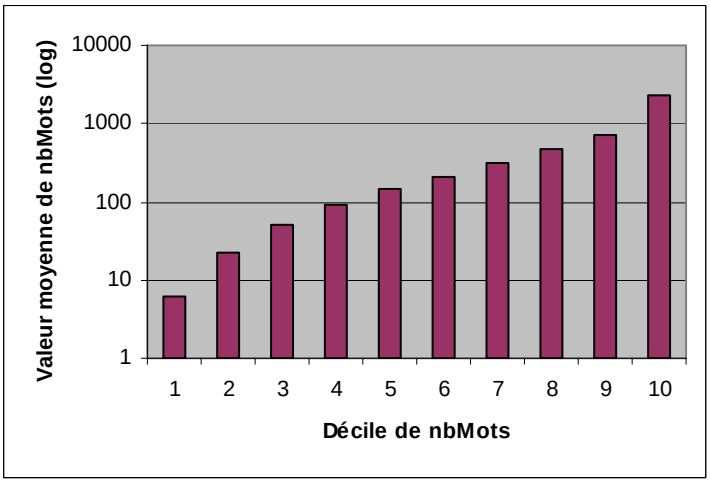


Figure 3.4 . Nombre moyen de mots par décile du nombre de mots (échelle logarithmique)

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).

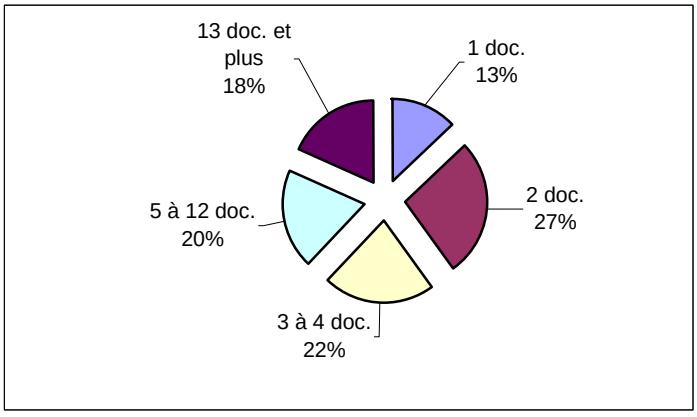


Figure 3.5. Représentation des formes de fréquence supérieure à 1 dans les documents

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).

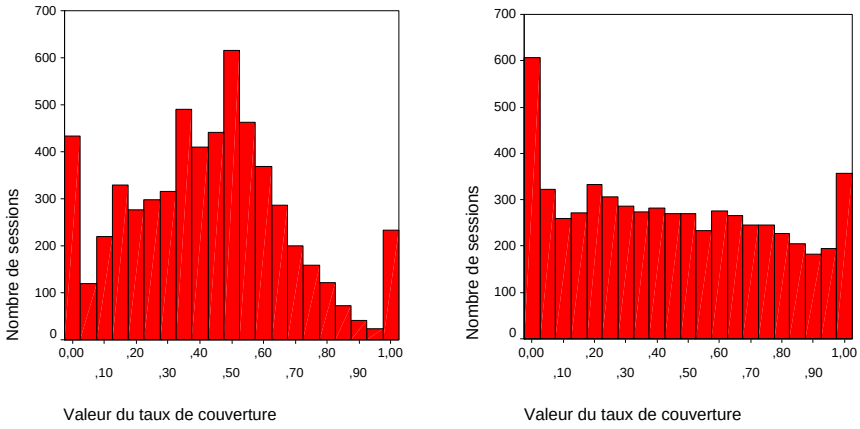


Figure 3.6. Répartition des valeurs du taux de couverture par session en nombre d'URL (à gauche) et en durée (à droite)

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).

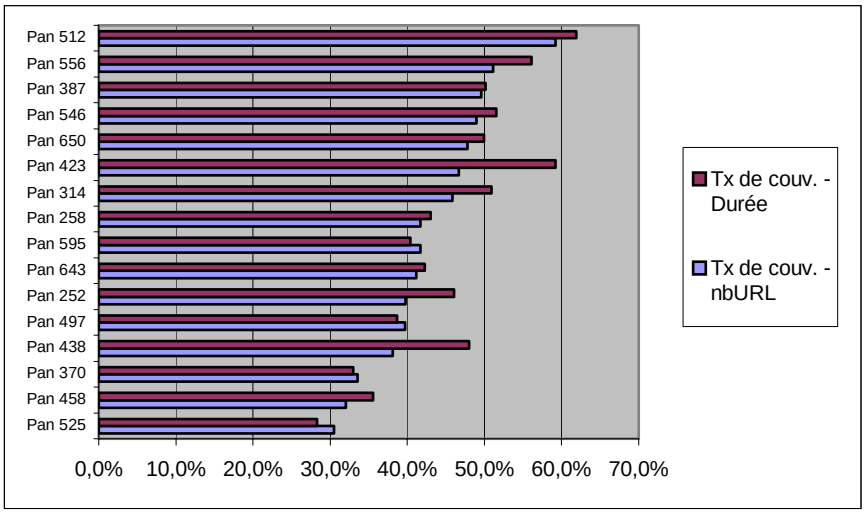


Figure 3.7. Couverture moyenne des sessions de chaque panéliste BibUsages

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.

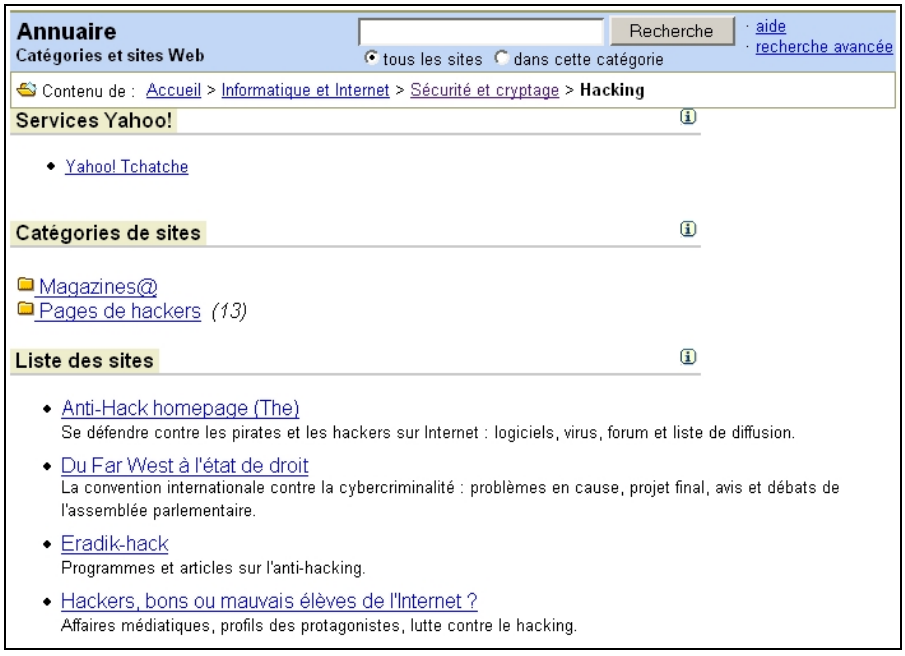


Figure 3.10. Un exemple d'annuaire : Yahoo France

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).

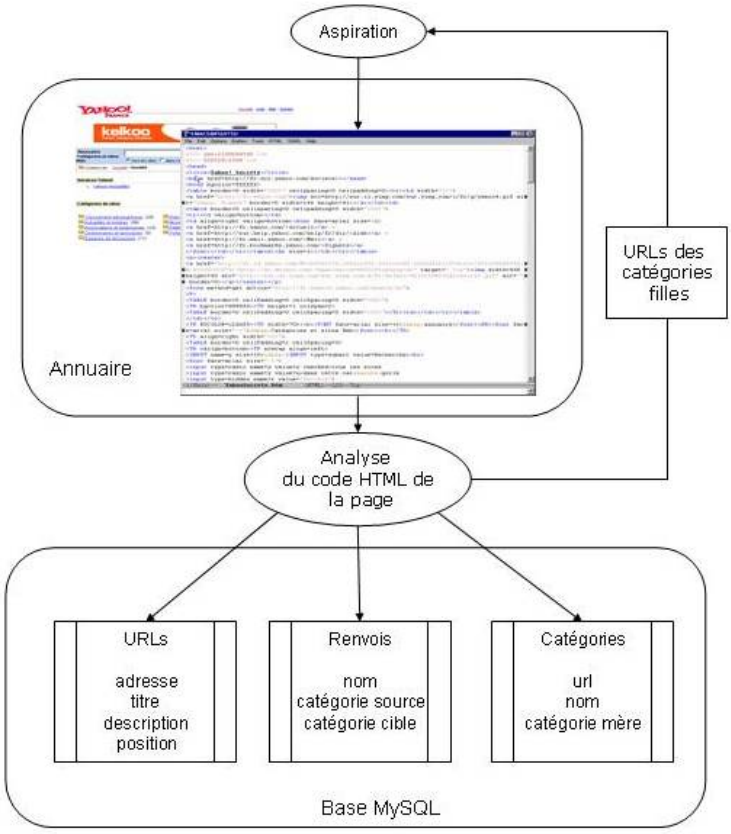


Figure 3.11. Aspiration des annuaires : fonctionnement général

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

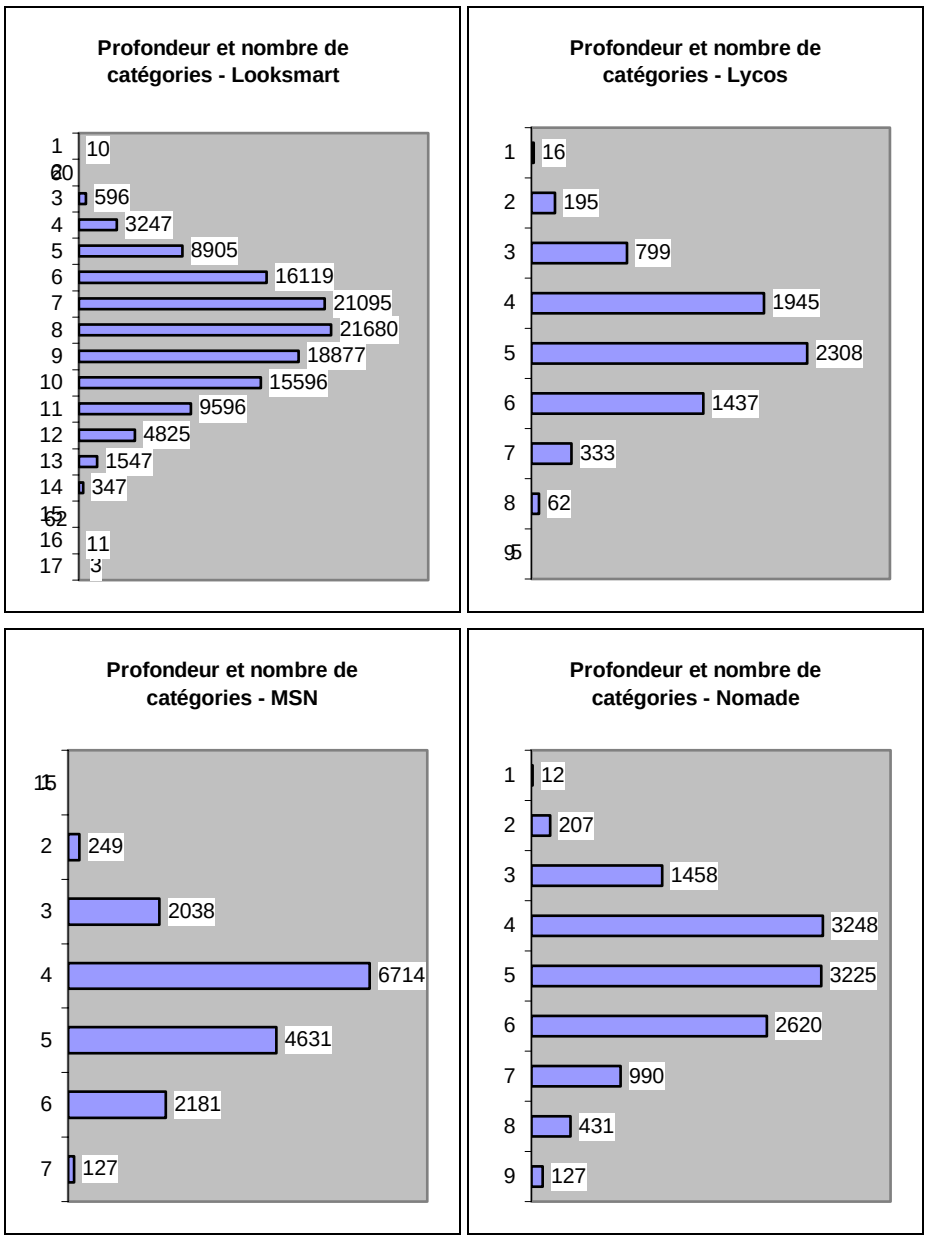
La profondeur d'un annuaire n'est donc pas directement liée au fait d'avoir un nombre important d'URL à présenter, mais semble plutôt résulter d'un choix organisationnel. Cette hypothèse est confirmée par l'examen du nombre moyen d'URL indexées par catégorie comportant au moins une URL (Tableau 3.19) : 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.

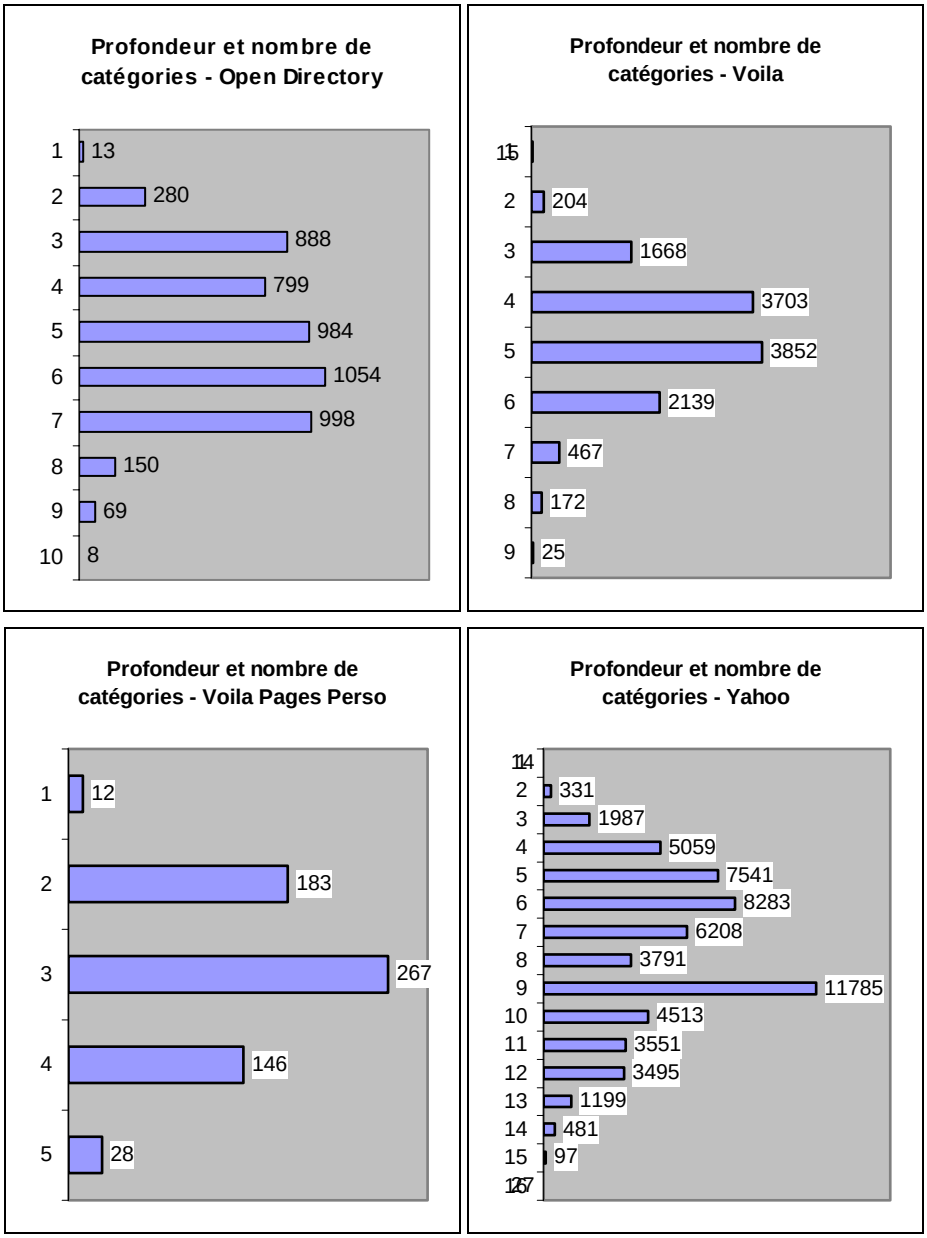
Tableau 3.19. Indexation des URL en 2002

	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.

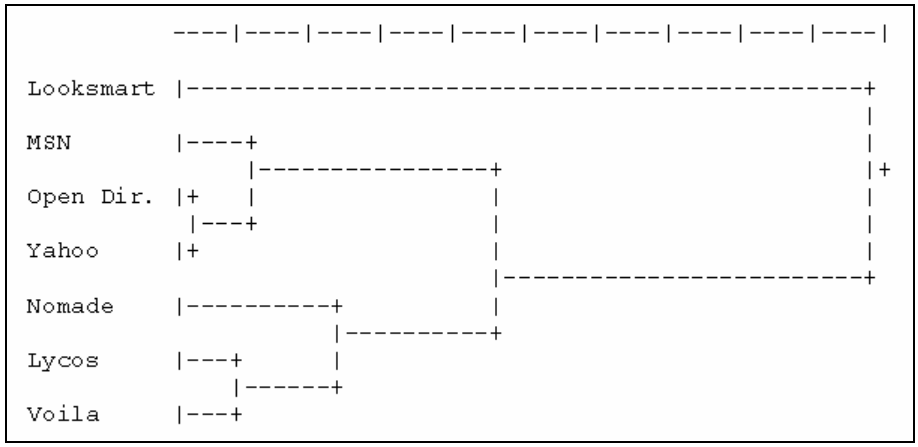


Figure 3.13. Classification des sept annuaires généralistes en 2002 sur la base des descriptifs des sites de la catégorie « Commerce et économie »

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.

Chapitre 4

Décrire et visualiser la dynamique des parcours

L'ancrage temporel et séquentiel des parcours sur le Web nous interdit de nous limiter à l'analyse des contenus visités : de la même manière qu'un texte n'est pas un sac de mots, une session de navigation n'est pas une collection de pages, mais un cours d'action où la visualisation de chaque page prend sens dans la dynamique générale du parcours. La sémantique des parcours se doit de rendre compte de cette dynamique : pour cela, nous proposons à la fois de recourir à des outils de fouille manuelle pour travailler au plus près de l'objet, ainsi que des éléments de représentation statistiques de la topologie des parcours. Pour cela, nous opposons aux travaux menés jusqu'alors sur la navigation une approche descriptive se situant dans le cadre des sciences humaines, dans le cadre d'une théorie de l'action. À ce titre, la session doit être replongée dans le contexte individuel de chaque internaute : ses pratiques du Web, ses usages d'Internet en général, pour autant que les déterminations globales de l'activité de navigation influencent directement les éléments locaux – forme, contenu.

4.1 Outils de fouille des données

L'analyse de données de trafic volumineuses ne doit pas sacrifier au « tout statistique » : il est indispensable de pouvoir mener un examen manuel des parcours pour en appréhender la logique et la complexité. Pour répondre à ce besoin, nous avons développé deux outils de manipulation des parcours qui trouvent naturellement leur place dans l'outillage d'analyse de la navigation.

4.1.1 Rejouer les parcours

Nous avons, face aux données de trafic, rapidement éprouvé le besoin de pouvoir « refaire » des parcours, c'est-à-dire de revoir dans l'ordre de la visite les différentes URL composant une session. Nous avons pour cela développé un outil baptisé *RePlay*

qui permet, pour une session donnée, de visiter l'ensemble de ses pages dans l'ordre et la temporalité d'origine.

L'application consiste à fournir une interface HTML pour reproduire la session dans la fenêtre d'un navigateur (voir Figure 4.1). Utilisant le mécanisme des *frames*, elle propose, dans une partie gauche de l'écran, la liste ordonnée des noms de sites visités dans une session ; chacun de ces éléments contient un lien hypertexte vers l'adresse de la page vue par le panéliste, qui s'affiche dans le *frame* de droite.

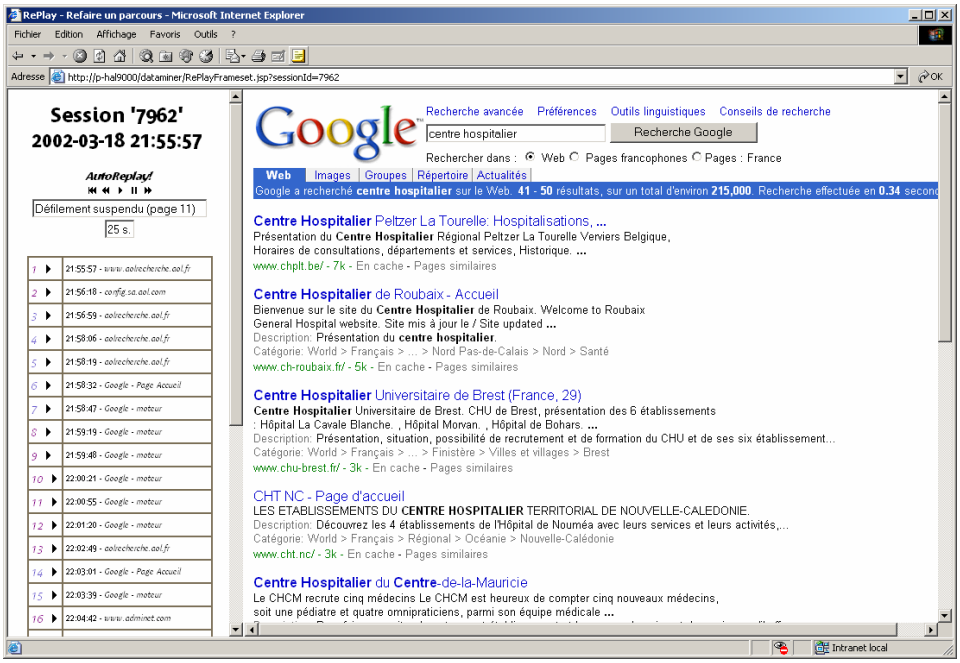


Figure 4.1. Interface de RePlay – vue générale

Deux modes d'utilisation sont proposés : dans le premier, l'utilisateur de *RePlay* clique sur chaque lien pour ouvrir la page visitée par l'internaute. Cet accès permet d'aller rapidement aux passages jugés intéressants et de vérifier rapidement certains contenus, ou de passer du temps sur d'autres. Le deuxième mode reproduit automatiquement les requêtes une à une en respectant les délais observés entre deux requêtes par l'internaute. L'utilisateur de *RePlay* regarde ainsi « défiler » la session sous ses yeux, et bénéficie de l'effet de temps passé sur chaque page.

Ces deux modes ne sont pas antagonistes : il est possible de lancer le défilement automatique à partir de n'importe quelle page de la session, et de l'interrompre à tout moment. Une console de commande *AutoRePlay* (voir Figure 4.2 ci-dessous) permet de contrôler le défilement, de le suspendre ou de le relancer à tout moment, et de forcer le passage à la page suivante, sur le mode de la lecture des pistes d'un CD audio. En outre, cette console affiche l'URL visitée, et le temps total que l'internaute a passé dessus ; certains temps de visualisation pouvant être très longs (en théorie, jusqu'à trente minutes étant donné le mode de calcul des sessions), la durée effective entre deux pages a été positionnée à 30 secondes maximum en mode défilement.

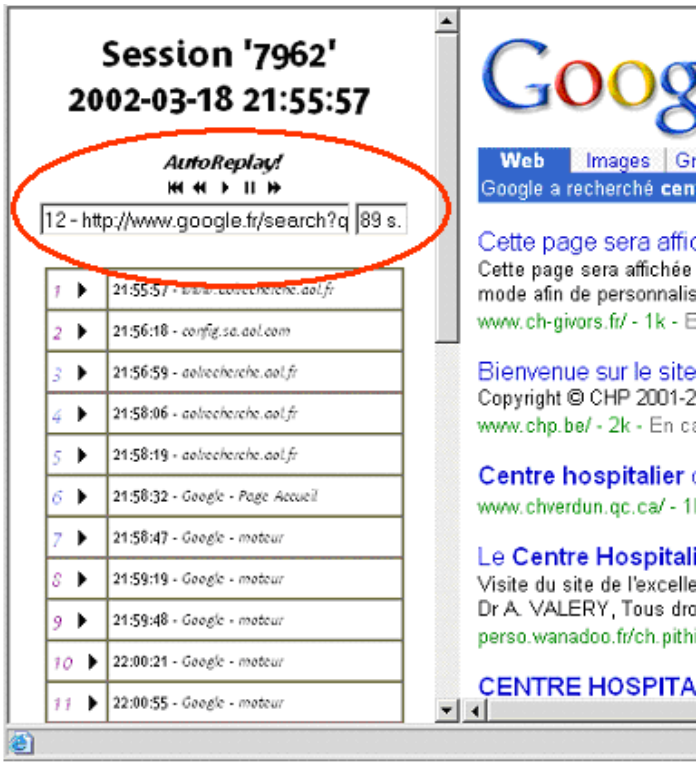


Figure 4.2. Interface de RePlay – détail

Bien évidemment, cette interface ne fait pas abstraction du problème du renouvellement ou de la disparition des pages Web, dans la mesure où *RePlay* effectue une requête en différé vers des URL accédées plusieurs mois auparavant ; en outre, on se heurte également au problème des accès restreints et authentifiés. Bref, nous rencontrons ici les mêmes aléas que lors de l'aspiration de pages (voir chapitre 3.2). En outre, le problème des *frames* n'est pas résolu par ce mode de représentation : comme nous l'avons vu au Chapitre 2, l'unité ergonomique perçue par l'utilisateur peut être le fruit de plusieurs requêtes, notamment lorsque la page affichée contient des pages imbriquées par le mécanisme des *frames*. Dans ce cas, *RePlay* effectue une requête par élément, et affiche le contenu de chaque *frame* séparément, ce qui nuit à la lisibilité globale du parcours.

Cet outil a été développé sous la forme de servlets Java qui interrogent directement la base de données de trafic, pour être intégré à la plateforme de fouille de données de trafic SensNet, et reconstruit dynamiquement une interface à chaque session demandée. Cette application nécessitant d'être connecté à cette base, une version *stand-alone* a été développée, qui permet, sous la forme de deux fichiers HTML, de transporter les résultats de l'application et de refaire les parcours depuis n'importe quelle machine disposant d'un accès Internet.

> ***Synthèse.*** *Le module* RePlay *permet de « rejouer » une session en revoyant en différé le contenu d'un parcours dans l'ordre et la temporalité de la visite par l'internaute. Si le module se heurte aux problèmes de l'évolution*

des pages et de l'accès restreint, il est néanmoins un puissant outil pour formuler et vérifier des hypothèses sur les parcours, et mettre à jour les logiques à l'œuvre dans les sessions.

4.1.2 Représentation graphique

L'intérêt d'une représentation synthétique des sessions sous forme graphique est double : d'une part, elle permet de formuler et de vérifier des hypothèses sur les parcours et les liens entre forme et contenu. D'autre part, la visualisation graphique d'une navigation a une valeur didactique qui n'est pas à négliger dans la présentation des travaux sur les parcours.

Pour répondre à ce besoin, nous avons développé deux outils capables de représenter les sessions sous la forme d'un graphe : dans les deux cas, les nœuds du graphe sont les pages ou les sites visités, et les arcs orientés représentent le passage d'une page ou d'un site à l'autre. Afin de rendre les graphes ainsi obtenus plus lisibles, nous avons représenté différemment les URL ou les sites taggués comme services dans les portails généralistes : ce n'est alors plus l'adresse, mais le nom du portail et le service, dans le cas d'un graphe d'URL, qui sont affichés. Quatre types de graphes sont produits, en fonction de l'échelle d'analyse à laquelle on se place :

- graphe de pages : les nœuds représentent les URL visitées ;
- graphe de pages/services : les nœuds représentent les URL visitées, mais les URL correspondant à un service sur un portail identifié par *CatService* sont regroupées et représentées comme telles, et sont colorées ; en outre, s'il s'agit d'un moteur de recherche, le contenu de la requête est affiché, et les différentes pages de résultat sont regroupées ;
- graphe de sites : les nœuds représentent les sites visités, et chaque nœud agrège l'ensemble des URL visitées sur un même site ;
- graphe de sites/services : comme dans le cas précédent, les nœuds représentent les sites, mais si ce site est un portail identifié par *CatService* et qu'un service est spécifié, on distingue dans le graphe les différents services utilisés sur le portail, qui donnent lieu à autant de nœuds différents.

Ne pouvant ni ne souhaitant développer un algorithme de mise en forme des graphes – champ de recherche qui dépasse de très loin notre étude – nous avons fait appel à deux solutions existantes pour la mise en forme.

Solution Graphlet

Le premier outil employé pour la mise en forme de graphes de sessions est le logiciel Graphlet[1]. Notre outil extrait des données de trafic, en fonction de la granularité souhaitée, la liste ordonnée des URL ou des sites d'une session et la date de chaque requête ; à partir de cette liste, il prépare un fichier au format supporté par Graphlet, où sont spécifiés les nœuds, les arcs, leurs labels, les couleurs des nœuds,

[1] Logiciel développé par l'Université de Passau (Allemagne) ; voir `http://www.infosun.fmi.uni-passau.de/Graphlet/`.

mais aucune coordonnée ; il faut ensuite ouvrir le fichier dans Graphlet et appliquer à notre graphe un des algorithmes de mise en forme proposés.

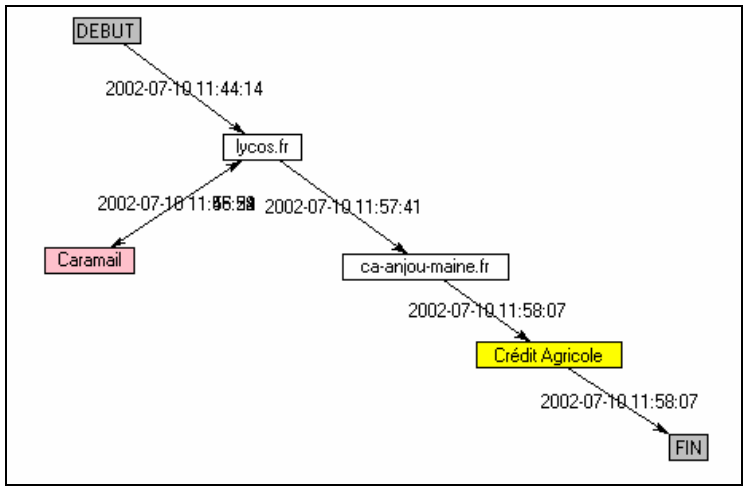


Figure 4.3. Exemple de graphe de session à l'échelle du site (SN2002, session 127666)

La Figure 4.3 présente un exemple de sortie obtenue après ce traitement, au niveau de granularité du site. La représentation permet d'apprécier la grande différence en termes de linéarité entre l'analyse à l'échelle de la page et à celle du site : la Figure 4.4 représente la même session, au niveau de la page cette fois. On constate que, si une linéarité globale se dessine, quatre « déviations » sont opérées, contre une seule au niveau des sites.

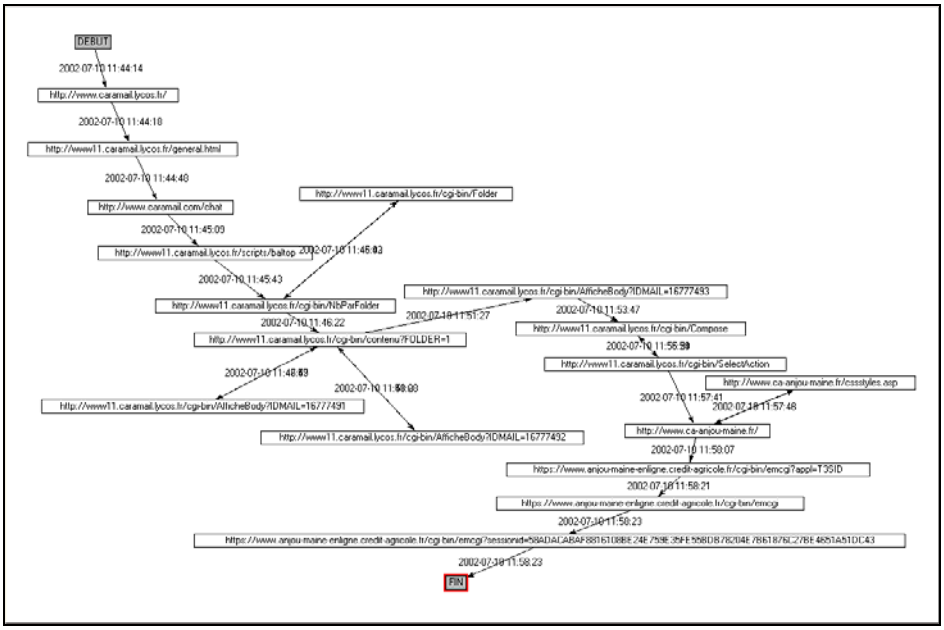


Figure 4.4. Graphe de session à l'échelle de la page (SN2002, session 127666)

Lorsque le nombre de nœuds distincts est supérieur à une dizaine, la lisibilité des graphes se trouve quelque peu altérée, mais les représentations restent cependant exploitables. La Figure 4.5 ci-dessous montre un exemple de graphe au niveau site particulièrement dense, avec beaucoup de sites visités et de retours en arrière, certains sites agissant comme de véritables pivots dans la navigation. La représentation qui en découle est touffue, pas toujours lisible, mais on peut néanmoins dégager une thématique générale à la session : dans cet exemple de session datée du 21 avril 2002, on voit clairement que, outre les portails généralistes et les moteurs, les sites visités sont liés à la politique et aux élections présidentielles. L'effet de balayage des ressources est très fort, l'ensemble des tendances politiques étant représentées.

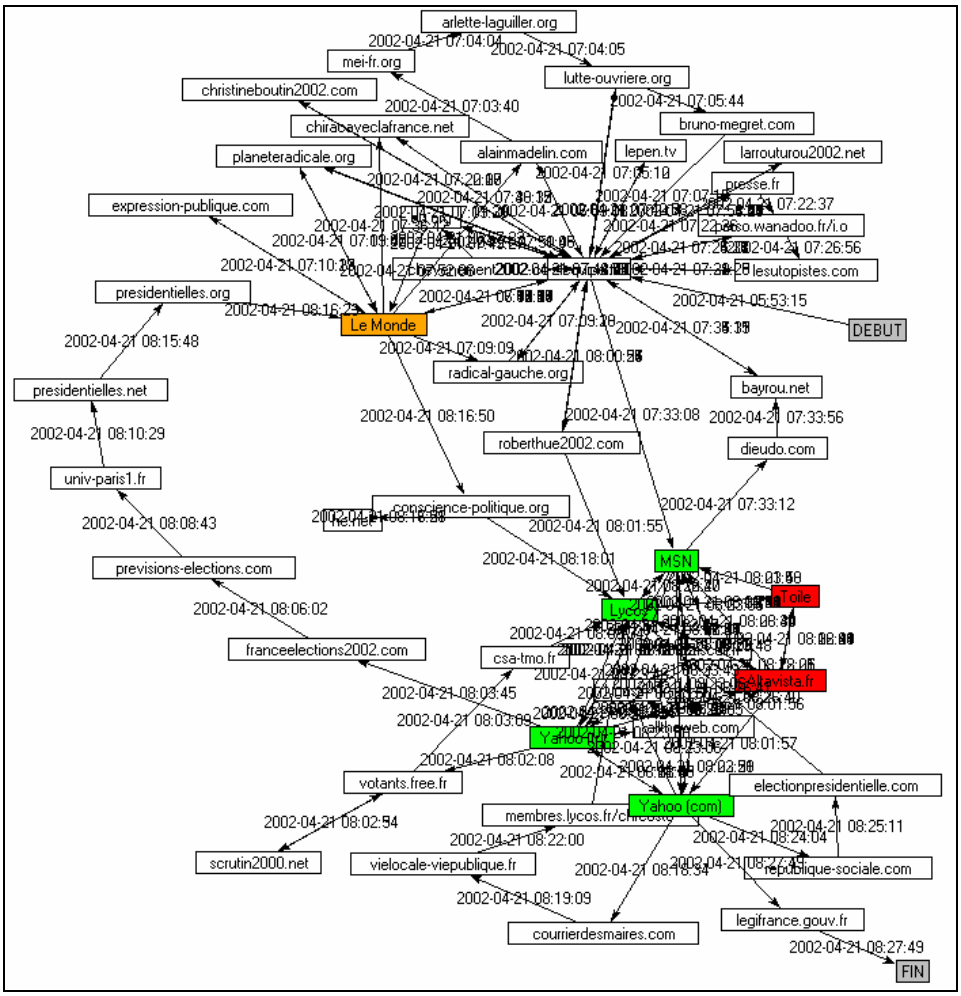


Figure 4.5. Graphe de session à l'échelle du site (SN2002, session 2461)

L'avantage de Graphlet est de proposer plusieurs algorithmes avancés de mise en forme ; le graphe produit peut être retouché finement, au niveau des labels, de la taille, du format des nœuds, etc. et le résultat peut être enregistré au format GML

(format Graphlet). Par contre, la production des graphes n'est pas automatisée : Graphlet ne peut pas être utilisé en ligne de commande, et il faut éditer chaque graphe non mis en forme dans l'interface graphique, appliquer un algorithme et le sauvegarder ensuite, ce qui est fastidieux lorsque l'on veut visionner un nombre important de graphes.

Solution Java

La deuxième solution à laquelle nous avons recouru pallie ce problème d'automatisation : le graphe est ici présenté dans une *applet* Java (voir Figure 4.6). Les paramètres relatifs aux différents nœuds et arcs du graphe sont passés à l'*applet* dans le fichier HTML qui l'appelle, ce qui permet ici aussi de faire une version *stand-alone* de l'outil, contenant les classes Java nécessaires et le fichier HTML relatif à la session représentée.

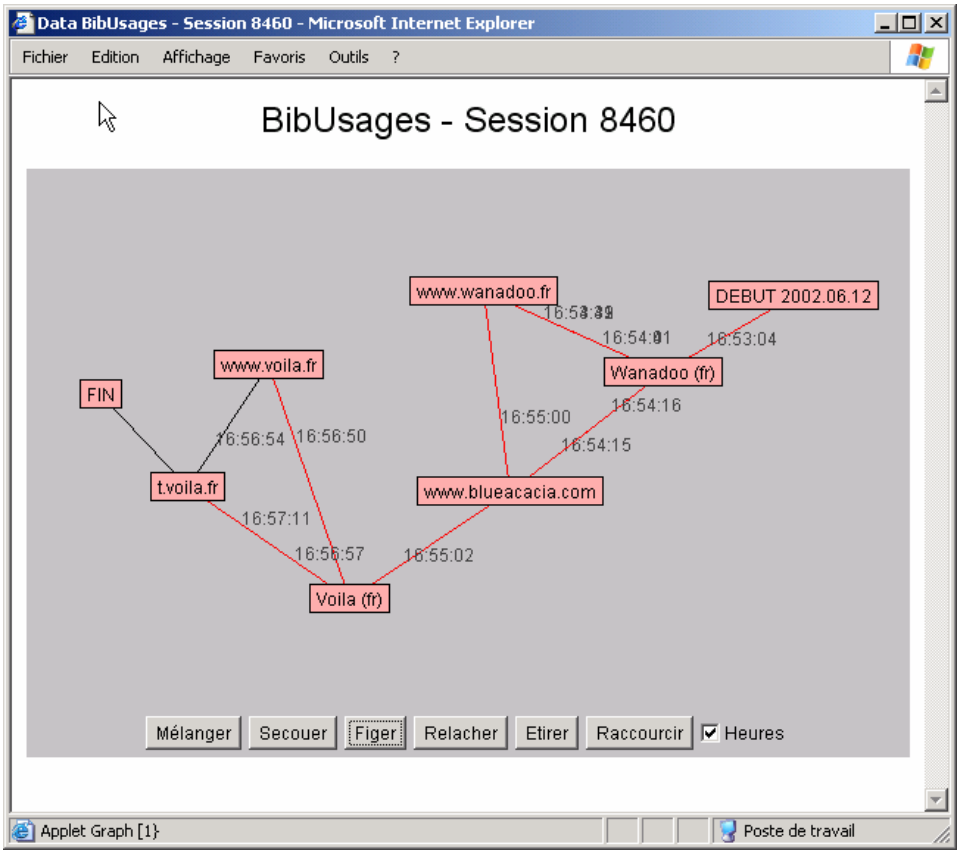


Figure 4.6. Graphe de session à l'échelle du site - Applet Java

L'inconvénient de cette solution est qu'elle ne propose pas un mécanisme de mise en forme élaboré : l'algorithme utilisé[1], assez rudimentaire, se base sur les distances entre nœuds, et tente d'étaler le graphe au mieux. Pour autant, le résultat est satisfaisant pour des graphes de faible dimension, d'autant plus que le graphe est facilement manipulable : l'utilisateur voit la mise en forme se faire dans l'*applet*, et peut intervenir pour déplacer un nœud, le figer, allonger ou raccourcir la distance entre les nœuds, et figer l'ensemble ou laisser se poursuivre la mise en forme. Cette souplesse compense en partie la faiblesse du procédé d'élaboration du graphe.

L'autre intérêt particulier de cette solution en Java est son interfaçage complet avec les données de trafic et les outils de fouille développés dans le projet SensNet. Là où il faut avec Graphlet générer un fichier GML, l'ouvrir puis le mettre en forme pour chaque session, ce qui est fastidieux, le graphe est produit dynamiquement dans une interface Web et sa production est totalement transparente pour l'utilisateur, qui fait l'économie des manipulations techniques.

Dans les deux cas, nous voyons l'agrément de cette représentation des parcours en même temps que ses limites : les graphes sont clairs lorsque les parcours sont assez courts et linéaires, mais rapidement illisibles lorsque les sessions s'allongent. En outre, nous rencontrons ici des problèmes inhérents aux données sur lesquelles nous travaillons, à savoir la différence entre ce que perçoit l'utilisateur (l'unité visuelle de la page) et les requêtes multiples qui peuvent en être à l'origine. Ainsi, comme nous l'avons vu pour *RePlay*, l'utilisation de *frames* par les concepteurs de sites génèrera au moins trois requêtes, et trois nœuds sur le graphe, alors qu'il s'agit, du point de vue de l'utilisateur, d'une seule page. Cela étant, l'outil s'est montré très utile et a globalement répondu à nos attentes.

À l'usage, cet ensemble d'applications de fouille minutieuse des données s'est avéré précieux, en ce qu'il rend plus palpables les logiques de navigation en les replaçant dans un pseudo-contexte d'utilisation. Il a également, et c'est son principal intérêt, permis de formuler certaines hypothèses, parmi lesquelles :

- le comportement des utilisateurs varie grandement en fonction des services accédés, en particulier en termes de longueur de session (temps et nombre d'URL).
- on tend à observer une forme d'opposition entre comportement « prédateur » (courte session, le panéliste sait où il va ou ce qu'il cherche avec précision) et « fureteur » (session plus longue, plus diversifiée, où le suivi des liens entre pages semble tenir d'un certain opportunisme).
- la session n'apparaît pas comme une unité cohérente du point de vue du contenu, et l'on observe dans un certain nombre de sessions des formes de « coq à l'âne » assez radicaux.
- la vision très « mono-tâche » de la navigation Web, que l'on retrouve dans un certain nombre d'études, est mise à mal par l'observation d'entrelacements au sein d'une même session de deux, voire plus, cours

[1] Cet algorithme est emprunté aux exemples d'*applets* proposés par Sun ; voir `http://java.sun.com/applets/jdk/1.1/demo/GraphLayout/`.

d'action distincts, qui peuvent correspondre à l'utilisation de plusieurs fenêtre du navigateur utilisées simultanément.

Toutes ces hypothèses sont loin d'être exhaustives, et méritent bien évidemment d'être vérifiées. Elles traduisent surtout la démarche hypothético-déductive qui est la nôtre, et qui implique la possibilité d'aller et venir entre des représentations synthétiques de masse et un examen minutieux des données de navigation. C'est dans ce cadre que l'élaboration d'indicateurs statistiques rendant compte ce que nous percevons lors de la fouille minutieuse des données s'avère précieuse : face aux données volumineuses de trafic dont nous disposons, l'examen manuel systématique est impossible, et nous devons nous doter d'outils synthétiques de représentation des contenus et des formes de parcours pour la vérification massive des observations manuelles.

> ***Synthèse.*** *La représentation des parcours sous forme de graphes permet, comme* RePlay, *d'émettre et de vérifier des hypothèses sur les parcours, et alimente une approche hypothético-déductive. Elle offre également une vue synthétique des sessions à différentes échelles (page, site, service) qui permet d'appréhender leur complexité : détours, retours arrière, pages-pivots, etc. Une telle vue est particulièrement utile pour l'analyse de la topologie des parcours.*

4.2 Analyser la séquentialité

Si les outils de fouille permettent d'approcher au plus près les données de navigation et de formuler des hypothèses de travail, l'analyse de données de trafic volumineuses nécessite la construction de représentations et de descriptions synthétiques des parcours. Ces descriptions permettent de vérifier les hypothèses sur les comportements de navigation à l'aide de traitements statistiques et formels sur les parcours, et ouvrent la voie vers l'élaboration de profils-types de sessions et d'outils de classification automatique des parcours.

4.2.1 Parcours Web : travaux existants

L'existence de nombreux systèmes hypermédia avant le Web a donné lieu à nombre de travaux sur la navigation, en particulier dans le champ des sciences cognitives. Les recherches portent alors principalement sur la modélisation de l'utilisateur à travers l'étude de ses parcours dans un système hypermédia donné ; les applications sont alors tournées vers les recommandations de conception et surtout la mise en place d'hypermédias adaptatifs (*adaptive hypermedia*). On trouvera dans [Brusilovsky 1996] un panorama très complet des problématiques, des méthodes et des applications relatives aux hypermédias adaptatifs avant l'émergence du Web, complété en 2001 dans [Brusilovsky 2001] pour les études centrées sur le Web. Les champs d'applications principaux sont centrés autour des sciences de l'éducation (application à des encyclopédies, des méthodes d'apprentissage multimédia) et à la recherche d'information, proche de l'ingénierie documentaire.

Dans ce cadre de recherche, les éléments qui nous intéressent ont trait au travail descriptif nécessaire à la modélisation des comportements d'utilisateurs : d'une part les méthodologies et indicateurs construits pour rendre compte des navigations dans les systèmes hypermédias, et d'autre part les différents profils-types qui ont pu être extraits des observations. Les modèles construits proprement dits ne retiennent pas notre attention : nous ne visons pas à modéliser les comportements pour faire de la prédiction, mais à disposer d'outils permettant de décrire et d'analyser les comportements observés dans le cadre de pratiques en situations réelles.

Browsing patterns et modélisation de l'utilisateur

Ceci étant posé, que retiendrons-nous des travaux orientés *user modeling* ? En premier lieu, au niveau minimal des actions de navigation, Canter, Rivers et Storr proposent dans [Canter *et al.* 1985] d'identifier quatre formes de bases dans la navigation (voir Figure 4.7 ci-dessous) :

- *pathiness* (chemin) : un chemin est un parcours qui ne passe pas deux fois par le même nœud ;
- *ringiness* (anneau) : un anneau est un parcours qui retourne à son point de départ ;
- *loopiness* (boucle) : une boucle est un anneau qui ne contient pas de sous-anneau ;
- *spikiness* (pointe) : une pointe est un anneau qui retourne à l'origine en repassant par les nœuds intermédiaires.

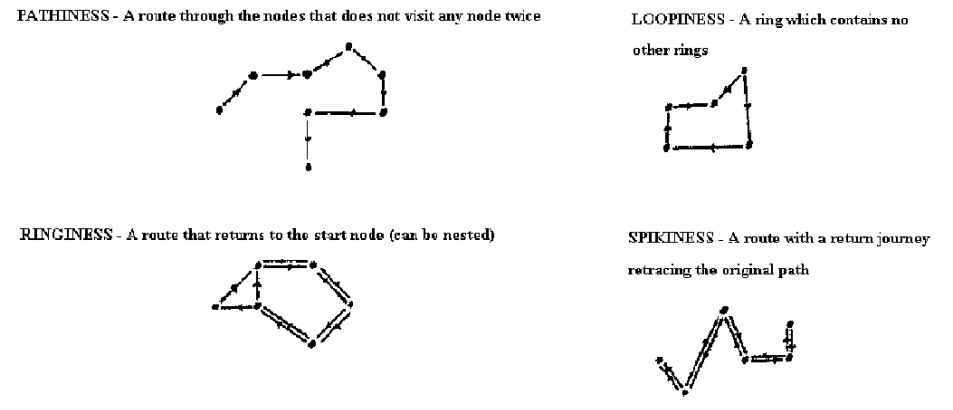


Figure 4.7. Formes de bases de navigation présentées dans [Canter et al. 1985]

Ces quatre formes sont dénombrées dans les sessions de navigation, à quoi s'ajoutent deux ratio : 1) le ratio du nombre d'éléments visités rapporté au nombre total d'éléments, et 2) le ratio du nombre d'éléments visités rapporté au nombre total de visites. Les auteurs obtiennent ainsi six indicateurs pour représenter de manière formelle la navigation, et les analysent pour dégager des modèles de navigation-type. Notons ici que la reconnaissances de ces quatre formes est loin d'être triviale, en particulier l'extraction des *boucles* et des *anneaux* dont les imbrications peuvent être complexes ; [Mullier 2000] propose ainsi une méthode de reconnaissance automatique de ces formes sur la base de réseaux de neurones.

Ce type de recherches sur les hypermédias a trouvé un prolongement naturel et productif sur l'hypertexte particulier que constitue le Web : on trouve une littérature relativement abondante traitant de l'analyse des parcours d'utilisateurs d'un point de vue *site-centric*, sur la base de l'analyse des *logs* des serveurs Web. S'appuyant sur les techniques mentionnées ci-dessus ou mobilisant des analyses statistiques sur la base de chaînes de Markov, d'analyse de séries temporelles ou d'outils de *data mining*, les travaux cherchent à découvrir des motifs récurrents de navigation (*browsing patterns*) sur un site donné. Les applications sont essentiellement orientées vers la conception (amélioration de l'architecture et de l'ergonomie d'un site), l'analyse de la fréquentation (rubriques les plus visitées, segmentation des sessions par type de visite), l'optimisation des serveurs et l'élaboration de contenus adaptatifs (prédiction, proposition de liens et contenus personnalisés à la session). Un tel engouement s'explique par les enjeux économiques sous-jacents à ces recherches : les sites à vocation commerciale souhaitent disposer de données les plus précises possibles sur leur fréquentation, afin de savoir quelles pages sont les plus visitées, comment les utilisateurs y arrivent, et comment les faire « rester » plus longtemps sur le site.

Toute une série de travaux a suivi cette voie en conservant les paradigmes issus des études sur les hypermédias ; ces recherches se situent dans le champ des sciences cognitives et s'orientent vers la modélisation de l'utilisateur en situation de navigation sur le Web. Nous renvoyons à la lecture de [Modjeska 1997] pour un panorama des travaux effectués dans ce domaine, certes un peu daté, mais qui rend bien compte des problématiques soulevées par cette approche, qui fait la part belle aux perceptions, aux « structures cognitives » et aux « modèles mentaux » de l'utilisateur.

Dans la plupart des cas, il s'agit d'études centrées-utilisateur sur des panels restreints, parfois tournées vers l'« usabilité » d'un site en particulier, le problème de la « désorientation » des utilisateurs[1] mais le plus souvent orientées vers la recherche d'information. Ce paradigme, directement hérité de l'ingénierie documentaire, domine encore la recherche sur la navigation Web, où les contenus sont assimilés à des documents contenant des « molécules informationnelles », avec en arrière-plan une vision orientée « exécution de tâche » et résolution de problème (problématique héritée de l'Intelligence Artificielle).

Nous ne passerons pas en revue l'ensemble de ces études, mais proposons d'en décrire une qui nous paraît particulièrement représentative des problèmes qui sont posés dans ce cadre. Il s'agit du travail de D. Mullier, D. Hobbs et D. Moore ([Mullier *et al.* 2002]) : sur la base des indicateurs de [Canter *et al.* 1985], les auteurs appliquent une méthode basée sur des réseaux de neurones pour reconnaître ces motifs dans des données de navigation dans un hypermédia ([Mullier 2000]), « et les interpréter (lorsque c'est possible) ». Ce système est appliqué à l'analyse de la navigation en situation de recherche d'information : on demande à onze étudiants d'effectuer une recherche sur un thème donné (l'astronomie) pour répondre à une question précise (quel est le plus gros satellite du système solaire), puis de naviguer à leur guise dans ce domaine dont ils n'ont pas tous la même connaissance. Les auteurs

---

[1] Problème du type « lost in hyperspace », abordé sous l'angle des modèles mentaux côté site et côté utilisateur ; voir en particulier [Xu *et al.* 2001] et [Danielson 2003].

ont notamment observé des motifs topologiques différents en fonction de l'expertise des volontaires, les experts présentant des navigations avec plus de boucles que les autres.

Les conclusions sont alléchantes, rejoignent nos problématiques et nous intéressent directement. On peut toutefois reprocher à cette étude de ne pas donner plus d'informations sur les contenus visités : quels sites sont vus, leur nombre, les utilisateurs connaissaient-ils les sites qu'ils ont visités (ce qui influence directement leur navigation dans ces sites), etc. Faute d'apporter des précisions sur ces éléments, l'étude se révèle peu convaincante.

Mais plus encore, c'est la question de l'interprétation des motifs observés qui pose problème à nos yeux : les travaux effectués dans ce champ amènent la plupart des auteurs à définir des taxinomies de « stratégies de navigation ». Il est ainsi souvent fait référence aux quatre classes définies dans [Canter *et al.* 1985] :

- *scanning* (feuilletage) : l'utilisateur passe en revue un nombre important de pages sur un thème donné sans passer beaucoup de temps sur chaque page, de manière superficielle ;
- *browsing* (navigation) : l'utilisateur suit un chemin jusqu'à parvenir à son but ;
- *searching* (recherche) : l'utilisateur cherche un document ou une information en particulier ;
- *exploring* (exploration) : l'utilisateur explore une zone ou un domaine particulier jusqu'à en épuiser les ressources ;
- *wandering* (errance) : l'utilisateur suit un parcours déstructuré et sans but précis.

Comme le note [Bidel *et al.* 2003], il n'y a pas de consensus général parmi les différentes recherches sur une typologie des stratégies de navigation, et chaque auteur est amené, en fonction du matériau d'expérimentation sur lequel il base son étude, à proposer des catégories différentes. Pour autant, la grande majorité des études mettent leurs participants dans la situation de chercher dans un site donné ou sur le Web pour répondre à une question précise : cette situation d'expérimentation, quasi-prototypique pour avoir été répétée chez les uns et les autres, réduit drastiquement la réalité de la navigation sur le Web. Elle enferme contenus proposés, modes d'accès et activité de navigation dans le paradigme de la recherche d'information : ce faisant, elle conclut à des équivalences entre motifs de navigation, tâche et motivation de l'utilisateur qui sont à nos yeux abusives et réductrices. Dès lors, nous laisserons volontiers de côté ces approches orientées modélisation pour notre analyse. À cela nous opposons une approche descriptive qui s'attache à replacer les modes de navigation dans le cadre de pratique avérées, et à prendre en compte la singularité des situations, des contenus et des individus.

Web Usage Mining et analyse de *logs*

Aux côtés des recherches orientées vers la modélisation des utilisateurs, le champ du *Web Usage Mining* a vu se développer un courant plutôt centré sur l'analyse de données de trafic proprement dites. Ces travaux sont massivement centrés-serveur :

exception faite des quatre études que nous avons déjà évoquées au Chapitre 1[1], l'ensemble de travaux d'analyse de *logs* de navigation porte presque systématiquement sur des traces recueillies au niveau des serveurs Web. Nous renvoyons à la lecture du compte-rendu du Workshop, *WEBKDD'99: Workshop on Web Usage Analysis and User Profiling* ([Masand & Spiliopoulou 2000]) et des panoramas proposé dans « Web Mining Research: a survey » de R. Kosala et H. Blockeel ([Kosala & Blockeel 2000]) et dans « Web Mining – Accomplishments & Future Directions » ([Srivastava *et al.* 2003]) pour une vue assez complète et relativement récente des recherches menées dans ce cadre.

De manière générale, les méthodes utilisées font fortement appel aux outils d'analyse statistiques et à la théorie des graphes ; on se reportera volontiers à [Roddick & Spiliopoulou 2002] pour un panorama des méthodes de fouille de données appliquées à l'analyse des données temporelles. On trouve dans ce champ un nombre important d'études ; nous ne les détaillerons pas dans leur ensemble, mais citerons trois travaux qui nous semblent particulièrement intéressants et représentatifs :

- *HPG (Hypertext Probabilistic Grammar)* : Borges, Levene (en particulier [Borges & Levene 1998], [Levene & Loizou 1999] et [Borges & Levene 2000]). Pour Borges et Levene, le but est de proposer des techniques permettant d'identifier des *web trails*, c'est-à-dire des séquences de liens suivis par l'utilisateur. Pour cela, le site Web étudié est modélisé comme une « grammaire régulière » (*regular grammar*) dont les états correspondent aux pages Web et la production de règles aux hyperliens. Les sessions de navigation sont incorporées dans ce modèle afin de construire une *Hypertext Probabilistic Grammar* (HPG) à laquelle on peut appliquer des techniques de *data mining*. Ces techniques sont appliquées à des *logs* côté serveur, et ont surtout pour but d'aider les webmestres à améliorer leurs sites. Le point crucial de leur recherche est l'établissement de règles reflétant des régularités dans la navigation. Pour cela, des heuristiques « à grain fin » sont développées pour trouver l'accord entre la justesse des règles et leur nombre. Dans ce travail, les chaînes de Markov sont utilisées pour l'analyse et la prédiction.
- *WebMiner* : Cooley, Mobasher, Srivastava (en particulier [Cooley *et al.* 1997], [Cooley *et al.* 1999a] et [Mobasher *et al.* 2000]). Les travaux de ces chercheurs sont centrés sur l'application des techniques de *data mining* aux usages du Web. Dans [Cooley *et al.* 1999a], ils présentent le système WebMiner qui inclut la préparation des données pour l'analyse et met en œuvre ces techniques afin de modéliser le parcours de l'utilisateur. Dans [Cooley *et al.* 1999b], R. Cooley propose le système *WebSIFT* qui utilise le contenu et la structure d'un site pour identifier les résultats potentiellement intéressants de Web Usage Mining. Dans [Mobasher *et al.* 2000], B. Mobasher poursuit ces travaux dans l'objectif d'une personnalisation des

[1] Il s'agit de [Catledge & Pitkow 1995], [Cunha *et al.* 1995], [Tauscher & Greenberg 1997a] et [Cockburn & McKenzie 2000].

sites, c'est-à-dire d'une adaptation des contenus renvoyés en fonction des chemins suivis sur un site.

- *WUM (Web Usage Miner)* : Spiliopoulou, Faulstich et Winkler. Dans [Spiliopoulou *et al.* 1999], ces trois chercheurs présentent un outil, Web Usage Miner, capable d'agréger sous forme d'arbre les différents chemins suivis au sein d'un site, de la page d'entrée à la dernière page visitée. L'ensemble des données est stocké dans un format qui permet de les interroger *via* un langage proche du SQL, MINT. Il est ainsi possible de calculer, sous forme de requête, la probabilité qu'un utilisateur voie telle page, à partir d'un parcours ayant traversé telle ou telle page, à telle ou telle position, l'éventail des combinaisons se révélant illimité. L'outil est également pourvu d'une interface graphique permettant de visualiser les parcours et les résultats de requêtes sous forme d'arbres.

Ces approches sont intéressantes en ce qu'elles traitent réellement l'aspect séquentiel et parfois temporel des parcours (pondération par la durée passée sur la page). Toutefois, les outils et méthodes centrés-serveur ne peuvent pas être directement transposés à l'analyse de données centrées-utilisateur, pour deux raisons principales. La première tient à la redondance nécessaire dans les données : dans des *logs* de serveurs, les différentes URL sont vues un nombre assez important de fois pour observer des régularités, tandis que seule une minorité de pages et de sites sont vus dans plus d'une session pour un utilisateur donné. Le second problème tient à ce que l'approche centrée-serveur part du principe que le contenu des pages naviguées est connu, et appliquent des méthodes d'analyse statistique sur des séries de symboles représentant les pages. L'interprétation des motifs de navigation devient dès lors, en dehors de toute qualification de contenu, quasi-impossible sitôt que l'on passe du côté de l'utilisateur.

Certains travaux s'efforcent cependant d'inclure cette dimension dans l'analyse : ainsi, Acharyya et Ghosh ajoutent à l'analyse statistique « classique » des *logs* une information de « concept » rattachée à chaque page ([Acharyya & Ghosh 2003]). L'objectif est ici de prendre en compte un « changement de centre d'intérêt » de l'utilisateur au cours de la session, et de pré-segmenter les sessions sur la base des contenus visités ainsi que de mieux prédire les liens qui seront suivis en fonction de la position dans l' « arbre de concepts » (*concept tree*). Les auteurs notent une augmentation significative du taux de prédiction en ayant recours à cette méthode.

Dans la même optique, Heer et Chi présentent dans [Heer & Chi 2002] une approche de la navigation *site-centric* incluant données d'usage (*logs* du serveur), de contenu (contenu textuel des pages) et de topologie (structure de liens entre les pages) des sites. L'analyse combinée de ces trois sources de données est appliquée dans un premier temps aux données recueillies auprès d'un échantillon de 21 volontaires à qui les auteurs ont demandé d'effectuer une liste de tâches sur le site Web de Xerox ; dans un second temps, Heer et Chi utilisent les *logs* entiers du serveur lui-même sur une journée. La première étape permet de régler les pondérations correctes pour chacune des trois modalités descriptives ; la seconde conduit à la classification des sessions sur la base du contenu (poids : 0,75) et des liens (poids : 0,25) des pages. Neuf classes sont générées, qui représentent les thèmes et associations de pages typiquement visités sur le site de Xerox : achat en ligne,

support technique, catalogue des produits, etc. Il est intéressant de noter que la classe la plus importante (42% des sessions) est relative à la page d'accueil, ce que les auteurs interprètent comme le reflet d'une navigation repassant fréquemment par cette page.

Ces deux études nous intéressent directement : elles ouvrent le chemin d'un croisement entre contenus visités et formes de parcours, même si nous ne souscrivons pas à la méthode des *concept trees* employée dans [Acharyya & Ghosh 2003]. Pour autant, elles restent tributaires de l'approche côté serveur : comme l'a montré [Padmanabhan *et al.* 2001] en travaillant au niveau intermédiaire d'un fournisseur de contenus disposant de *logs* relatifs à plusieurs sites commerciaux, le point de vue server-centric est partiel et biaisé, et les conclusions que l'on peut tirer de ce type d'approches sont toujours à considérer avec prudence, en particulier lorsqu'elles tendent à dresser des utilisations-types et des comportements de navigation généraux. Dans le champ du commerce électronique, il a ainsi été montré dans [Licoppe *et al.* 2002], basé sur des données de trafic centrées-utilisateur et sur des entretiens avec des internautes, que les consommateurs en ligne ont un comportement très volatil, oscillant entre achat réfléchi et achat d'impulsion, et que l'achat en ligne implique la mobilisation de ressources hors des sites de e-commerce (moteurs, comparateurs, etc.) et hors Web. Ce retour de la sociologie des usages sur l'étude de la navigation et des usages du Web montre, s'il était nécessaire, la nécessité de se placer résolument du côté de l'utilisateur et de la complexité des pratiques dans et hors Web pour appréhender les comportements de navigation.

> ***Synthèse.*** *Les travaux menés dans le champ des sciences cognitives sur la navigation dans des hypertextes ont mis à jour des motifs élémentaires de navigation qui nous intéressent directement, même si les conclusions en termes de comportements des utilisateurs ne nous satisfont pas complètement. Les recherches centrées-serveur proposent quant à elles des méthodes intéressantes d'analyse de séquences de navigation, mais leur transposition dans une approche centrée-utilisateur reste problématique.*

4.2.2 Indicateurs topologiques

À la plupart des travaux existants sur la navigation Web, nous opposons un double décalage : en premier lieu, nous adoptons une approche centrée-utilisateur qui nous amène à considérer l'ensemble des parcours sur le Web effectués par des utilisateurs identifiés. D'autre part, notre approche vise la description et non la modélisation, et se situe résolument dans le cadre des sciences humaines. Nous ne rejetons pas les méthodologies statistiques élaborées qui ont pu être développées jusqu'alors dans d'autres travaux, mais elles sortent du champ de notre travail. Pour traiter la complexité et la spécificité de nos données de trafic, ainsi que la diversité des contenus et des comportements observés, nous proposons en contrepartie des indicateurs simples de la topologie et du rythme des parcours qui, combinés aux descriptions des contenus visités, permettent de rendre compte de manière compréhensive de l'activité de navigation.

Échelle d'analyse et descripteurs

Les descriptions des sessions que l'on peut construire s'appuient sur les éléments minimaux qui les composent : requêtes HTTP, composant des pages, regroupées au sein de sites. Il importe à ce niveau d'analyse de voir ce que l'on retient de ces descriptions élémentaires et de la manière de les combiner.

Comme nous l'avons déjà souligné, l'URL ne correspond pas systématiquement à la page en tant qu'unité ergonomique. Pour les analyses centrées-serveur, ceci ne pose pas de problème insurmontable : il est possible de corriger localement les *logs* des serveurs pour tenir compte des différents systèmes de publications et modes d'organisation de chaque site. Pour l'analyse centrée-utilisateur, il en va tout autrement, car nous ne savons pas, pour chaque site, comment celui-ci est organisé et si la correspondance entre page et requête (donc URL) est juste ; l'exemple de session donné dans le Tableau 4.1 illustre ce phénomène.

Tableau 4.1. Session à plat au niveau des URL (SN2002 - session 435)

Site	Date	URL
Wanadoo (fr)	16:34:53	http://www.wanadoo.fr/bin/frame.cgi
	16:34:55	http://www.wanadoo.fr/personnalisation/bin/webauth_aff.cgi
	16:34:55	http://www.wanadoo.fr/common/abonnes/menu_accueil.html
monster.fr	16:36:10	http://www.monster.fr/
	16:36:51	http://offres.monster.fr/
	16:36:53	http://offres.monster.fr/
	16:39:42	http://offres.monster.fr/getjob.asp?JobID=14153005&col=&cy=&brd=&lid=&fn=&q=&AVSDM
Google	16:40:32	http://www.google.fr/
	16:40:43	http://www..google.fr/search?q=Hachette+Livre+%28Edition%29+&hl=fr&btnG=Recherche+G(…)
	16:41:42	http://www.google.fr/search?q=Hachette+Livre+(Edition)+&hl=fr&cr=countryFR&start=10&sa=N
hachette.com	16:42:51	http://www.hachette.com/HomePageFO/francais/site/index.htm
	16:42:52	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:42:53	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:42:53	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:42:53	http://www.hachette.com/HomePageFO/servlet/CtlHome?URL=site/myi-home.jsp
	16:42:54	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:42:54	http://www.hachette.com/HomePageFO/servlet/CtlHome?URL=site/myi-home.jsp
	16:42:54	http://www.hachette.com/HomePageFO/francais/site/blanc.htm
	16:44:26	http://www.hachette.com/HomePageFO/francais/site/page/Frameset_Groupe.jsp?page=carrieres
	16:44:29	http://www.hachette.com/HomePageFO/francais/site/page/NavHaut.htm
	16:44:30	http://www.hachette.com/HomePageFO/francais/site/page/NavGauche_Groupe.jsp
	16:44:31	http://www.hachette.com/HomePageFO/francais/site/page/Frame_Carrieres.jsp?rub=metier1
	16:44:32	http://www.hachette.com/HomePageFO/francais/site/page/NavHautInter_Carrieres.jsp?Nrubrique=1
	16:44:33	http://www.hachette.com/HomePageFO/francais/site/page/NavHaut.htm
	16:44:35	http://www.hachette.com/HomePageFO/francais/site/CAR/CAR01_ACCUEI_F.htm
	16:45:06	http://www.hachette.com/HomePageFO/francais/site/CAR/CAR06_COMMER_F.htm
	16:45:18	http://www.hachette.com/HomePageFO/servlet/CtlOffres?ACTION=0
	16:45:28	http://www.hachette.com/HomePageFO/francais/site/OFF/OFF04_STAGES_F.jsp
	16:45:29	http://www.hachette.com/HomePageFO/servlet/CtlOffres?ACTION=0
	16:45:32	http://www.hachette.com/HomePageFO/servlet/CtlOffres?ACTION=0
	16:45:44	http://www.hachette.com/HomePageFO/francais/site/CAR/CAR01_ACCUEI_F.htm

Dans cette session envisagée à l'échelle de l'URL, on compte 31 requêtes passées par l'internaute pour 26 URL distinctes, certaines étant envoyées deux fois. Les trois requêtes introductives envoyées à Wanadoo correspondent à une page unique, la page d'accueil du portail, que l'internaute a sans doute conservée en page de démarrage de son navigateur. L'analyse au niveau de la page se trouve brouillée par des données orientées trafic et peu fiables à l'échelle de la page : impossible, dès lors, de comparer les visites de sites en nombre de pages vues, car chaque webmestre aura mis en place des systèmes de *frames* différents, ce qui rend les résultats non homogènes.

Au-delà de ce problème, on questionnera volontiers l'approche « nombre de pages », qui a été beaucoup employée dans la mesure d'audience à ses débuts : quand bien même les données de trafic nous permettraient de compter exactement les pages vues du point de vue de l'utilisateur, quel sens donner à un tel décompte ? Certes, avoir ou ne pas avoir visité une page, voilà déjà un indice de fréquentation indéniable ; mais si l'on veut aller plus loin et entrer dans une logique comparative d'analyse de la fréquentation des sites Web, le décompte des pages soulève plus de problèmes qu'il n'en résout. Les pages, nous avons eu l'occasion de le constater dans l'examen préliminaire du corpus constitué à partir des données BibUsages, diffèrent en termes de taille, de fonction, de types de contenus : de très longues et de toutes petites, des textes volumineux et de simples formulaires, etc. De ce point de vue, la visite de la page ne prend sens que dans la dynamique de la navigation, et uniquement dans la mesure où nous pouvons en décrire le contenu thématique et/ou fonctionnel. De ces éléments, nous tirons deux conclusions méthodologiques importantes.

En premier lieu, il est nécessaire d'adopter la bonne échelle d'analyse en fonction des informations dont on dispose. Nous avons déjà quelque peu abordé la question de l'échelle de représentation des parcours dans la présentation des outils de visualisation, et avions vu que des échelles différentes donnent des résultats visuels sensiblement variables. Cette question se pose de manière plus brutale encore lorsque l'on souhaite construire des représentations chiffrées des parcours : en particulier, la linéarité de la navigation diffère grandement, pour un parcours donné, selon que l'on considère les pages visitées, les sites accédés, les services utilisés.

Dans l'exemple donné au Tableau 4.1, nous avions observé que la session est non linéaire au niveau de la page ; toutefois, si l'on observe cette session au niveau du site, elle est strictement linéaire, l'internaute ne revenant pas sur un site déjà vu au cours de la session. La session est alors décomposable en quatre pas : 1) Wanadoo (fr), 2) monster.fr, 3) Google et 4) hachette.com. On notera à cet instant les bénéfices du préformatage et de l'enrichissement des données brutes : d'une part, les domaines `www.monster.fr` et `offres.monster.fr` sont regroupées en un seul et même site, 'monster.fr' ce qui est cohérent avec une logique centrée-utilisateur ; d'autre part, l'identification des portails généralistes avec *CatService* permet de repérer que les pages vues sur `www.wanadoo.fr` sont relatives au portail Wanadoo (fr), taggué comme 'Portail généraliste', et que celles sur `www.google.fr` se rapportent à Google, 'Moteur de recherche'.

On peut également exploiter plus finement les descriptions fournies par *CatService* (voir Tableau 4.2). On sait alors que les trois URL vues sur Wanadoo correspondent à la page d'accueil, tandis que sur Google, le mouvement est décomposable en 1) l'accès à la page d'accueil (une URL), et 2) une requête contenant les mots-clefs « Hachette Livre (Edition) » (deux URL).

Tableau 4.2. Session agrégée au niveau du site et des services (SN2002 - session 435)

Date	site/portail	Nb URL	Durée	Service	mots-clefs
16:34:53	Wanadoo (fr)	3	7"	Page Accueil	
16:36:10	monster.fr	4	4' 12"		
16:40:32	Google	1	11"	Page Accueil	
16:40:43	Google	2	2' 8"	moteur	Hachette Livre (Edition)
16:42:51	hachette.com	21	2' 54"		

En fonction des différentes sessions et des différentes échelles d'analyse, les résultats seront sensiblement différents en termes de linéarité. Devant le défaut de fiabilité des données au niveau de la page, on s'attachera plutôt par la suite à travailler au niveau du site (ou du portail), et du service lorsque celui-ci est identifié dans *CatService* ; ceci est par ailleurs cohérent avec la mobilisation des descriptions des annuaires, qui décrivent majoritairement les données de trafic à l'échelle du site.

Deuxième élément méthodologique, hors de toute qualification des contenus, on préfèrera s'attacher à la durée de visite plutôt qu'au nombre d'URL visitées. Dans l'exemple de session ci-dessus, on constate ainsi que si l'internaute a demandé quatre URL sur `monster.fr` contre 21 sur `hachette.com`, il a passé plus de quatre minutes sur le premier site, contre moins de trois minutes sur le second. Bien évidemment, rien ne nous dit qu'il ne s'agit pas là d'un effet de feuilletage plus rapide des contenus proposés par `hachette.com` par rapport à ceux de `monster.fr`. Quoi qu'il en soit, le problème est insoluble dès lors que l'on ne dispose pas d'information précise sur les contenus visités.

Dans cette perspective, si l'on agrège les données de navigation à l'échelle des sites, la durée apparaît comme une donnée préférable au nombre de pages. Certes, cet indicateur n'est pas absolu : il faut en particulier se garder de conclure à une équivalence entre durée et importance dans la navigation ou intérêt de l'utilisateur. Certains sites comme les moteurs de recherche peuvent fonctionner comme des lieux de passage où l'utilisateur ne va pas rester longtemps : il ne faut pas en conclure pour autant que leur présence est négligeable dans le parcours. Pour autant, la durée est, dans les données dont nous disposons, le moins mauvais indicateur ; c'est au moment de l'interprétation des résultats qu'il importera d'y prêter une attention particulière.

Construction d'indicateurs topologiques simples

La dimension temporelle est un des éléments fondamentaux de la sémantique des parcours. Elle se situe à deux niveaux : d'un côté, il s'agit de prendre en compte les durées de visites et le temps passé sur chaque page ou chaque site, ce que nous venons de voir. De l'autre, il importe d'examiner l'ordre dans lequel les contenus sont accédés et la valeur qu'ils prennent dans la dynamique du parcours.

Pour cela, nous avons tenté de nous munir d'outils de fouille permettant de vérifier ces hypothèses, ainsi que de mettre en place des outils statistiques à même de rendre compte de la dynamique des parcours. L'approche que nous mettons en place est simple, et repose sur la construction d'indicateurs permettant de représenter les « formes » et le rythme des parcours de manière synthétique. Les indicateurs doivent permettre de représenter certains aspects particuliers de la session :

- s'il est linéaire ou non ;
- le nombre et la longueur des détours ;
- l'importance de ces détours dans la temporalité de la session ;
- distinguer et quantifier les points de fixation de la session (centres de formes en rosace).

Le parti pris de la simplicité et de la robustesse qui est le nôtre nous amène à n'envisager que partiellement la dimension séquentielle : seul un outillage complexe permettrait de tenir ensemble, dans un même objet statistique, les éléments de forme et de contenus des parcours Web[1]. Dans les indicateurs que nous proposons, la séquentialité est analysée en dehors des éléments de contenu : nous formalisons les sessions Web comme une séquence de symboles représentant les éléments visités. À partir de cette représentation, nous construisons les indicateurs suivants :

- N : longueur de la session (nombre de pas) ;
- n : nombre d'éléments uniques vus dans la session ;
- $r = \frac{n}{N}$: taux moyen de linéarité du parcours, qui vaut 1 s'il est linéaire et se rapproche de 0 au fur et à mesure que cette linéarité diminue ;
- R : nombre d'éléments revisités, c'est-à-dire vus plus d'une fois ;
- $c = \frac{N-n}{R}$: nombre moyen de revisites par élément revisité. Cet indicateur représente la concentration des revisites sur un ou plusieurs éléments du parcours, et permet de détecter des navigations « en étoile » : dans l'exemple de la Figure 4.8, pour les deux navigations N=12, n=9, et r=1,3, mais c est différent (3 vs. 1).

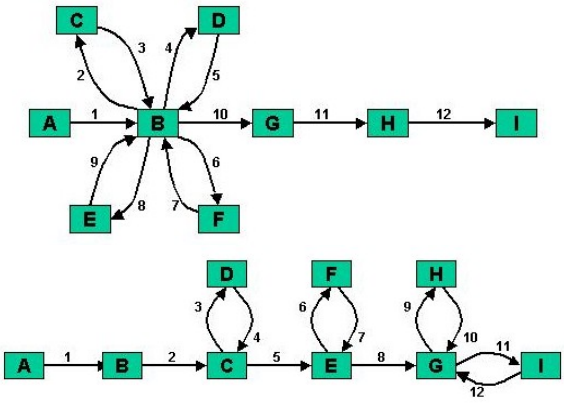


Figure 4.8. Concentration des revisites

[1] Nous pensons en particulier aux graphes colorés, qui permettent d'attacher des informations aux nœuds et aux arcs du graphe, tout en tenant compte du caractère orienté et temporel de l'objet.

Nous construisons également des indices qui prennent en compte les durées passées sur chaque élément visité, tout en tenant compte des séquences de navigation :

- *T* : durée totale de la session ;
- durées moyenne et médiane passées sur chaque pas de la session ;
- *T1* : le temps passé sur les éléments de la session vus une seule fois ;
- $d = \frac{T1}{T}$: part du temps passé sur des éléments vus une fois dans l'ensemble de la session. Cet indicateur est proche du taux de linéarité *r*, mais s'applique aux durées : il vaut 1 si la session est linéaire, et 0 si elle ne l'est pas du tout.

Nous avons également souhaité avoir des informations qualitatives sur la façon dont les pages sont revisitées. En particulier, nous avons voulu mesurer l'emploi de la fonction *Back* des navigateurs (retour d'une page en arrière). Pour cela, nous avons développé un algorithme spécifique capable d'identifier les séquences de *back* et de les isoler du reste de la session. Pour chaque session, nous en générons une nouvelle représentation qui correspond au parcours sans les mouvements de *back*. Par exemple, une session "A→B→C→D→E→D→C→F" sera transformée en "A→B→C→F", la séquence "C→D→E→D→C" étant réduite à "C". La Figure 4.9 illustre ce travail de réécriture des sessions, et montre en particulier la différence entre les séquences de type *back* et les boucles, non exclues dans ce traitement.

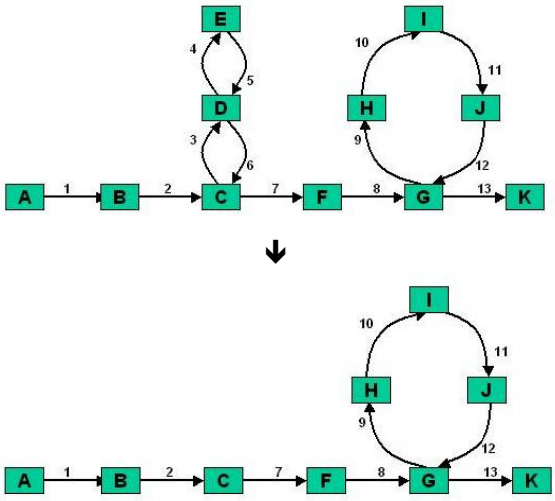


Figure 4.9. Réécriture d'une session sans les mouvements de Back

Ainsi, nous produisons une nouvelle série d'indicateurs relatifs à l'utilisation du *Back* et aux sessions dont les *back* ont été ôtés :

- *B* : nombre de séquences de type *Back*, quelle que soit leur longueur ;
- N_b : la longueur du parcours (nombre de pas) une fois les séquences de *back* ôtées ;

- $b = \frac{N_b}{N}$: part des actions de type *Back* dans le nombre total de pas dans la session. Plus l'indice est proche de 0, plus les actions *Back* occupent de place dans la session.

La quantification des actions de type *Back* est intéressante à double titre. À l'échelle de la page, elle correspond à l'utilisation d'une fonctionnalité des navigateurs, et rend compte d'un mode d'utilisation des interfaces et du déroulement du parcours. Au niveau du site, la correspondance avec la fonctionnalité des navigateurs n'opère que si une seule page est vue sur chaque site de la séquence d'aller-retour, et ne renvoie donc pas tant à une fonctionnalité de l'IHM qu'à renforcer l'identification de sites-pivots au sein d'une navigation en étoile.

Comme nous l'avons discuté précédemment, la session peut être abordée à deux niveaux de complexité : page / site, et nombre d'éléments / durée. En conséquence, chacun des indicateurs décrits ci-dessus est dédoublé, selon l'échelle retenue. La liste d'indicateurs finale obtenue est donnée au Tableau 4.3 ci-dessous.

Tableau 4.3. Liste des indicateurs topologiques et temporels retenus

	Indicateur	Description
	D	Durée de la session (en secondes)
Échelle : page	P	Nombre de pages visités (nombre de pas)
	p	Nombre de pages distinctes visitées
	R_{page}	Nombre de pages distinctes vues plus d'une fois
	r_{page}	Taux de linéarité – échelle page
	c_{page}	Taux de concentration – échelle page
	B_{page}	Nombre d'action de type *Back* – échelle page
	b_{page}	Part des actions de type *Back* dans la session – échelle page
	t_{page}-*moy*	Durée moyenne sur chaque page vue
	t_{page}-*med*	Durée médiane sur chaque page vue
	$D1_{page}$	Durée totale sur les pages vues une seule fois
	d_{page}	Part du temps passé sur les pages vues une seule fois
Échelle : site	S	Nombre de sites visités (nombre de pas)
	s	Nombre de sites différents visités
	R_{site}	Nombre de sites distincts vus plus d'une fois
	r_{site}	Taux de linéarité – échelle site
	c_{site}	Taux de concentration – échelle site
	B_{site}	Nombre d'action de type *Back* – échelle site
	b_{site}	Part des actions de type *Back* dans la session – échelle site
	t_{site}-*moy*	Durée moyenne sur chaque site vu
	t_{site}-*med*	Durée médiane sur chaque site vu
	$D1_{site}$	Durée totale sur les sites vus une seule fois
	d_{site}	Part du temps passé sur les sites vus une seule fois

Si ces indicateurs simples ne rendent pas compte de la complexité des formes de sessions dans son ensemble, ils en donnent un bon aperçu. Ils permettent d'établir des premières segmentations élémentaires des sessions sur la base de leur topologie ; couplés aux outils d'examen manuel des parcours et aux descripteurs de contenu, ils

doivent permettre de croiser forme et contenu de parcours et de parvenir à une segmentation des activités de navigation.

> ***Synthèse.*** *Nous avons élaboré des indicateurs statistiques simples pour représenter la topologie et la temporalité des parcours, à l'échelle de la page comme du site : linéarité, concentration des revisites, temps passé sur les pages et les sites dans et hors des retours, utilisation de la fonction* back *sont représentés de manière synthétique. Ces indicateurs alimentent une analyse praxéologique de la navigation.*

4.3 Contextualisation

La description de la sémantique des parcours que nous élaborons place la session au cœur de l'analyse, et en fait son objet privilégié. Gardons toutefois à l'esprit que dans l'activité de navigation comme ailleurs, le global définit le local : dans cette perspective, on s'efforcera autant que possible de replacer les parcours dans le double contexte de l'offre de contenu et du profil de l'utilisateur.

4.3.1 Contexte global du Web

Au niveau macro-analytique, avant d'avancer dans l'analyse des parcours sur le Web de page en page et de site en site, il semble important d'avoir une connaissance de l'arrière-plan structurel du Web. Un certain nombre de travaux ont été menés sur la structure du Web, modélisant celui-ci comme un graphe dont les noeuds sont les pages et les arcs les liens d'une page vers l'autre, mais on a vu peu d'études systématiques avec des robots parcourant l'ensemble du Web. Parmi ces dernières, une certaine controverse a semblé exister entre les résultats des différentes équipes ; le travail de Broder *et alii*, présenté en 2000 ([Broder *et al.* 2000]) s'impose dans les débats et apparaît comme le plus exhaustif et le plus fiable de tous.

Le résultat de ces analyses, dont nous reproduisons la représentation graphique (Figure 4.10 ci-dessous), montre une forme en « nœud papillon » :

- au centre, un réseau très fortement interconnecté, qu'il est possible de parcourir facilement ;
- à gauche, des sites qui pointent vers la nébuleuse centrale mais qu'il est difficile de rejoindre car peu de liens permettent de s'y rendre, typiquement des sites personnels à faible notoriété ;
- à droite, au contraire, une série de sites désignés par les pages du groupe central mais dont il est difficile de sortir car ils renvoient peu ou pas vers d'autres sites : ce groupe est essentiellement composé de sites commerciaux, qui contiennent peu de liens externes ;
- quelques passages directs de la partie gauche à la partie droite qui « évitent » la partie fortement interconnectée du Web.
- enfin, de petits composants fortement interconnectés mais sans liens avec le reste du Web.

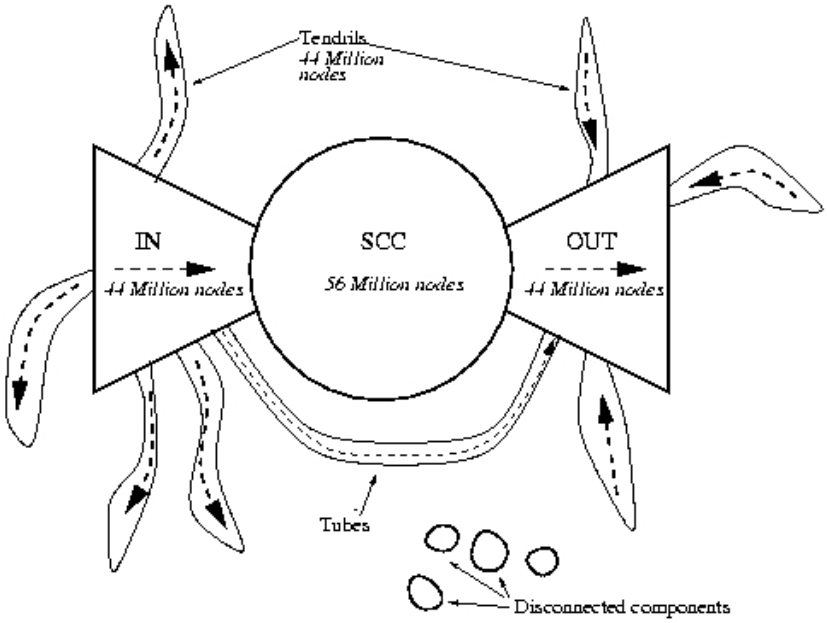


Figure 4.10. Structure du Web représentée dans [Broder et al. 2000]

Ce travail se trouve prolongé dans [Baeza-Yates & Poblete 2003], qui offre un éclairage longitudinal sur l'évolution du Web chilien entre 2000 et 2002. S'appuyant sur les travaux de Broder, l'auteur montre que la croissance constante de la Toile chilienne s'accompagne d'une complexification des sites et d'un accroissement du « noyau central » fortement interconnecté.

Dans le même champ d'analyse, on remarquera également le travail présenté dans [Faloutsos *et al.* 1999], qui souligne la pertinence des lois statistiques de type loi de puissance, « power-law » (i.e. de Pareto) pour l'analyse de la topologie du Web ; [Adamic 1999], [Adamic & Huberman 2001] et [Adamic 2001], où Adamic observe que le « Small World Web » est un monde de taille restreinte, au sens ou les pages et les sites sont regroupés en petits réseaux fortement interconnectés et de nature communautaire ; et enfin [Maurer *et al.* 2000] qui analyse la structure particulière que constituent les *Web rings*.

Ces observations sont particulièrement intéressantes : dans une navigation de site en site, l'impossibilité de sortir d'un site ou au contraire l'impossibilité d'aller sur un site car aucun lien n'y mène sont des facteurs importants. Il serait par exemple très utile de savoir quelle est la position de chaque site visité par un internaute sur le graphe de Broder. Si cette ambition dépasse le cadre de notre travail, notons ici l'intérêt d'un tel projet.

> ***Synthèse.*** *L'analyse macroscopique du Web sous l'angle des liens hypertextes, des flux de navigation et des cliques de sites peut apporter un arrière-plan pertinent pour l'analyse des parcours. Cette piste, complexe à mettre en œuvre, ne sera pas exploitée ici, mais on gardera à l'esprit ces éléments macro-structurels sous-jacents à la navigation.*

4.3.2 Contexte de l'utilisateur

Pratiques Web, pratiques Internet

Un premier élément de contextualisation de la navigation au niveau de l'individu consiste à replacer chacune de ses sessions dans le corpus global de ses parcours sur le Web. Ce type d'étude, qui nécessite une durée d'observation suffisamment longue pour être probante, est aussi rare que le sont les travaux centrés-utilisateur : dans [Catledge & Pitkow 1995] (trois semaines d'observation, 107 utilisateurs) ou [Crovella & Bestavros 1996] (cinq mois et demi, pour 37 postes équipés du dispositif de recueil de trafic), les sessions sont considérées en masse et ne sont pas rapportées à tel ou tel utilisateur.

Par contre, dans [Tauscher & Greenberg 1997a] et [Tauscher & Greenberg 1997b] (23 utilisateurs durant six semaines), les auteurs se penchent spécifiquement sur la « revisite » de pages Web, et s'attachent à examiner pour chaque internaute la fréquence de visite des pages accédées ; l'analyse montre qu'en moyenne 58 % des requêtes pour un utilisateur donné pointent vers une page qu'il a déjà visitée, en même temps que le « vocabulaire » des URL ne cesse de croître avec le temps. L'étude présentée dans [Cockburn & McKenzie 2000] (70 utilisateurs observés pendant quatre mois) va dans le même sens : pour chaque internaute observé, Cockburn et McKenzie constatent une croissance assez régulière du vocabulaire (les pages) dans le temps, une forte corrélation entre le nombre de visites et le vocabulaire, ainsi qu'une distribution de type zipfienne de la visite des pages (peu de pages sont visitées très régulièrement, beaucoup le sont rarement).

Ces premiers constats statistiques conduisent d'ores et déjà à considérer que la navigation amène sans cesse à voir de nouveaux sites, mais à n'en revoir que très peu : dans le corpus de pages et de sites de chaque internaute, on peut distinguer de manière nette les sites vus une fois seulement et ceux soumis à des visites routinières. Routinier ne signifie pas forcément fréquent : un internaute peut être amené à consulter systématiquement le ou les mêmes sites dans un contexte donné, tout en ne se trouvant que rarement dans cette situation précise : préparer ses vacances en allant sur des sites de voyagistes, réserver un billet de train, connaître un itinéraire routier, chercher un emploi sont autant d'activités qui, sans être fréquentes, peuvent être contextuellement régulières.

Nous souhaitons approfondir ce point : derrière la notion de corpus de sessions et de sites pour chaque individu, se profile la question de la construction des territoires personnels sur le Web. Découverte de sites, évolution des modes de navigation, modes d'appréhension des sites en fonction de leur place dans les routines de l'utilisateur, apprentissage ou perfectionnement dans le maniement des outils Web sont autant de questions qui renvoient à une éthologie de la navigation et un examen des modalités de la pratique en contexte. En outre, l'offre étant en perpétuel renouvellement, tant au niveau des sites que des types de contenus et de services proposés (« webisation » des outils de communication, outils de publication simplifiés, élargissement de l'offre de contenus culturels, commerciaux, etc.), l'internaute est soumis à la nécessité d'adapter ses comportements à cette évolution et se trouve potentiellement en perpétuelle situation d'apprentissage et de découverte. Nos données de trafic nous invitent à ce type d'approche : centrées-

utilisateur et exceptionnelles par leur taille et leur durée (plusieurs milliers d'internautes suivis durant plusieurs mois ou plusieurs années), elles ouvrent légitimement la voie d'une approche longitudinale approfondie.

Un autre élément de contextualisation individuelle des pratiques tient à l'examen de l'activité Internet dans son ensemble. Il s'agit là d'un élément fondamental dans l'approche retenue pour les projets TypWeb, SensNet ou BibUsages : la navigation sur la Toile s'insère et s'entrelace avec les autres outils Internet, comme la messagerie électronique, le *chat*, les jeux en ligne, le téléchargement, etc.

Dans le cadre du projet TypWeb, V. Beaudouin a établi une segmentation d'une cohorte d'internautes résidentiels fondée sur l'utilisation du Web et des outils de communication :

> Deux grands groupes d'internautes se distinguent : ceux qui accordent une place prépondérante au Web dans leurs usages d'Internet et ceux qui favorisent au contraire l'usage des services de communication. Dans chacun de ces groupes se constituent des axes de différenciation, en fonction de l'intensité d'usage pour le premier groupe et en fonction du ou des outils de communication utilisés (mail classique, WebMail, *chat*, messagerie instantanée…) pour le second groupe. Les utilisateurs de *chat* et messageries instantanées se recrutent surtout chez les jeunes, et se distinguent par leur capacité à articuler au cours d'une même session consultation du Web, utilisation de la messagerie et conversations synchrones.[1]

Dans une étude sur l'entrelacement des médias dans la constitution des publics de l'émission Loft Story (voir [Beaudouin *et al.* 2003a]), nous avions également constaté la synchronisation de l'activité Internet avec le flux télévisuel et la création de communautés sur support électronique extrêmement actives mobilisant le Web comme support de publication, les salons de *chat* comme espace de débat collectif et d'échanges inter-individuels, et la messagerie comme outil d'échange d'images ou d'informations.

Ces éléments rappellent que l'activité de navigation s'inscrit dans un double entour : celui des contenus et des services Internet en général, avec lesquels elle se trouve étroitement entrelacée, et l'entour plus global des pratiques hors Internet. Si ce deuxième point nous échappe, nous prendrons en compte autant que faire se peut l'usage global d'Internet dans la description des utilisateurs, afin de voir dans quelle mesure cette dimension générale permet d'expliquer des comportements de navigation sur le Web. Gardons à l'esprit que la navigation sur le Web s'apparente à une activité « comme une autre » et qu'elle s'inscrit dans l'univers des pratiques de l'individu, ce qui justifie encore, s'il était besoin, une approche praxéologique et contextualisée de la navigation.

---

[1] [Beaudouin *et al.* 2002], p. 6.

Éléments socio-démographiques

Les études menées dans le champ de la sociologie des usages nous rappellent que les pratiques relatives à Internet ne sont pas neutres socialement[1]. En dehors du constat d'une fracture numérique, qui touche en particulier la capacité à s'équiper en terminaux pour les ménages à faibles revenus, [Lelong 2003] fait remarquer qu'« il reste à en préciser les multiples dimensions qui ne se limitent pas aux inégalités d'accès aux nouvelles technologies ». L'auteur précise que « l'analyse de leurs usages permet notamment d'évaluer l'importance de l'âge, du sexe, du milieu social dans cette appropriation ou ce rejet des nouveaux outils que sont principalement l'ordinateur et Internet. ». L'utilisation et l'appropriation de l'outil informatique et, corrélativement, des outils Internet sont encore aujourd'hui très sexuées, même si les femmes investissent aujourd'hui ce terrain longtemps demeuré masculin. Comme le remarque Jouët dans [Jouët 2003] à propos des TIC en général, « les catégories binaires – technologie/homme, relation/femmes – sont plus complexes qu'il n'y paraît. On observe ainsi une inversion des qualités attribuées à chaque sexe : les femmes traditionnellement associées à la subjectivité et à l'émotion, font preuve d'une grande rationalité dans leurs usages, alors que les hommes, traditionnellement rangés du côté de l'objectivité et de la rationalité, donnent libre cours à leur émotion et à leurs affects dans leur relation à la machine »[2].

En termes de milieux sociaux, il a été montré dans le cadre du projet TypWeb que les usages des outils de communication sont variables selon l'âge et la catégorie socio-professionnelle de l'utilisateur (voir [Beaudouin *et al.* 2002]). En matière d'outils de communication, les internautes utilisant de façon privilégiée le courrier électronique sont plus fréquemment des cadres et professions intermédiaires ; à l'inverse, les outils de communication synchrone (*chat*, messagerie instantanée) sont préférentiellement utilisés par les jeunes et les individus appartenant à des foyers dont le chef de famille est employé ou ouvrier. L'hypothèse avancée par Beaudouin pour expliquer cette différence, relate [Lelong 2003], repose sur la distance à la culture légitime et les barrières à l'écrit propres aux jeunes adultes issus de milieux modestes : « ainsi s'expliquerait leur préférence pour des échanges écrits rapides, quasi conversationnels, sans traces durables et donc moins exposés que le mail à des jugements sociaux valorisant la correction orthographique et grammaticale, et érigeant la lettre manuscrite en modèle de communication. »

La navigation sur le Web est également influencée par ces clivages sociaux : « Dans les familles des milieux favorisés, la pratique de lecture des livres est valorisée et structurée selon les schémas de la culture humaniste et classique. […] Plus on descend dans l'échelle sociale, plus les lycéens décrivent Internet comme un gisement de connaissances digne de foi et supérieur aux autres. »[3]

---

[1] Voir [DiMaggio *et al.* 2001] pour un panorama des recherches en sociologie sur Internet, structurées autour de cinq thématiques : les inégalités (*digital divide*), les communautés et les collectifs, les implications politiques, l'impact sur les organisations, et la diversité culturelle.

[2] [Jouët 2003], p. 81.

[3] [Lelong 2003], p. 114.

Cette inscription sociale des pratiques Web n'entre que partiellement dans notre champ d'investigation : notre objectif n'est pas de différencier les pratiques sur la base des catégories socio-professionnelles, mais ces variables peuvent localement être mobilisées pour expliquer et interpréter des comportements. Elle agit également comme garde-fou : on s'attachera à tenir compte des déterminants socio-économiques dans l'examen des centres d'intérêt, des contenus des parcours, de l'expertise, tant il est vrai que le capital social, culturel et technique est un déterminant important des pratiques.

> ***Synthèse.*** *L'usage du Web n'échappe pas plus que toute autre activité à des déterminations sociales, qui influencent les parcours sur le plan des contenus autant que des modalités. Pour autant, c'est surtout dans l'usage de la Toile que l'on cherchera les éléments de contextualisation les plus pertinents. L'étude des pratiques d'Internet en général et l'analyse de la structure et des modes d'appréhension des territoires personnels alimentent ainsi une approche éthologique et contextuelle des parcours sur le Web.*

Conclusion

En positionnant l'étude des parcours sur le Web dans le champ des sciences humaines et sociales, nous nous écartons sensiblement des travaux qui ont pu être menés jusqu'alors sur la navigation : tentatives de modélisation des comportements d'utilisateurs fortement influencées par les sciences cognitives d'une part, et point de vue centré-serveur d'autre part. Notre démarche est descriptive, et se concentre sur l'activité de navigation du côté de l'utilisateur, avec la diversité des situations et des contenus que cela implique. Pour appréhender cette complexité, nous avons élaboré des outils complémentaires de fouille des données de trafic : de manière qualitative, la possibilité de visualiser et de refaire les parcours permet à la fois de formuler et de vérifier des hypothèses ; les indicateurs topologiques simples que nous avons mis en place autorisent quant à eux des traitements statistiques de masse sur la forme, le rythme et la temporalité des parcours. Couplés aux descriptions de contenus au niveau de la page et du site d'une part, et aux éléments de contextualisation individuelle des pratiques d'autre part, ils forment une base solide pour l'analyse des parcours, leur description, leur segmentation et leur interprétation.

II

Usages et comportements de navigation sur le Web

A partir des données brutes de trafic, nous nous sommes dotés d'un outillage complet pour l'analyse des parcours : représentation des contenus à l'aide des annuaires Web et de l'application *CatService*, indicateurs statistiques rendant compte de la forme et de la temporalité des sessions, éléments de contextualisation sous l'angle des territoires personnels et des pratiques Internet, modules de visualisation et d'examen manuel des parcours. On dispose ainsi d'une description complète des parcours intégrant, de la page à l'utilisateur, la description de la production et de l'offre sur le Web et les modalités de leur appréhension par les internautes.

Cette seconde partie montre la mise en œuvre de ces outils et descripteurs sur trois panels d'internautes centrés-utilisateur pour l'analyse des usages et des comportements de navigation sur le Web. Dans un premier temps, nous centrons notre analyse sur une vue locale et décontextualisée des parcours afin de construire une segmentation des sessions basée uniquement sur leur forme et leur contenu. Les cinq parcours-type que nous identifions sont ensuite examinés en regard des pratiques d'Internet et des territoires personnels des utilisateurs sur le Web ; cet examen contextuel de la navigation nous amène à identifier des modes prototypiques d'appréhension des sites Web en fonction de leur contenu et de leur valorisation par l'utilisateur.

Chapitre 5

Contenus et formes de parcours

Nous présentons dans ce chapitre les trois sources de données que nous avons exploitées pour cette étude : un panel résidentiel général large en 2002, une cohorte suivie à domicile de 2000 à 2002, et un échantillon restreint fin 2002 centré sur les usages des bibliothèques électroniques. Après avoir décrit la composition de ces panels, leurs spécificités et les résultats particuliers que nous souhaitons en retirer, nous donnons un portrait général des usages du Web en termes d'intensité, de régularité, de thématiques et de services qui nous amène à établir une première segmentation des parcours.

5.1 Description des panels

Nous travaillons sur trois sources de données complémentaires. La première porte sur l'année 2002 et concerne un panel large de 3 372 individus ; la seconde, longitudinale, est centrée sur 597 internautes français observés de 2000 à 2002 ; la dernière est un panel de taille plus modeste formé de 72 utilisateurs de bibliothèques électroniques.

5.1.1 Panel SensNet 2002

Le panel SensNet 2002 dont nous disposons est issu des données de trafic recueillies par NetValue dans le cadre du projet SensNet (voir Annexe 1). Cette cohorte est constituée de 3 372 internautes dont l'activité Web est observée de janvier à octobre 2002, soit 10 mois d'observation. Ce panel, désigné dans la suite du document par l'abréviation 'SN2002', a été élaboré avec la méthode des quotas par la société NetValue : il est globalement représentatif de la population des internautes français sur la période. En effet, nous ne disposons pas ici du panel de NetValue proprement dit : celui-ci, conçu pour offrir une vue représentative de la population des internautes à tout moment à des fins de mesure d'audience, est corrigé chaque mois pour y intégrer les nouveaux internautes et assurer sa représentativité. Notre échantillon suit une logique quelque peu différente : il s'agit d'étudier une cohorte

d'internautes et d'observer l'évolution des usages de ces individus sur une période donnée.

Le panel SN2002 est alors représentatif de la population des internautes français en janvier 2002, mais cette représentativité s'affaiblit au fil des mois. C'est au chapitre de l'ancienneté de la pratique que l'échantillon est le plus biaisé par rapport à la population des internautes en général (voir Tableau 5.1) : les primo-accédants sont absents des données, les internautes « récents » sous-représentés (5,5 % du panel connecté pour la première fois en 2001), la moitié du panel a deux à trois ans de pratique, et le tiers était connecté en 1998 ou avant.

Tableau 5.1. SN2002, ancienneté de la pratique Web

Année de première connexion	% du panel
Avant 1997	8,8 %
1997	7,9 %
1998	18,2 %
1999	26,8 %
2000	26,8 %
2001	5,5 %
Ne sait pas	6,0 %

Malgré cette sur-représentation des « anciens internautes », le profil du panel en termes d'âge, de sexe et de type de connexion suit globalement ceux de l'ensemble des internautes français à la même époque (voir Tableau 5.2).

Tableau 5.2. SN2002 : caractéristiques générales

		SN2002 en janvier 2002	Internautes français en décembre 2001
Homme / femme		56,4 % / 43,6 %	58 % / 42 %
Âge	Moins de 15 ans	7,8 %	4,6 %
	15-24 ans	22,8 %	24,7 %
	25-34 ans	22,3 %	23,1 %
	35-49 ans	31,5 %	31,8 %
	50-64 ans	12,6 %	12,5 %
	Plus de 65 ans	3,0 %	3,3 %
Bas débit / haut débit (autre, n.s.p)		90,1 % / 7 % (2,9 %)	91,1 % / 8,9 %

Le taux d'équipement du panel en accès Internet à haut débit est similaire à celui de l'ensemble de la population, alors que nous aurions pu nous attendre, chez ces internautes confirmés, à une part importante d'abonnements au câble ou à l'ADSL : ceci est sans doute dû au fait que c'est en 2002 que l'accès haut débit s'est diffusé largement en France, notre panel étant observé ici en janvier 2002.

En termes de catégories socio-professionnelles (CSP), le panel SN2002 est également très proche des caractéristiques de l'ensemble des internautes français à la même époque (décembre 2001) ; par rapport à la population française en général, les professions intermédiaires sont surreprésentées, au détriment des ouvriers et des agriculteurs (voir Tableau 5.3).

Tableau 5.3. SN2002 : CSP des panélistes

Occupation	Nb. panélistes	Part du panel	Internautes français, déc. 2001
Retraité(e)	186	5,5 %	6,2 %
Sans profession, au foyer	416	12,2 %	9,4 %
Temporairement sans emploi (chômage, maladie, etc.)	62	1,8 %	1,6 %
Étudiant(e)	727	21,4 %	23,3 %
Ouvrier	146	4,3 %	4,4 %
Agriculteur, pêcheur	8	0,2 %	0,3 %
Employé(e)	594	17,5 %	18,8 %
Profession intermédiaire (cadre, chef de service, chef de groupe, technicien, etc.)	851	25,0 %	24,2 %
Profession libérale	69	2,0 %	1,5 %
Dirigeant (PDG, DG, directeur, cadre sup., etc.)	282	8,3 %	8,4 %
Propriétaire d'une entreprise, artisan, commerçant, autre travailleur indépendant	49	1,4 %	1,5 %
Non renseigné	8	0,2 %	-

On retrouve là des éléments présentés maintes fois dans des études et des réflexions autour du thème de la « fracture numérique », nous n'irons donc pas plus avant sur ce sujet. Dans le cadre présent, nous nous contenterons de constater que malgré une présence forte d'anciens internautes, notre échantillon est représentatif en termes de CSP.

> ***Synthèse.*** *La taille, la représentativité et la durée d'observation du panel SensNet 2002 permettent de travailler dans la masse et d'observer des comportements récurrents entre sessions et entre individus. Ce panel sert de toile de fond à la mobilisation d'autres jeux de données moins représentatifs.*

5.1.2 Panel longitudinal 2000-2002

L'échantillon longitudinal sur lequel nous travaillons est également issu des données du panel NetValue : il s'agit d'un sous-ensemble du panel SN2002, dont on observe 597 individus présents dans le panel de 2000 à 2002 (données issues des projets TypWeb et SensNet) ; nous le désignons par la suite par l'abréviation 'SN00-02'. Le mode de sélection des internautes de cette cohorte laisse de côté les abandonnistes, et favorise plus encore que dans l'échantillon précédent les internautes anciens et « fidèles », qui ont ancré l'usage d'Internet dans leurs pratiques. On ne s'étonnera pas, dès lors, de constater que la moyenne d'âge de la cohorte SN00-02 est centrée sur la tranche 35-49 ans (voir Tableau 5.4 ci-dessous), alors que les âges sont beaucoup plus répartis pour l'ensemble des internautes français à la même période.

Tableau 5.4. Panel SN00-02 : caractéristiques générales

		SN00-02 en 2000	SN00-02 en 2002
Homme/femme		67,5 % / 32,5 %	
Lieu d'habitation	Rural	16,7 %	15,2 %
	2 000 à 19 999 hab.	13,4 %	14,4 %
	20 000 à 100 000 hab.	10,2 %	10,4 %
	100 000 et + hab.	38,6 %	37,8 %
	Région parisienne	21,1 %	22,3 %
Âge	Moins de 15 ans	10,4 %	3,6 %
	15-24 ans	21,5 %	24,3 %
	25-34 ans	18,2 %	19,3 %
	35-49 ans	35,5 %	36,0 %
	50-64 ans	13,2 %	14,5 %
	Plus de 65 ans	1,3 %	2,3 %
Bas débit / haut débit (n.s.p)		nc.	89 % / 9 % (2 %)

Corrélativement, les catégories socio-professionnelles sont plus proches de celles des internautes français avant 2000 qu'en 2002 (voir Tableau 5.5) : les agriculteurs sont absents de la cohorte, et la part des ouvriers et des employés diminue au profit des professions intermédiaires et libérales.

Tableau 5.5. Panel SN00-02 : CSP des participants de SN00-02

Occupation	SN00-02		Internautes français, déc. 2001
	En 2000	En 2002	
Retraité	3,8 %	4,6 %	6,2 %
Sans profession, au foyer	11,4 %	5,8 %	9,4 %
Temporairement sans emploi (chômage, maladie...)	2,5 %	3,1 %	1,6 %
Étudiant	23,4 %	23,3 %	23,3 %
Ouvrier	1,8 %	1,7 %	4,4 %
Agriculteur, pêcheur	0 %	0 %	0,3 %
Employé	15,0 %	16,5 %	18,8 %
Profession intermédiaire (cadre, chef de service, technicien, etc.)	31,5 %	33,7 %	24,2 %
Profession libérale	1,5 %	2,1 %	1,5 %
Dirigeant (PDG, DG, directeur, cadre supérieur...)	6,9 %	7,1 %	8,4 %
Propriétaire d'une entreprise, commerçant	2,1 %	2,1 %	1,5 %

Pour ce qui est de l'analyse des parcours Internet, attendons-nous également à ce que les contenus visités reflètent en partie les centres d'intérêt d'une population au pouvoir d'achat moyen ou élevé, et que certains services et modes de communication soient sous-représentés.

Ces données présentent un intérêt très important : il n'existe pas à ce jour d'étude d'usages d'Internet portant sur une période d'observation aussi longue, ni sur une cohorte aussi importante. De telles données permettent de répondre à des questions jusqu'alors en suspens, en particulier sur l'évolution des pratiques au fil du temps et, pour le cadre plus précis de la navigation Web, l'établissement de routines et de visites fréquentes de certains sites.

Synthèse. *Le panel SensNet 2000-2002 offre une durée d'observation inédite d'une cohorte d'internaute résidentiels. De telles données longitudinales autorisent notamment une analyse de la structure et de l'évolution des territoires personnels sur le Web.*

5.1.3 Panel BibUsages

Le dernier lot de données de trafic est apporté par le projet BibUsages, mené par France Télécom R&D et la Bibliothèque Nationale de France, qui vise à étudier l'usage des bibliothèques électroniques en ligne par le grand public[1]. Si ces données ne sont pas représentatives des pratiques Web en général, leur intérêt est ailleurs : il s'agit de donner un éclairage particulier sur des pratiques non observables dans les échantillons précédents du fait du peu d'internautes concernés. En outre, on perd en quantité ce que l'on gagne en qualité : la méthodologie menée ici est plus complète que pour les panels SN2002 et SN00-02, et permet de mener des études qualitatives fines.

Phases du projet et méthodologie

Le projet BibUsages s'est déroulé en trois phases :

1. Enquête de cadrage *via* un questionnaire en ligne (mars 2002).
2. Constitution d'un panel et recueil de trafic Web pour ce panel (avril-décembre 2002).
3. Enquête qualitative par entretiens (octobre 2002).

Dans la première phase de l'expérimentation, un questionnaire a été soumis aux visiteurs du site Gallica en mars 2002 durant trois semaines. Il a permis à la fois d'avoir une connaissance plus précise du public de Gallica, et de recruter les volontaires pour faire partie du panel d'utilisateurs dont le trafic Web a été enregistré. Le questionnaire utilisé est fourni en annexe (voir Annexe 4).

Outre les caractéristiques socio-démographiques des répondants, le questionnaire s'articule autour de deux thématiques principales : d'une part, l'usage de Gallica (fréquence des visites, rubriques consultées, etc.), et d'autre part les usages d'Internet en général (intensité d'usage, services utilisés, types de sites visités, etc.). À la fin du questionnaire, les répondants se sont vus proposer de participer au panel d'utilisateurs mis en place. Au terme de cette première étape, 2 340 personnes ont répondu au questionnaire, et 589 ont donné leur accord de principe pour faire partie du panel d'utilisateurs, soit près d'un quart.

Dans un deuxième temps, un panel représentatif de la population totale des répondants au questionnaire a été constitué, composé de 72 volontaires qui ont téléchargé et installé le dispositif de recueil de trafic. Les premières données de trafic nous parviennent fin mai 2002, mais l'installation du dispositif de recueil n'est achevée sur la quasi-totalité des postes que début juillet. À partir de juillet, l'ensemble des volontaires avaient installé le dispositif de recueil de trafic sur leur

[1] Projet soutenu par le Réseau National de Recherche en Télécommunications ; voir Annexe 1 pour une description complète du projet.

poste ; leur activité Web a été enregistrée pendant six mois, jusqu'en décembre 2002[1].

En complément des données de trafic, des entretiens semi-directifs ont été menés auprès de 16 des 72 participants à l'expérimentation[2]. Les entretiens ont été organisés en juillet et réalisés en octobre 2002. Sur les 72 membres actifs du panel, 50 ont été contactés afin d'obtenir leur accord de participation aux entretiens, et ce en s'intéressant aux membres les plus actifs en fonction de leur trafic Internet.

Le questionnaire de Gallica de mars 2002 et le trafic enregistré par Audinet[3] permettent de dégager des profils d'usages pour chaque utilisateur interviewé, en particulier autour des types de contenus visités et de l'intensité des pratiques. Néanmoins les modalités et les contextes d'utilisation restent à approfondir. Les entretiens comblent ce vide en permettant d'une part de confirmer les pratiques telles qu'elles émergent de l'analyse du trafic et d'autre part de les inscrire dans leur contexte (usages d'Internet en général – et pas seulement du Web – et les pratiques hors-ligne). Il s'agit ainsi de voir comment la consultation des bibliothèques numériques s'inscrit dans une pratique générale d'Internet dans un contexte donné, et de mieux connaître les différents types d'usages, les motivations et les modalités de la pratique avec comme appui les données de trafic.

Pour mener ces entretiens, une grille d'entretien reprenant en filigrane ces objectifs, fournie en annexe (voir Annexe 4, Matériau d'enquête BibUsages), a été élaborée autour de trois axes :

1. Une première partie centrée sur l'utilisation générale d'Internet permet de mieux cibler le profil général de l'internaute dans ses pratiques : durée, motivations, contexte de l'utilisation, modalités des recherches et traitement de l'information. En outre, cela permet d'avoir des renseignements sur les usages hors Web (*chat*, mail, forums, *peer-to-peer*…) que la sonde Audinet, dans la version utilisée pour cette étude, n'enregistre pas.
2. La deuxième partie de la grille se concentre sur l'usage des bibliothèques électroniques et de Gallica en particulier. Il s'agit de connaître les contextes d'utilisation des fonds numériques, les méthodes de recherche et les modalités de traitement de l'information. Dans la discussion, on cherche également à pressentir des difficultés et à obtenir des propositions d'amélioration dans la conception du site Gallica.
3. La troisième partie de l'entretien se concentre sur les pratiques « off-line » : il s'agit ici de relier l'utilisation des bibliothèques électroniques à celle des

---

[1] Nous tenons particulièrement à remercier les personnes qui ont accepté de participer à cette étude, en installant le dispositif de recueil de trafic et en participant aux entretiens que nous avons menés. C'est grâce à leur collaboration et au temps qu'ils ont consacré à ce projet que celui-ci a pu être mené à bien.

[2] La plupart des entretiens ont été menés par France de Charentenay (BnF) ; nous avons mené ou participé à la plupart d'entre eux.

[3] Sonde de recueil de trafic Internet développé par France Télécom R&D ; voir chapitre 2.1.1, « Technologies de recueil de données ».

bibliothèques classiques et, plus largement, aux pratiques de lecture et aux pratiques culturelles des interviewés.

Pour chaque entretien, une fiche descriptive du panéliste a été élaborée à partir de ses réponses au questionnaire de Gallica (mars 2002) et des statistiques de son trafic Internet déjà recueilli *via* Audinet. En s'appuyant sur ces informations, l'entretien permet de confirmer et d'expliciter les pratiques observées.

La richesse de la méthodologie suivie réside dans l'exploitation conjointe et croisée des données issues des trois phases du projet. Le questionnaire donne une description précise de la population d'ensemble et nous permet ainsi de bien situer l'analyse plus fine des phases 2 et 3 dans un cadre général. L'approche qualitative permet de valider des hypothèses issues de la phase d'analyse de trafic, de même que l'analyse de trafic permet de consolider les conclusions issues de l'analyse des entretiens, en les appuyant sur des mesures objectives des usages.

Profil général à l'issue du questionnaire

En premier lieu, on note une part importante de visiteurs étrangers : parmi les répondants à l'enquête, 31,4% déclarent résider à l'étranger ; en outre, pour 60% d'entre eux, le français n'est pas leur langue maternelle (mais ils la pratiquent assez pour répondre au questionnaire). Gallica s'impose ainsi comme un point d'accès à des fonds francophones depuis l'étranger.

Dans le « lectorat » de Gallica, tel qu'il ressort de l'enquête, on constate les tendances fortes suivantes :

- un niveau d'études élevé : 33,8 % entre Bac+2 et Bac+4, et 38,3 % de diplômés de troisième cycle ;
- une représentation majoritaire des plus de cinquante ans, qui comptent pour 32,7 % des répondants, au détriment des moins de trente ans (11,8 %) ;
- une sur-représentation des cadres de la fonction publique, très loin devant les autres catégories (l'enseignement supérieur est le secteur d'activité le plus représenté).

En ce qui concerne les usages d'Internet, les personnes interrogées déclarent un haut degré de pratique, utilisant les services Web de façon très régulière (recherche d'informations ou opérations bancaires ou boursières, achats en ligne), tout cela majoritairement à partir d'un accès à domicile. En même temps, il faut remarquer que ces mêmes utilisateurs sont d'anciens internautes (37% d'entre eux sont connectés à Internet depuis 1997 ou antérieurement). On voit donc que le profil général des utilisateurs rencontrés dans cette enquête se situe dans un contexte de régularité et de fidélité, de connaissance de l'outil, avec un usage personnel à partir du domicile et dans des sessions plutôt longues.

Pour analyser la spécificité du public de Gallica par rapport aux internautes français, nous recourons aux chiffres données par NetValue en décembre 2001, que nous comparons aux caractéristiques des répondants au questionnaire résidant en France.

Tableau 5.6. « Gallicanautes » résidant en France et internautes français

		France - données NetValue décembre 2001	Répondants BibUsages résidant en France (mars 2002)
Homme / femme		58% / 42%	69,3% / 30,7%
Urbain		81,0%	86,4%
Âge	Moins de 15 ans	4,6 %	0,5 %
	15-24 ans	24,7 %	9,3 %
	25-34 ans	23,1 %	19,2 %
	35-49 ans	31,8 %	32,8 %
	50-64 ans	12,5 %	31,1 %
	Plus de 65 ans	3,3 %	7,1 %
Bas débit / haut débit		91,1% / 8,9%	67,2% / 32,8%

Les caractéristiques générales (voir Tableau 5.6) montrent que les gallicanautes sont, par rapport aux internautes français, plutôt des hommes, globalement plus âgés (sur-représentation des tranches 50-64 ans et plus de 65 ans). L'écart le plus important concerne le type d'équipement Internet : le haut débit est fortement sur-représenté, avec un taux d'équipement de près de 33% chez les utilisateurs de Gallica, contre 9% en général à la même époque. Les gallicanautes sont également des internautes plus « anciens », avec 67,2% des répondants résidant en France disposant d'une connexion depuis 1999 au moins.

Tableau 5.7. Spécificités des gallicanautes résidant en France : CSP[1]

France - données NetValue (déc. 2001)		Questionnaire BibUsages en ligne (mars 2002)	
Retraité	6,2 %	Retraité	10,1 %
Sans profession, au foyer	9,4 %	Au foyer	1,3 %
Temporairement sans emploi	1,6 %	Temporairement sans emploi	2,5 %
Étudiant	23,3 %	Étudiant	18,4 %
Ouvrier	4,4 %	Ouvrier	0,3 %
Agriculteur, pêcheur	0,3 %	Agriculteur, exploitant	0 %
Employé	18,8 %	Employé, personnel de service	6,5 %
Profession intermédiaire	24,2 %	Technicien, agent de maîtrise, cat. B	10,0 %
Profession libérale	1,5 %	Profession libérale	6,0 %
Propriétaire d'une entreprise	1,5 %	Commerçant, artisan	0,5 %
		Cadre du secteur privé	13,8 %
Dirigeant	8,4 %	Cadre de la fonction publique (cat. A)	26,5 %
		Chef d'entreprise, cadre dirigeant	4,0 %

L'examen des professions et catégories socio-professionnelles des répondants au questionnaire résidant en France (voir Tableau 5.7) complète l'analyse en montrant une sur-représentation des CSP supérieures (cadres, enseignants), tandis qu'au sein des étudiants, les 2e et 3e cycle sont sur-représentés parmi les visiteurs de Gallica.

[1] Les grilles de CSP employées par NetValue et dans le questionnaire BibUsages étant différentes, nous proposons des équivalences, sans garantir leur concordance.

Composition du panel BibUsages

Le panel BibUsages, composé d'internautes chez qui nous avons installé le dispositif de recueil de trafic Web, a été composé de telle manière qu'il reflète le plus fidèlement possible les caractéristiques des répondants au questionnaire présenté sur Gallica en mars 2002. Cependant, en raison de contraintes d'organisation, le panel est exclusivement composé de personnes résidant en France : les répondants étrangers à l'enquête n'y sont donc pas du tout représentés. Nous donnons ici les caractéristiques de ce panel, qui reprend de manière générale celles des « gallicanautes » présentées précédemment.

En termes socio-démographiques, le panel BibUsages est essentiellement masculin (¾ d'hommes). La moyenne d'âge est de 46 ans ; toutefois, cette moyenne ne met pas en évidence la sur-représentation des panélistes de plus de 55 ans, avec plus de 31 % du panel se situant dans les deux dernières tranches d'âge contre 19 % pour l'enquête de Gallica. Les panélistes exercent majoritairement une activité professionnelle (82 %) et pour 45 % d'entre eux, occupent un poste cadre de la fonction publique, dans le secteur de l'enseignement ou de la recherche. Dans le cadre des panélistes n'ayant pas d'activité rémunérée (18 %), plus de 60 % sont des retraités, alors qu'ils n'étaient que 23,4 % dans l'enquête générale. De plus il faut noter que 20 % d'entre eux sont des étudiants de deuxième cycle. En ce qui concerne la provenance géographique du panel, la diversité des régions françaises a été respectée avec une sur-représentation de Paris et la région parisienne.

Plus de 47 % des panélistes utilisent Internet depuis 1997, de manière quotidienne et dans la majorité des cas de leur domicile (plus de 62 %). Ils sont largement équipés d'accès haut-débit, puisque 29 des 72 participants ont une connexion par ADSL ou Câble. Le questionnaire en ligne nous apprend que les panélistes utilisent principalement Internet pour rechercher des informations (30 %) et échanger à partir des différents modes de communication (mail, *chat*, forum). Leurs principaux centres d'intérêts sur le Web sont assez hétérogènes, avec une prédominance de l'art et de la littérature, des sciences sociales et de la recherche documentaire ou bibliographique.

Le panel est une population d'internautes fidèles à Gallica : plus de 47 % déclarent en être à plus de dix visites. Ils fréquentent le site de manière régulière et pour la plupart de chez eux. Leurs visites sont d'une durée (déclarée) supérieure à 10 minutes : 38 % disent rester entre 10 à 30 minutes et 32 % plus d'une demi-heure. Les panélistes ont connu Gallica par différents moyens : brochure de la Bibliothèque Nationale de France, lien d'un autre site, requête *via* un moteur de recherche ou annuaire Web.

Profil général des personnes interrogées en entretien

Le profil socio-démographique des seize personnes interrogées reste dans la lignée du panel, ainsi que de la population générale de l'enquête de Gallica (voir Tableau 5.8 ci-dessous). La plupart des interviewés occupent des postes de cadre de la fonction publique ou du secteur privé ; la moyenne d'âge se situe autour de 48 ans avec une fourchette allant de 33 à 76 ans, et la majorité habite en milieu urbain.

Tableau 5.8. Récapitulatif des profils des panélistes interrogés en entretien

Utilisateur	Sexe	Département	Tranche d'âge	Profession	Type de connexion à domicile	Ordinateur partagé avec d'autres membres de la famille	Nbre total de sessions réalisées	Lieu d'installation d'Audinet Domicile/ Travail	Ancienneté Internet
Utilisateur A	F	16	55/59	Enseignante du supérieur	Haut débit	Oui	905	D	1998
Utilisateur B	F	75	30/34	Indépendante Correctrice	Haut débit	Oui	110	D/T	1997
Utilisateur C	F	35	55/59	Cadre de la fonction publique	Haut débit	Non	pb. install.	D	1997
Utilisateur D	H	94	35/39	En Formation Sciences Sociales	Haut débit	Oui	642	D	1997
Utilisateur E	H	83	70/74	Retraité (cadre du privé)	RTC	Oui	61	D	2000
Utilisateur F	H	77	50/54	Enseignant du primaire	RTC	Oui	150	D	1998
Utilisateur G	H	75	40/44	Enseignant (Histoire)	Haut débit	Oui	626	D	1999
Utilisateur H	H	75	45/49	Cadre du privé (informatique)	Haut débit	Oui	28	D	1999
Utilisateur I	H	83	45/49	Employé, personnel de service	RTC	Oui	667	T	1998
Utilisateur J	H	33	50/54	Cadre de la fonction publique	Numéris	Non	278	D	2000
Utilisateur K	H	27	30/34	Cadre du privé	Haut débit	Non	36	T	1997
Utilisateur L	H	69	30/34	Cadre de la fonction publique (CNRS)	RTC	Non	218	T	1997
Utilisateur M	H	59	50/54	Enseignant du supérieur	Haut débit	Oui	381	D	1997
Utilisateur N	F	92	55/59	cadre du privé (assurance)	RTC	Non	262	D	1997
Utilisateur O	H	91	75/79	Retraité (comptable)	RTC	Non	658	D	2000
Utilisateur P	H	15	65/69	Retraité (enseignant du secondaire)	Haut débit	Non	101	D	1997

En ce qui concerne l'utilisation d'Internet, on remarque qu'il s'agit d'anciens internautes, avec en moyenne déjà plus de 3 ans de pratiques, et ayant un usage fréquent et une consommation importante (déclarée dans le questionnaire en ligne, et confirmée par les données de trafic).

Les participants ont en général installé Audinet chez eux. Pour certains, cet ordinateur est un ordinateur familial dont ils sont l'utilisateur principal ; il ressort des entretiens que l'utilisation par les autres membres du foyer reste occasionnelle. Les choix individuels se sont portés sur l'équipement le plus utilisé pour l'installation de la sonde, c'est pourquoi seules deux personnes ont installé ce logiciel sur leur lieu de travail.

C'est sur les données de trafic issues de ces trois panels – SN2002, SN00-02 et BibUsages – que nous appliquons nos outils et nos méthodes d'enrichissement et d'analyse de trafic afin de décrire les parcours des utilisateurs sur le Web. Toutefois, nous ne mettons pas ces trois sources de données sur le même plan : le panel SN2002 est le point focal de notre analyse, car il s'agit à la fois du plus récent, du plus volumineux et du plus représentatif des trois. Les panels SN00-02 et BibUsages sont mobilisés en appui, pour vérifier si les résultats obtenus sur le panel de référence sont observables chez ces deux autres populations. En outre, les données sur trois ans sont mobilisées pour les études longitudinales, afin d'observer des effets d'habitude et de constitution de territoires, tandis que les données BibUsages nous permettent de faire un éclairage ciblé sur une population atypique d'internautes anciens, intensifs, fortement équipés en haut débit et consommateurs de contenus culturels.

> ***Synthèse.*** *Le panel BibUsages, de plus petite taille que les précédents, est centré sur la population des internautes visiteurs de bibliothèques numériques en ligne. Sa construction suit une méthodologie en « entonnoir » : un questionnaire cerne le profil général de cette population, dont un panel est extrait pour faire du recueil de trafic, et des entretiens avec seize panélistes viennent compléter qualitativement ces données. Les trois panels (SN2002, SN00-02 et BibUsages) sont complémentaires, et chacun répond à des problématiques spécifiques.*

5.1.4 Usages généraux d'Internet

Avant d'entrer plus précisément dans l'analyse des sessions Web et des parcours de site en site, nous souhaitons nous doter de descriptions sur les usages généraux d'Internet au sein de nos trois panels. Ces données doivent servir de cadrage pour ce qui va suivre, et replacent les parcours Web dans la perspective plus générale de l'accès au réseau. Car si nous avons fait le choix ici de ne traiter que des parcours sur le Web, nous n'oublions pas qu'une approche plus vaste se doit d'analyser l'ensemble de l'activité de chaque utilisateur en considérant que les ruptures techniques entre les dispositifs qui sous-tendent l'Internet ne sont pas forcément perçues par l'utilisateur et qu'il y a pour lui une véritable continuité entre Web, messagerie, forum, etc. autour d'un objet unique, l'ordinateur.

Panels SensNet : profils d'usage différenciés

Les données SensNet concernent l'ensemble des protocoles mobilisés dans l'accès à Internet : on dispose d'informations sur l'activité Web, mais aussi sur la messagerie, les jeux en ligne, le *chat*, l'échange de fichiers, etc. Ces traces complètes de l'activité sur le Net permettent de construire une description des internautes en fonction de leur utilisation ou non de ces différents services, et de l'intensité d'usage qui y est attachée.

Pour cela, le panel SensNet 2002, représentatif des internautes sur les dix mois d'observation, nous sert de trame de fond : pour chaque panéliste, on examine le nombre de sessions pour chaque type d'activité (Web, mail, *peer-to-peer*, etc.) de janvier à octobre 2002. Le travail de classification sur cette base est compliqué par le fait que les variables ne sont pas homogènes : certains types d'activité comme le Web ou la messagerie sont très présents et partagés par la quasi-totalité des internautes, tandis que le *chat* ou le téléchargement concernent peu d'individus, et impliquent moins de sessions. Pour contourner cette difficulté, nous avons discrétisé chaque variable de manière *ad hoc*, en tenant compte à la fois du fait d'avoir utilisé ou non un protocole, et du nombre de sessions que son utilisation implique à l'échelle du panel (voir Tableau 5.9). Certains services comme les forums sont ainsi discrétisés en « utilisation / pas d'utilisation », tandis que l'usage du Web est décomposé en trois modalités : peu intensif en deçà de 20 sessions sur les dix mois, intensif au-delà de 130 sessions.

Tableau 5.9. Variables et modalités pour la description des usages d'Internet

Variable	Modalités	Intervalles (nb de sessions) sur 10 mois	Répartition des panélistes
Chat/IRC	Pas de chat/IRC	0	70,4 %
	Chat/IRC faible	1-2	13,0 %
	Chat/IRC intense	3 et plus	16,6 %
Web	Web - peu	0-19	29,2 %
	Web - moyen	20-129	41,4 %
	Web - intense	130 et plus	29,4 %
WebMail	Pas de WebMail	0	28,1 %
	WebMail - peu	1-4	24,6 %
	WebMail - moyen	5-39	29,6 %
	WebMail - intense	40 et plus	17,7 %
WebChat	Pas de WebChat	0	67,4 %
	WebChat - faible	1-2	15,3 %
	WebChat - intense	3 et plus	17,3 %
Mail	Pas de mail	0	31,5 %
	Mail - peu	1-8	28,4 %
	Mail - moyen	9-69	25,8 %
	Mail - intense	70 et plus	14,2 %
Peer to peer	Pas de p2p	0	78,1 %
	p2p faible	1-15	12,8 %
	p2p intense	16 et plus	9,2 %
Forums	Pas de forum	0	92,8 %
	Forum	1 et plus	7,2 %
Téléchargement	Pas de téléchargement	0	57,8 %
	Téléchargement faible	1-2	20,0 %

	Téléchargement moyen	3-8	13,3 %
	Téléchargement intense	9 et plus	8,9 %
Audio-vidéo	Pas d'audio-vidéo	0	53,0 %
	Audio-vidéo faible	1-2	20,2 %
	Audio-vidéo moyen	3-8	13,2 %
	Audio-vidéo intense	9 et plus	13,6 %
Jeux	Pas de jeux	0	90,6 %
	Jeux	1 et plus	9,4 %

Clef de lecture : un panéliste ayant réalisé deux sessions de Chat/IRC sur les dix mois d'observation sera rattaché à la modalités des « faibles utilisateurs » de ce service.

Chaque panéliste est ainsi décrit par 10 variables discrètes relatives à son intensité d'usage de chaque type de protocole sur Internet. L'analyse en composantes multiples pratiquée sur cette base et la classification sur les dix facteurs premiers nous permet d'identifier quatre classes d'utilisateurs distinctes, que représente la Figure 5.1 ci-dessous[1].

Un premier groupe d'utilisateur se distingue par la faible intensité et le peu de diversité de ses usages (classe n°2 sur le graphique). Chez ces utilisateurs occasionnels, qui représentent 33 % du panel, on ne compte que 2,9 sessions par mois en moyenne (médiane : 1 session), contre 145 pour l'ensemble du panel (médiane : 70). Les services utilisés se limitent alors aux « fondamentaux » : leur activité sur le Net se résume à un peu de Web, et un peu de messagerie.

Un second groupe constitué de 35 % du panel (classe n°1 sur le graphique) présente des usages à la fois plus intenses et plus diversifiés. Cette classe des « internautes ordinaires », qui compte 9,3 sessions par mois en moyenne (médiane : 7,2) a une activité moyenne en ce qui concerne le Web, le mail et le WebMail ; un peu de téléchargement et de contenus audio/vidéo complètent ce menu où l'on ne trouvera ni *peer-to-peer*, ni outils avancés de communication (*chat*, WebChat, forum).

Ces services sont l'apanage d'un troisième groupe d'internautes (classe n°3, 22 % du panel), chez qui la pratique des outils de communication interpersonnelle est particulièrement intense. Corrélativement, l'usage du Web est également très fort chez ces « internautes communicants », qui pratiquent autant la messagerie et le *chat* sur le Web que sur les outils dédiés.

Le dernier groupe, comptant pour 10 % du panel, s'oriente bien plus vers des activités ludiques : le *peer-to-peer*, absent ou très faible dans les trois autres groupes, est ici particulièrement intense, de même que les services audio/vidéo et le téléchargement de fichiers. En ce qui concerne la messagerie instantanée, ces internautes intensifs se distinguent non seulement par une utilisation forte, mais surtout le délaissement du WebChat au profit du *chat* sur logiciel spécifique (ICQ, Messenger, etc.). Sans être déterminante, la pratique des jeux en réseau est également forte dans ce groupe, où elle touche 40 % des individus, contre 10 % dans l'ensemble du panel. Ces utilisateurs intensifs sont enfin ceux qui utilisent le plus le Web, avec 39 sessions par mois en moyenne pour chaque panéliste.

[1] Nous utilisons pour ces calculs le logiciel SPAD (www.cisia.com).

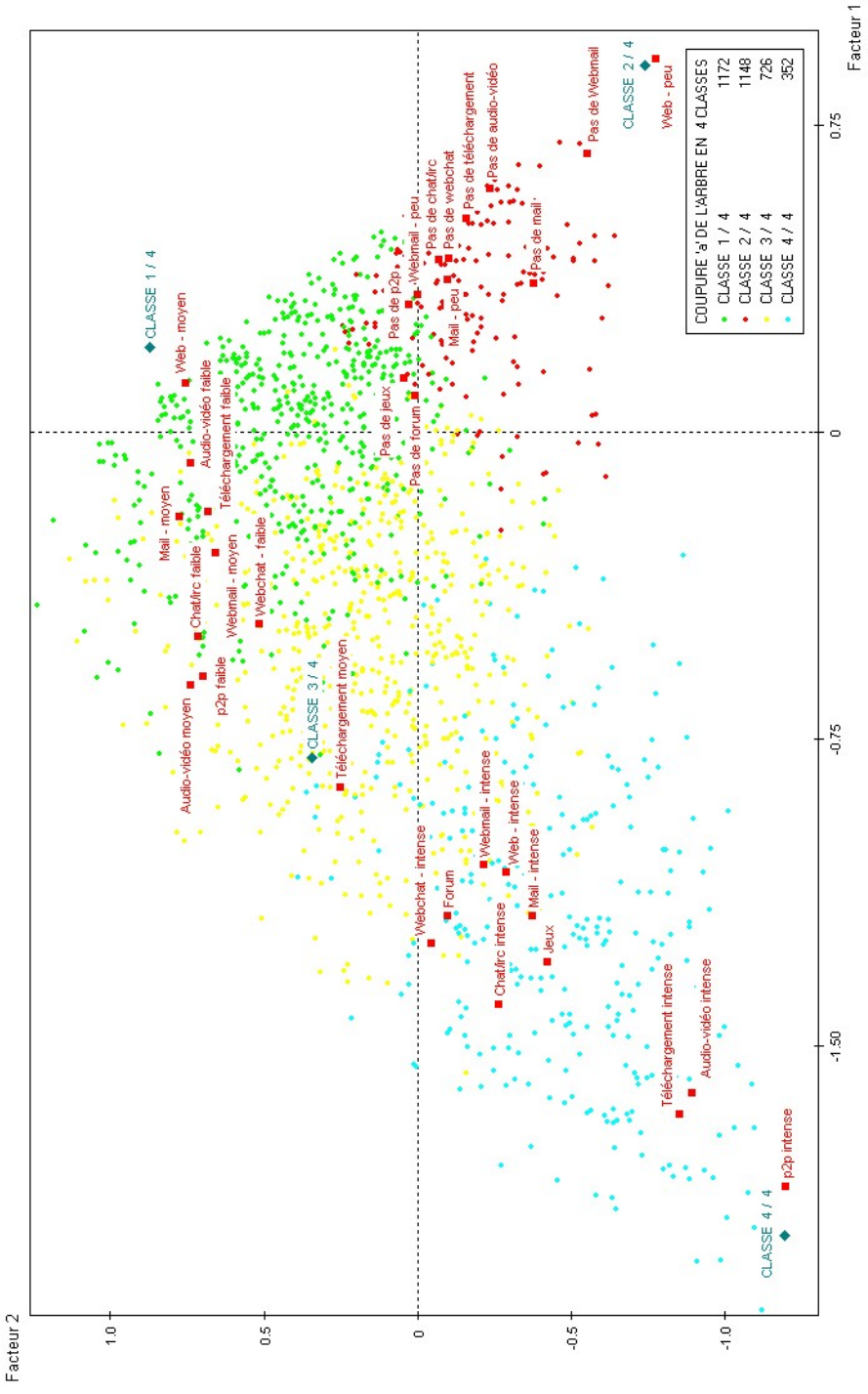


Figure 5.1. Classification des panélistes SN2002 sur la base de leur activité Internet - 4 classes, axes 1 et 2

Les internautes du panel SN00-02 suivis sur trois années constituent, en 2002, un sous-groupe du panel représentatif SensNet 2002. On observera avec intérêt dans quel groupe-type d'usagers ces internautes anciens se positionnent en 2002 : on peut supposer que plus de trois années d'utilisation suivie d'Internet impliquent un ancrage et une intensification des pratiques. Il ne semble pourtant pas que ce soit toujours le cas : la répartition des internautes du panel SN00-02 dans les quatre catégories-types d'utilisateurs suit globalement celle du panel général SN2002 (voir Tableau 5.10).

Tableau 5.10. Répartition des individus dans les groupes-types d'internautes

	Panel SN2002	Sous-panel SN00-02
Utilisateurs occasionnels	33,3 %	24,2 %
Internautes ordinaires	34,8 %	34,8 %
Communication interpersonnelle	21,5 %	27,7 %
Usages ludiques	10,4 %	13,3 %
Total	100 %	100 %

Ces internautes « anciens » sont plus actifs que les autres : ils sont globalement plus présents dans les trois groupes d' « internautes ordinaires », « communication interpersonnelle » et « usages ludiques » que l'ensemble du panel SN2002, au détriment des usages occasionnels. Cependant, un quart d'entre eux se rattache à cette dernière classe, contre un tiers dans l'ensemble du panel SN2002. Ce résultat est intéressant : il montre que l'ancienneté n'est pas un moteur systématique d'intensité d'usage, et que l'utilisation d'Internet peut être ancrée dans les pratiques des individus tout en demeurant une ressource mobilisée de façon occasionnelle.

BibUsages : utilisateurs avertis pour usages ciblés

Le dispositif de recueil de trafic utilisé dans BibUsages a été paramétré pour ne recueillir que les données relatives à l'activité Web, en conséquence de quoi il nous est impossible de pratiquer sur ce panel la même segmentation que pour les données SensNet. Cela étant, nous disposons des 2 340 réponses au questionnaire soumis en ligne, qui permettent de dresser un panorama rapide des usages d'Internet des gallicanautes.

En premier lieu, les répondants au questionnaire se présentent comme des internautes très réguliers et très fidèles : 73 % d'entre eux déclarent utiliser Internet quotidiennement, tandis qu'ils ne sont que 1,5 % à moins de 3 connexions par mois (voir Tableau 5.11).

Tableau 5.11. Questionnaire BibUsages : fréquence d'utilisation d'Internet

À quelle fréquence utilisez-vous Internet ?	% du total
Tous les jours	72,9 %
2 à 5 fois par semaine	21,4 %
Environ une fois par semaine	4,2 %
1 à 3 fois par mois	1,2 %
Moins souvent	0,3 %

En ce qui concerne les usages des différents services disponibles sur Internet (voir Tableau 5.12), les visiteurs de Gallica interrogés semblent se placer plutôt dans

le groupe des « internautes ordinaires » : la recherche d'information sur le Web est pratiquée par la quasi-totalité des répondants, et certaines pratiques comme l'achat ou le recours aux services bancaires en ligne dénotent des usages avancés du Web.

Tableau 5.12. Questionnaire BibUsages : usages d'Internet

Quel usage avez-vous d'Internet ? (plusieurs réponses possibles)	% du total
Recherche d'information	98,8 %
Communication : *chat*, groupes de discussion, messagerie, ...	57,9 %
Achat ou opération financière, dont	48,2 %
- Achat en ligne	34,2 %
- Opérations et consultations bancaires ou boursières	33,1 %
Téléchargement, dont	40,1 %
- Téléchargement de logiciels	32,6 %
- Téléchargement de musique et/ou de vidéo	21,4 %
Jeux en ligne	5,7 %
Autre usage	9,3 %

L'absence d'informations sur l'intensité d'usages des différents services nous empêche d'avoir une vue plus précise : la pratique du téléchargement, qui touche 40 % des répondants, dont la moitié pour des fichiers audio/vidéo, pourrait rattacher les gallicanautes au groupe des « internautes ludiques », mais la faible part du jeu dans les pratiques laisse penser que ce n'est pas le cas, et l'on peut supposer que ces pratiques de téléchargement restent marginales en termes d'intensité.

Enfin, en ce qui concerne les outils de communication, ceux-ci sont utilisés par 58 % des répondants, mais la généricité de cette catégorie ne permet pas de conclure à la place de la communication dans les pratiques des répondants. Les entretiens avec seize membres du panel BibUsages laissent penser que celle-ci reste secondaire : au fil des discussions, il est apparu que l'usage de la messagerie n'arrive pas dans les premières considérations des interviewés. Elle est certes importante mais ne rentre pas dans les motivations personnelles d'accès à Internet : elle permet d'échanger essentiellement avec la famille géographiquement éloignée et les spécialistes rencontrés dans le cadre des recherches. Ce volet orienté vers la communication interpersonnelle est complété par un usage relatif à la recherche d'information, *via* l'abonnement pour certains à des *newsletters* et à des listes de diffusion[1].

Pour les trois panels, l'analyse générale des usages d'Internet montre des variations fortes en termes d'intensité entre les différents individus ; elle met surtout en avant le fait que le Web, à la différence d'autres services comme le *peer-to-peer* ou le *chat*, s'impose comme l'outil fondamental et le pivot de l'activité sur Internet. Non seulement le Web est le seul service mobilisé par les faibles utilisateurs, mais il accompagne également le recours à des outils plus spécifiques (audio/vidéo, téléchargement, etc.) où il est utilisé de manière intensive. L'analyse des parcours

[1] Nous n'avons pas particulièrement mis l'accent sur les usages des outils de communication dans cette étude : le dispositif de recueil de trafic utilisé n'était pas paramétré pour enregistrer cet usage ; par ailleurs, nous n'avons pas cherché à développer cet aspect lors des entretiens.

Web doit tenir compte de cette diversité, qui implique que la navigation s'insère dans des pratiques et des contextes d'usage bien différenciés.

> ***Synthèse.*** *La segmentation des internautes sur la base de l'intensité d'usage des différents outils Internet (Web, mail, messagerie instantanée, jeux, etc.) fait émerger quatre profils distincts : les utilisateurs occasionnels, qui font uniquement du Web et du mail en petite quantité ; les internautes « ordinaires », plus intensifs mais peu diversifiés dans leurs usages ; les communicants, qui ont régulièrement recours aux outils de communication interpersonnelle (*chat, *messagerie instantanée) ; et les utilisateurs ludiques, plutôt orientés vers les jeux ou le* peer-to-peer. *Mobilisé par les quatre groupes, le Web s'impose comme outil pivot de l'activité sur Internet, mais il s'insère, selon les pratiques des individus, dans des contextes d'usages différenciés.*

5.2 Volumétrie, temporalité et topologie des parcours

En première approche des parcours sur le Web, on s'intéressera aux éléments temporels et topologiques de la navigation : ceci permet d'appréhender la diversité des parcours, et de les replacer dans le contexte d'une description de l'activité de navigation préalable à l'examen des contenus et des services visités.

5.2.1 Intensités d'usage variées

Nous présentons ici les résultats de l'analyse des données de trafic en termes de nombre de pages, de sites, et de durée. L'objectif est de donner, avant d'aller plus loin, un panorama général du trafic réalisé par les membres de nos trois panels. En arrière-plan, il s'agit de répondre à une question d'ordre méthodologique : quelles informations peut-on extraire en termes d'usages de données non qualifiées sur le plan du contenu ? Les URL visitées sont ici manipulées comme des données symboliques, regroupées en sites, sans indication de contenu.

Volumétrie globale du trafic des différents panels

De manière globale, nos trois sources de données représentent près de 45 années de navigation Web cumulée répartie en plus de 681 000 sessions. Ces éléments de volumétrie cumulée, quelque peu absurdes en eux-mêmes, ne témoignent que de la taille importante des données à traiter. Ils montrent ici la nécessité de se doter d'outils d'analyse permettant de manipuler des données en masse, à travers l'élaboration d'agrégats et d'indicateurs, tout autant que d'aller dans le détail de chaque session.

Tableau 5.13. Trafic des trois panels : volumétrie générale

	BibUsages	SN00-02	SN2002
Nombre d'utilisateurs	72	597	3372
Nombre de sessions	17 083	261 634	403 129
Durée d'observation	6 mois	34 mois	10 mois
Nb moyen de sessions par mois et par util.	39,5	12,7	11,9
Nb pages vues (en millions)	1,41	15,66	26,72
Nb pages distinctes (en millions)	0,65	4,60	6,75

Les volumes globaux par jeu de donnée montrent des disparités importantes (voir Tableau 5.13) : pour le plus important, SN2002, près de 27 millions de pages sont vues au cours de 403 000 sessions Web, contre 1,4 millions en 17 000 sessions pour BibUsages. Ces éléments ne doivent pas masquer les véritables disparités entre échantillons en termes d'usages : si le trafic global de BibUsages est bien moindre que celui des panels SensNet, du fait du nombre de participants et de la durée d'observation, les utilisateurs de BibUsages sont plus intensifs que les autres. Ainsi, chaque panéliste de BibUsages réalise en moyenne 39,5 sessions par mois (médiane : 23,2), tandis que ceux de SN2002 en font en moyenne 11,9 (médiane : 5,7). Les utilisateurs du panel 2000-2002, dont nous avons vu qu'ils étaient majoritairement des « anciens internautes », sont plus proches des seconds que des premiers, avec 12,7 sessions par mois en moyenne pour chacun (médiane : 8,7).

À l'examen de ce premier résultat, il semblerait bien que l'ancienneté ne soit pas un facteur déterminant dans l'intensité de la pratique, mais que les centres d'intérêt – en l'occurrence, l'attrait pour les bibliothèques électroniques – soit un moteur d'intensité d'utilisation. Ceci est sûrement en partie vrai, mais doit être tempéré par un autre élément : nos trois panels sont résidentiels, mais dans le cas de BibUsages, nous avons moins d'actifs, sinon beaucoup d'actifs exerçant une partie de leur activité à domicile.

Un examen longitudinal permet de donner plus de profondeur à l'analyse de l'activité des panels : la pratique du Web tend-elle à s'intensifier au fil du temps, ou se stabilise-t-elle ? Deux hypothèses opposées sont possibles : d'une part, on peut supposer que l'accroissement et la diversification de l'offre de contenus et de services sur le Web amènent les internautes à intensifier leurs pratiques ; d'un autre côté, on peut postuler l'ancrage des pratiques autour de certains sites amenant une stabilisation globale des usages au fil du temps. La répartition de l'activité des trois panels tout au long des périodes d'observation montre des comportements assez différenciés, qui tiennent plus à la composition et au maintien des panels. Dans le cas de BibUsages, on constate, malgré une activité relativement soutenue de juin à décembre, une baisse constante du nombre de panélistes actifs par mois, qui passe de 65 au début de la période à 40 en décembre 2002. Ceci est principalement dû au fait que certains participants à l'expérimentation ont désinstallé la sonde Audinet et quitté le panel sans que nous en soyons informés.

Une baisse similaire de l'activité, quoique moindre, est observable pour les données issues du panel SensNet 2002 : le nombre de panélistes actifs par mois décroît doucement (2 500 en janvier, 2 200 en octobre), de même que le nombre de sessions par mois. Cependant, les utilisateurs quittant le panel ont été filtrés dans ces

données, et la baisse d'activité est ici imputable aux « abandonnistes », des utilisateurs dont l'usage d'Internet s'affaiblit jusqu'à cesser. Dans le cadre du projet TypWeb[1], portant sur 1 140 internautes observés tout au long de l'année 2000, ce phénomène avait déjà été observé et décrit : « Pour une fraction des internautes, les usages sont rares et tendent à se raréfier, voire à disparaître au fil des mois ; pour les autres au contraire, surtout pour les gros utilisateurs, la pratique tend à s'intensifier »[2].

Le panel longitudinal se comporte de manière bien différente : en retenant des internautes ayant eu une activité en 2000 et en 2002, nous sommes assurés de n'observer que des individus qui ont intégré l'usage d'Internet dans leurs pratiques. Pas d'abandonnistes dans ce panel, donc, ce que traduit le nombre relativement stable d'actifs dans le panel mois par mois (voir Figure 5.2). Les variations sont ici imputables aux rythmes saisonniers, avec en particulier des baisses d'activité en février et surtout en juillet-août. Le nombre de sessions est particulièrement influencé par ces variations saisonnières[3], mais il connaît une croissance globale nette sur l'ensemble des 34 mois d'observation. On voit ici une confirmation de l'hypothèse formulée dans [Beaudouin *et al.* 2002] : pour les internautes fidèles, l'intensification de la pratique semble globalement se poursuivre sur trois ans.

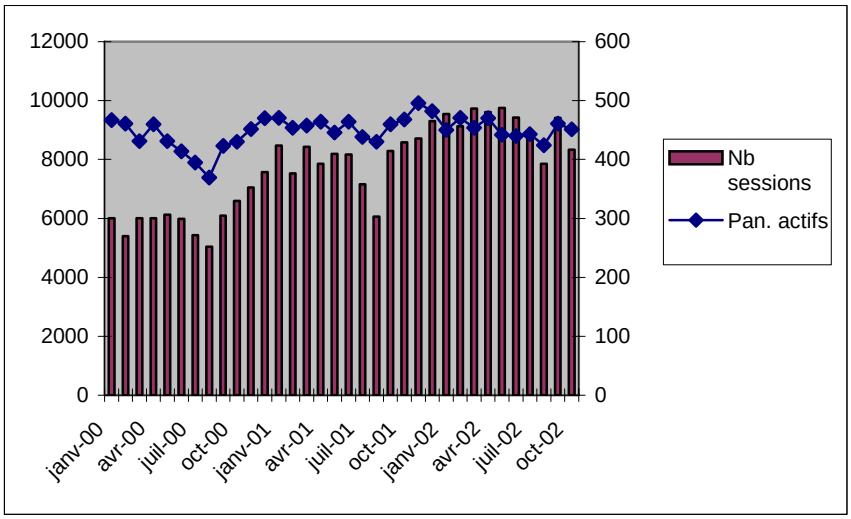


Figure 5.2. Activité du panel SN00-02

Comme nous avons pu le voir dans l'examen des usages globaux d'Internet, cette intensification globale n'est pas partagée par tous les individus. Pour quantifier ce phénomène, nous avons réalisé une régression linéaire du nombre et de la durée

---

[1] Voir Annexe 1 pour une description du projet TypWeb.

[2] [Beaudouin *et al.* 2002], p. 23.

[3] Il suffit, évidemment, d'une connexion dans le mois pour qu'un panéliste soit considéré comme actif, mais celui-ci peut être absent de son domicile pendant le reste du mois.

d'activité Web par mois pour chaque individu sur les 34 mois d'observation (voir Tableau 5.14).

Tableau 5.14. Pentes des régressions linéaires du nombre et de la durée des sessions par mois, SN00-02

		Nombre de sessions	Durée de sessions
Moyenne		0,206	15 min. 34 s.
Médiane		0,089	1 min. 22 s.
Quartiles	25	- 0,079	- 1 min. 37 s.
	50	0,089	1 min. 22 s.
	75	0,370	9 min. 50 s.

Pour la cohorte SN00-02, on observe en moyenne une session Web de plus par mois tous les cinq mois, et une durée d'activité Web augmentant de près de 15 minutes par mois. Ces valeurs moyennes sont fortement influencées à la hausse par une faible part des individus dont les pratiques se sont très fortement intensifiées : l'examen des valeurs médianes et des quartiles montre que, dans l'ensemble, le nombre de sessions augmente très lentement, environ une session mensuelle supplémentaire par an pour 1 min. 20 de navigation en plus par mois. En définitive, l'intensification de la pratique va plutôt dans le sens d'un allongement des durées de session (entre 1'20 et 9'50 minutes par mois pour un quart du panel, plus de 10 minutes pour un quart plus actif encore) que dans celui d'une augmentation du nombre de sessions, lequel ne connaît pas d'évolution significative, voire diminue faiblement pour un quart du panel. Des deux hypothèses que l'on opposait initialement, celle de l'intensification de la pratique n'est finalement vérifiée que pour une minorité des internautes ; pour les autres, l'ancrage de l'usage d'Internet dans les pratiques a conduit à une stabilisation globale du temps dévolu à cette activité.

Rythmes d'activité

A une échelle d'analyse différente, l'examen des rythmes d'activité dans la semaine et dans la journée permet de voir comment l'activité de navigation s'insère dans le cadre global des activités domestiques des trois panels résidentiels.

L'intensité d'usage varie au cours de la journée et de la semaine pour l'ensemble des trois sources de données, et ce de manière différente pour chaque panel : elle correspond à des rythmes d'activité globaux des utilisateurs. Pour les deux panels généralistes, les pics d'activité ont lieu le lundi et le mercredi (voir Figure 5.3 pour le panel SN2002). La cohorte suivie sur 34 mois se distingue uniquement par une activité particulièrement intense également le dimanche, en nombre de sessions comme en nombre de panélistes, tandis que l'échantillon SN2002 connaît peu d'activité ce jour.

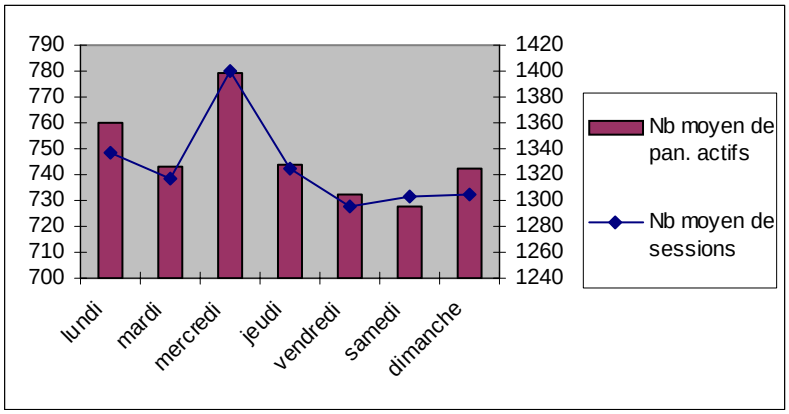


Figure 5.3. SN2002 – Activité par jour de la semaine

La répartition des heures d'activité montre une opposition bien plus nette entre semaine et week-end, relative au rythme de journées de travail. Pour les deux panels SensNet, les comportements sont similaires (voir Figure 5.4 pour les données SN00-02) : du lundi au vendredi, un pic d'activité très net se produit entre 18 et 23 heures, c'est-à-dire globalement à l'heure du retour au domicile. Le samedi et le dimanche, l'activité est bien plus répartie entre 10 heures et 22 heures, avec un creux d'activité vers 13-14 heures, et une hausse aux alentours de 17-18 heures, les soirées étant bien moins investies que le reste de la semaine.

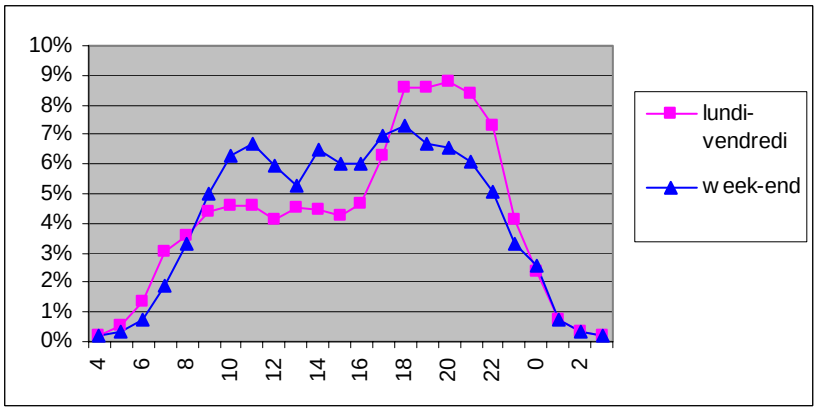


Figure 5.4. SN00-02 – Part des sessions par heure de début, semaine/week-end

Le profil du panel BibUsages est assez différent des deux autres : ce panel est mixte professionnel/résidentiel[1] et comporte un quart de non-actifs (retraités et étudiants en particulier), et son emploi du temps est structuré différemment de celui des données SensNet. L'activité Web reflète cette spécificité : l'activité hebdomadaire

[1] La sonde de recueil de trafic est dans certains cas installée sur le lieu de travail, et à l'inverse, des panélistes exercent leur activité professionnelle à domicile.

est répartie bien plus également entre les différentes journées. Le nombre de sessions est plus important du lundi au vendredi que le week-end, avec 30 à 32 panélistes actifs par jour contre 28 en moyenne. En ce qui concerne la répartition horaire, on ne constate aucune différence significative entre semaine et week-end : les sessions sont réalisées surtout entre huit heures le matin et 10 heures le soir, et leur nombre est assez stable entre ces deux horaires, malgré quelques pics d'activité à 11 heures, 14 heures et 20 heures.

Durée des sessions

L'activité de navigation s'insère donc fort naturellement dans la gestion globale du temps au quotidien et la disponibilité des utilisateurs. En revanche, elle n'est pas déterminée intrinsèquement par cette rythmique journalière : la durée des sessions est similaire pour les trois jeux de données, entre 30 et 35 minutes en moyenne (voir Tableau 5.15).

Tableau 5.15. Durées de sessions

	BibUsages	SN00-02	SN2002
Moyenne	30 min. 22 sec.	33 min. 57 sec.	35 min. 10 sec.
Médiane	13 min. 12 sec.	14 min. 21 sec.	14 min. 13 sec.
1er / 2e quartile	3 min. 7 sec.	4 min. 15 sec.	3 min. 48 sec.
2e / 3e quartile	13 min. 12 sec.	14 min. 21 sec.	14 min. 13 sec.
3e / 4e quartile	36 min. 57 sec.	35 min. 14 sec.	36 min. 43 sec.

Ainsi, malgré des activités de navigation de nature très différentes entre les trois panels (en particulier pour BibUsages), une séquence d'activité sur le Web est, dans sa durée, indépendante des éléments externes (individus, centres d'intérêt globaux) qui les composent, et la composante temporelle est bien plutôt liée à des éléments intrinsèques à la session. Pour les trois jeux de données, on trouve un premier quart des sessions très courtes, de moins de 3 à 4 minutes, tandis que la durée médiane se place autour d'un peu moins d'un quart d'heure ; lorsque la session s'allonge, la demi-heure est une limite dépassée dans un quart des cas uniquement, mais la durée peut alors être très importante, et atteint une heure en moyenne.

La gestion de leur « crédit-temps » par les utilisateurs est un de ces éléments : la grande diversité des sessions en termes de durée est liée à l'heure de début de session. Les quatre groupes de sessions – très courtes, courtes, longues, très longues – ne sont pas répartis de la même manière au fil de la journée (voir Figure 5.5 ci-dessous) ; nous travaillons ici sur les données SN2002, les plus représentatives et les plus récentes de nos trois jeux de données. On remarque que les sessions de plus de 35 minutes ne sont majoritaires qu'entre 14 heures et 17 heures, puis entre 20 et 22 heures, avec un pic particulier pour les sessions commençant entre 21 et 22 heures, et entre minuit et 1 heure du matin. Les autres tranches horaires laissent place aux sessions très courtes, mais surtout aux sessions à la durée comprise entre 4 minutes et un quart d'heure (2^{e} quartile).

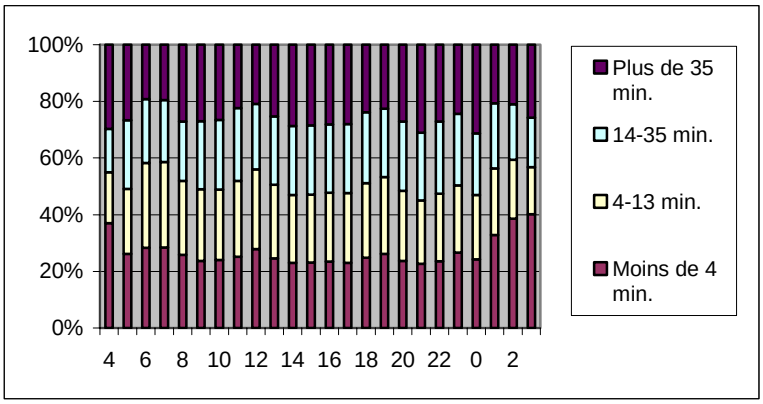


Figure 5.5. Durée des sessions et heure de début de session – SN2002, lundi-vendredi

Autre élément notable, on a constaté que les distributions sont très similaires entre le week-end et le reste de la semaine : alors que l'activité en nombre de sessions connaît des variations importantes entre ces deux moments de la semaine, la durée des sessions reste globalement très déterminée par l'heure de début de session quel que soit le jour de la semaine.

Il y a donc des périodes dans la journée où les sessions sont plus courtes, correspondant aux heures de retour du travail et d'avant-repas, dont on peut supposer qu'elles sont dévolues à certaines activités particulières sur le Web. Nous savons par ailleurs, par le biais d'études sur les modes de vie, que ces moments de la journée sont soumis à une nécessité particulière d'arbitrages entre un nombre important de sollicitations dans une période courte. À l'inverse, à partir de 22 heures, la disponibilité semble plus importante, et les sessions s'allongent. L'examen du contenu des sessions permettra par la suite de voir si ces sessions courtes sont liées à des tâches particulières ; contentons-nous pour l'heure de noter que la navigation sur le Web est une activité qui entre en concurrence avec d'autres au sein de la journée, et que sa pratique entre pour chaque individu dans l'économie générale du « crédit-temps ».

> ***Synthèse.*** *L'analyse, au sein des données volumineuses de trafic des trois panels, de la répartition de l'activité de navigation dans la semaine et dans la journée montre son inscription globale dans les pratiques dans et hors Web. Ce contexte global est fortement lié à la durée des sessions, et dénote l'implication d'activités différenciées au sein des sessions.*

5.2.2 Rythmes et formes de parcours

Nous avons jusqu'alors envisagé les sessions comme des blocs d'activité ; pour aller plus avant, nous entrons maintenant à l'intérieur des sessions, par le biais des indicateurs topologiques que nous avons élaborés (voir chapitre 4.2, « Analyser la séquentialité »). Nous introduisons ici le nombre de sites et de pages visités, le temps passé sur chaque page et chaque site, et l'enchaînement séquentiel des sites visités dans les sessions.

Sites visités

Le nombre de sites visités dans les sessions connaît des variations similaires à celles observées sur la durée, et les recoupe globalement. En moyenne, selon le jeu de données que nous manipulons, une session conduit à visiter entre 6 et 9 sites (voir Tableau 5.16), les sessions BibUsages se distinguant par un nombre plus important de sites que pour les deux autres échantillons. Mais comme pour la durée des sessions, la moyenne masque une importante diversité : la division en quartiles montre qu'un quart des sessions n'amène à visiter qu'un ou deux sites, tandis que seule la moitié des sessions entraîne la visite de plus de cinq sites. Pour le panel BibUsages, la moyenne de sites visités dans une session est plus importante du fait d'un plus grand nombre de sessions avec beaucoup de sites, les limites entre 1er / 2e et 2e / 3e quartiles étant similaires pour les trois jeux de données.

Tableau 5.16. Nombre de sites/portails visités par session

	BibUsages	SN00-02	SN2002
Moyenne	8,35	6,94	6,53
Médiane	4	4	4
1er / 2e quartile	2	2	2
2e / 3e quartile	4	4	4
3e / 4e quartile	9	8	7

Cette variation du nombre de sites visités dans chaque session est liée à celle déjà observée pour leur durée dans le cadre du rythme hebdomadaire. Le nombre de sites visités dans chaque session par période de la semaine ne connaît pas de différence notable entre le week-end et les autres jours. En revanche, la confrontation du nombre de sites visités et des rythmes d'activité journaliers déjà observés pour la durée des sessions est plus productive : pour les données SN2002, du lundi au vendredi, le nombre de sessions de plus de cinq sites diminue dans les tranches 12 h – 13 h et 19 h – 20 h, tandis que les autres sessions restent constantes à la même heure. Inversement, entre 20 h et 21 h, les sessions avec peu de sites décroissent, tandis que les autres connaissent un pic. Le week-end, cette différence de rythme entre sessions contenant un à quatre sites et sessions de plus de cinq sites est également observable : pic vers midi et 19 heures pour les premières, vers 15 heures et 18 heures pour les secondes.

Pour les données BibUsages, dont nous avons vu qu'elles ne comportent pas de différence marquée de rythme d'activité entre week-end et reste de la semaine, on retrouve des éléments similaires, mais avec des seuils différents. Alors que le panel généraliste opposait les sessions autour du seuil de cinq sites visités, les comportements sont ici similaires jusqu'à dix sites. C'est au-delà de dix sites visités que les décalages journaliers sont observables, ce qui semble dénoter des différences de pratique. Nous avons déjà observé que pour les panélistes de BibUsages, la durée médiane des sessions est, heure par heure, globalement inférieure à celle observée pour les deux autres panels ; il apparaît ici que, replongées dans le cadre des pratiques quotidiennes et de leur rythme, une session BibUsages « rapide » effectuée à des heures où la navigation entre en concurrence forte avec d'autres activités, dure à la fois moins longtemps et amène à voir plus de sites, comme si la navigation était accélérée, plus ciblée.

Durée de session et nombre de sites visités dans la session sont étroitement liés, c'est une évidence ; pour autant, la précédente comparaison entre données BibUsages et SN2002 invite à examiner plus finement cette corrélation. Si dans un temps donné, il semble bien difficile de voir plus d'un certain nombre de sites, en particulier pour les sessions courtes, il est intéressant de voir si ce lien est conservé pour l'ensemble des sessions, en croisant durées des sessions et nombre de sites visités (voir Tableau 5.17).

Tableau 5.17. Répartition des durées de sessions par nombre de sites visités

		1-2 sites	3-4 sites	5-10 sites	Plus de 10 sites
BibUsages	0-3 min.	67,2 %	27,8 %	7,4 %	0,6 %
	4-14 min.	18,5 %	37,6 %	30,3 %	7,9 %
	15-34 min.	9,1 %	22,6 %	33,4 %	26,3 %
	35 min. et plus	5,3 %	12,0 %	28,9 %	65,2 %
SN2002	0-3 min.	57,2 %	19,1 %	3,1 %	0,2 %
	4-14 min.	24,9 %	39,7 %	24,5 %	4,2 %
	15-34 min.	12,2 %	26,8 %	37,4 %	22,5 %
	35 min. et plus	5,6 %	14,4 %	35,0 %	73,1 %
SN00-02	0-3 min.	58,8 %	18,6 %	2,9 %	0,3 %
	4-14 min.	25,5 %	40,2 %	25,1 %	4,8 %
	15-34 min.	11,2 %	27,6 %	40,0 %	26,2 %
	35 min. et plus	4,5 %	13,6 %	32,0 %	68,8 %

Clef de lecture : dans les données BibUsages, 67,2 % des sessions comprenant 1 ou 2 sites durent moins de 4 minutes, 18,5 % durent entre 4 et 14 minutes.

Les différences entre jeux de données semblent attester un important effet d'ancienneté de la pratique : plus les internautes sont anciens et « avertis », plus ils voient de sites dans un temps donné. Les sessions d'un ou deux sites durent moins de 4 minutes dans 67 % des cas pour les données BibUsages, contre 57 % pour le panel généraliste SN2002. L'effet de rapidité est confirmé dans les sessions de plus de 10 sites : 65,2 % de ces sessions durent plus de 35 minutes pour BibUsages, contre 73,1 % pour SN2002. Dans tous les cas, le panel SN00-02 a une position intermédiaire.

Si l'on travaille à l'échelle de la page, on observe plus finement ce phénomène, ce que présente la Figure 5.6 (nous ne présentons pas le calcul pour les données SN00-02, similaire à celui pour SN2002). À mesure que le nombre de pages vues dans les sessions augmente, la durée des sessions croît régulièrement, avec un pic pour le dernier décile (valeurs extrêmes ou aberrantes). Parallèlement, pour BibUsages, le temps passé sur chaque page vue est relativement constant pour la première moitié des sessions, puis décroît de 34 à 20 secondes par page sur les 5 derniers déciles. L'effet « d'accélération » n'est ainsi perceptible que pour les sessions les plus longues, le rythme de visualisation des pages étant relativement constant pour les autres sessions.

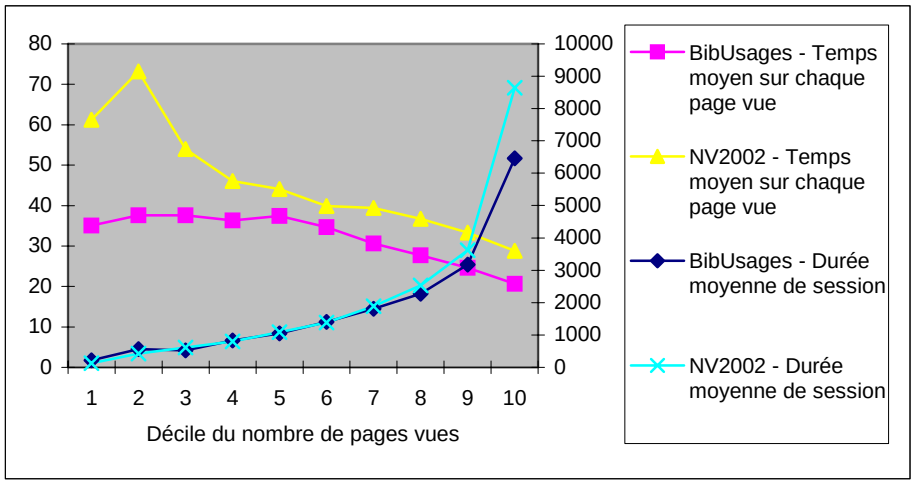


Figure 5.6. Nombre de pages vues et durées (BibUsages et SN2002)

Les sessions de SN2002 connaissent elles un effet d'accélération constant au fur et à mesure que l'utilisateur voit plus de pages dans la session. Il semble que les panélistes SensNet passent plus de temps pour appréhender le contenu des pages, possible reflet de l'effet d'apprentissage et d'expertise : il est probable que les utilisateurs de BibUsages, plus anciens et plus expérimentés, s'orientent plus facilement et plus rapidement à l'intérieur des sites et des pages Web.

Linéarité des parcours

L'examen des indicateurs topologiques permet d'aller plus loin dans l'analyse, en introduisant les notions de détour, de revisite de sites et de pages et de linéarité. Dans quelle mesure les sessions sont-elles linéaires, et si elles ne le sont pas, quelle est la portée des retours au sein des sessions, et quelles différences dans les modes de navigation cela dénote-t-il ? Comme nous l'avons présenté au Chapitre 4, nous distinguons deux types de linéarité : celle, globale, se situant au niveau de chaque requête adressée par le panéliste, dite « inter-pages », et une autre où les URL sont regroupées par site/portail, dite « inter-sites ».

Tableau 5.18. Part des sessions linéaires dans l'ensemble des sessions

	BibUsages	SN00-02	SN2002
Sessions inter-pages	24,5 %	17,1 %	18,6 %
Sessions inter-sites	31,3 %	35,3 %	36 %

Les deux modes de calcul apportent des résultats différents selon les jeux de données (voir Tableau 5.18). Pour le panel BibUsages, les sessions linéaires représentent environ un quart de l'ensemble des sessions dans les deux cas ; pour les deux autres panels, l'écart est important entre les linéarités inter-pages, qui concerne près de 18 % des sessions, et inter-sites, représentant près d'un tiers des sessions. Ainsi, si les retours sur des sites déjà vus au sein d'un parcours touchent globalement un tiers des sessions, à l'intérieur des sites, il semble que les utilisateurs de BibUsages aient des trajectoires sensiblement plus rectilignes que les autres.

Ici encore, on peut y lire un effet d'ancienneté de la pratique : le panel SN2002, où la proportion de nouveaux et récents internautes est la plus forte, est celui dont les sessions inter-pages sont les moins directes. En revanche, l'intensité de la pratique n'est pas liée à ce phénomène : nous avons examiné dans les données SN2002 la part des sessions linéaires en fonction du nombre total de sessions réalisées par panéliste sur les 10 mois d'observation. Au terme de ce calcul, aucune différence notable entre les d'utilisateurs : dans l'ensemble, pour chaque panéliste, la part des sessions linéaires inter-pages oscille entre autour de 29 %, celle des sessions linéaires inter-sites entre 31 et 36 %.

On ne s'étonnera pas de ce que les sessions linéaires sont plus courtes et comportent moins de sites que les autres : plus on voit de sites dans une session, plus celle-ci s'allonge, et plus la probabilité de revenir sur un site est importante. Ceci se vérifie pour les trois jeux de données que ce soit avec les sessions inter-pages (voir Tableau 5.19) ou les sessions inter-sites (voir Tableau 5.20).

Tableau 5.19. Sessions inter-pages linéaires et non linéaires (moyennes et médianes)

		Nb URL distinctes	Nb sites distincts	Durée totale	Temps par page vue
Sessions linéaires	BibUsages	7,1 (5)	2,6 (2)	2'43 (1'05)	32 (19)
	SN00-02	5,5 (3)	2,2 (2)	4'34 (1'05)	51 (29)
	SN2002	5,4 (3)	2,1 (1)	4'45 (1'09)	52 (30)
Sessions non linéaires	BibUsages	68 (36)	11,8 (7)	39'01 (21'23)	32 (21)
	SN00-02	40,4 (25)	8,8 (6)	39'40 (18'52)	42 (24)
	SN2002	42,6 (25)	8,5 (5)	41'19 (19'07)	43 (24)

De page en page, les temps moyen et médian passés sur chaque page sont similaires dans les sessions linéaires et non linéaires pour les sessions BibUsages, tandis que la revisite de pages amènent les panélistes NetValue à passer moins de temps sur chaque page et, en quelque sorte, à accélérer le rythme de la navigation.

Tableau 5.20. Sessions inter-sites linéaires et non linéaires (moyennes et médianes)

		Nb de sites distincts	Nombre de sites vus	Durée totale	Temps par site vu
Sessions linéaires	BibUsages	1,8 (1)	1,8 (1)	7'37 (1'44)	4'34 (1'07)
	SN00-02	2 (2)	2 (2)	8'48 (3'05)	4'57 (1'34)
	SN2002	1,9 (1)	1,9 (1)	9'22 (2'45)	5'40 (1'39)
Sessions non linéaires	BibUsages	11,3 (7)	35 (16)	40'44 (22'42)	2'01 (1'07)
	SN00-02	9,6 (6)	25,4 (13)	47'35 (24'34)	2'49 (1'30)
	SN2002	9,1 (6)	28,4 (13)	49'42 (25'00)	2'48 (1'33)

Ce phénomène est également observable à l'échelle des sites (voir Tableau 5.20) : dans les sessions non linéaires, le nombre de sites différents visités est sensiblement plus important (médianes à 7 ou 8, contre 1 ou 2 pour les parcours directs). On notera que les valeurs médianes du temps passé sur chaque visite de site sont systématiquement inférieures aux valeurs moyennes, mais surtout que l'écart entre moyenne et médiane est bien plus important pour les sessions linéaires : dans ce cas, quelques séquences de très longue durée influencent le calcul de la moyenne, valeurs

extrêmes bien moins présentes dans les sessions non linéaires[1]. Enfin, ici encore, le trafic BibUsages se démarque des deux autres par des durées plus faibles des séquences sur les sites et des sessions en général.

Ces éléments confirment la triple corrélation : nombre de sites différents / durée / linéarité. Nous avons déjà observé une relation quasi-linéaire entre durée de sessions et nombre de sites distincts vus dans la session ; l'examen de la proportion de sessions linéaires par décile de la durée de la session et par décile du nombre de sites visités montre une relation tout autre. À l'échelle de la page, plus la session dure, plus la part des sessions linéaires diminue, de manière proportionnelle ; à l'échelle du site, la distribution n'est pas la même, et prend une allure plus zipfienne.

Les résultats sont très similaires pour les trois échantillons, et restent stables si l'on croise linéarité et nombre de sites visités. Dans tous les cas, les sessions courtes, comportant peu de sites (s'il n'y en a qu'un, tout est joué d'avance) sont linéaires à 80 % en inter-pages et 92 % en inter-sites ; ces proportions décroissent très vite à mesure que la session s'allonge, et au septième décile, 3 % sont linéaires en inter-pages, et 12 % en inter-sites. Ainsi, au-delà de quatre sites différents visités dans une session et de dix minutes de navigation, les deux tiers des sessions comportent au moins un retour sur un site déjà visité dans la session, et 90 % amènent à voir une même page au moins deux fois.

> ***Synthèse.*** *La diversité des sessions en termes de durée se retrouve dans le nombre de pages et de sites différents visités dans la session : à mesure que le nombre de pages vues dans les sessions augmente, leur durée croît régulièrement. Ces éléments sont corrélés à la linéarité des parcours : dès que la session s'allonge, la ligne brisée se généralise, à l'échelle de la page comme du site. Une première opposition se fait jour entre des sessions courtes et linéaires, qui regroupent un tiers des sessions à l'échelle du site, et semblent plus ciblées, et des sessions longues et non-linéaires reflétant des comportement plus diversifiés.*

Quantifier et qualifier les revisites à l'échelle de la page

Les sessions non linéaires sont donc relativement fréquentes dans nos corpus de sessions, et font partie intégrante des modes de navigation sur le Web puisqu'elles concernent entre 75 et 80 % des sessions. Toutefois, l'opposition linéaire/non linéaire masque une grande diversité au sein des sessions non linéaires, qu'il importe d'examiner plus en détail. Entre le simple retour sur une page dans une longue session quasi-rectiligne et la visite massive et répétée d'une même page dans une navigation en étoile, par exemple la page de réponse d'un moteur de recherche, ce sont différents modes de navigation qui sont en jeux. L'examen des taux de linéarité calculés pour chaque session permet d'apprécier cette diversité, en donnant une

---

[1] Il n'est pas impossible que ces très longues séquences linéaires ne soient pas la trace d'une navigation volontaire et supervisée de la part des internautes, mais plutôt de requêtes adressées automatiquement (rafraîchissement automatique de pages, bandeaux publicitaires alternés, etc.).

estimation de la part de la session consacrée à des pages ou des sites déjà vus, en nombre de pages et de sites ainsi qu'en durée.

Tableau 5.21. Valeurs des taux de linéarité r_{page} *et* r_{site} *par corpus (sessions non linéaires)*

	Sessions inter-pages non linéaires		Sessions inter-sites non linéaires	
	Moyenne	Médiane	Moyenne	Médiane
BibUsages	0,69	0,75	0,47	0,47
SN00-02	0,63	0,66	0,50	0,50
SN2002	0,63	0,67	0,48	0,50

Les valeurs moyennes des taux de linéarité r_{page} et r_{site} pour les sessions non linéaires sont relativement comparables pour les trois jeux de données : en moyenne, la linéarité à l'échelle de la page est de l'ordre de 0,65, et de 0,5 à l'échelle du site (voir Tableau 5.21). La distribution des valeurs de r_{page} montre une concentration autour de la moyenne (proche de la médiane)[1]. En d'autres termes, dès lors que l'utilisateur revisite des pages, un tiers des pas de la session correspond à des pages vues plusieurs fois en moyenne, et ce pour les trois jeux de données. Tout au plus notera-t-on que pour les sessions du panel BibUsages, la linéarité est globalement plus importante au niveau de la page, mais l'est moins au niveau du site.

Le taux de linéarité calculé sur la base de la durée de visite des pages, d_{page}, rend compte du temps passé sur les pages vues une fois dans la session : il permet d'analyser plus finement la revisite en la quantifiant sur une échelle de temps. L'indicateur montre qu'en moyenne la moitié du temps de la session concerne des revisites, avec des taux avoisinant 0,5 (voir Tableau 5.22).

Tableau 5.22. Valeurs de d_{page} *pour les sessions inter-pages non linéaires*

		BibUsages	SN00-02	SN2002
Moyenne		0,59	0,53	0,53
Médiane		0,65	0,55	0,55
Quartiles	25	0,40	0,32	0,32
	50	0,65	0,55	0,55
	75	0,82	0,76	0,76

Clef de lecture : en moyenne, pour le panel BibUsages, 59 % de la durée d'une session non linéaire sont consacrés à des pages vues une seule fois dans la session.

La distribution des valeurs du taux d_{page} calculé sur la durée est beaucoup plus uniforme que celle observée avec le calcul basé sur les éléments visités (voir Figure 5.7 ci-dessous) : le pic autour de la moyenne est moins sensible, et une continuité forte s'établit entre sessions linéaires et sessions non linéaires.

---

[1] La distribution du taux de répétition moyen r_{page}, pour les sessions non linéaires montre un pic autour de la valeur 0,75, et un déficit de valeurs entre 0,75 et 1 (sessions linéaires). Nous pensons que cette distribution ne correspond pas tant à une rupture entre sessions linéaires et sessions non linéaires qu'à un biais issu du mode de calcul de l'indicateur : en effet, le nombre important de sessions dont le nombre de pages vues est de trois, quatre ou un multiple faible de ces valeurs (8, 9, 12), favorise des ratios de 0,66 ou 0,75 lorsqu'une page est revue sur trois par exemple.

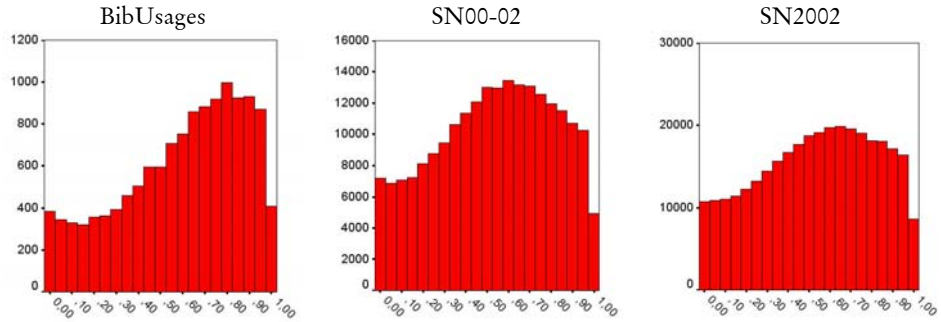


Figure 5.7. Fréquences des valeurs de d_{page} *pour les sessions inter-pages non linéaires*

On observe une différence sensible entre les sessions BibUsages et les autres : le taux de linéarité est globalement plus élevé, ce qui signifie que les utilisateurs de ce panel passent moins de temps sur les pages vues plus d'une fois. Ce constat, qui rejoint celui d'un rythme plus élevé dans la navigation, tend à montrer que ces internautes sont plus sélectifs et plus rythmés dans leurs visites. Cet effet est confirmé par le profil intermédiaire des données SN00-02 entre BibUsages et SN2002, panel intermédiaire en termes d'ancienneté et d'expérience.

Comment s'organisent structurellement les revisites à l'échelle de la page ? Touchent-elles certaines pages en particulier, ou s'appliquent-elles à l'ensemble des pages vues plusieurs fois dans la session ? Le taux de concentration c_{page} qui rend compte de ces phénomènes est égal à 1 dans un quart des cas (chaque page revisitée est revue une seule fois), et reste globalement compris entre 1 et 2. Toutefois, les taux de concentrations élevés pour le dernier quart montrent un écart de comportement entre des sessions avec pages-pivots (concentration forte) et d'autres où la revisite est plus étalée sur les différentes pages vues (voir Tableau 5.23).

Tableau 5.23. Taux de concentration c_{page}

		BibUsages	SN00-02	SN2002
Moyenne		3,35	2,34	2,35
Médiane		1,65	1,57	1,61
Quartiles	25	1	1	1
	50	1,65	1,57	1,61
	75	2,54	2,23	2,36

Clef de lecture : pour le panel BibUsages, une page revisitée est revue en moyenne 3,35 fois dans la session.

Cet écart entre deux profils structurels de revisite est fortement lié au taux de linéarité : moins les sessions sont linéaires, plus les revisites se concentrent sur une ou plusieurs pages. Nous pouvons supposer que cet écart entre sessions peu linéaires, avec un fort taux de concentration, et sessions très linéaires est le reflet de stratégies et de configurations de navigation différentes : soit l'on est dans une configuration de « prédateur » où l'utilisateur sait précisément où il va et agit rapidement, soit l'on est dans une forme de parcours de type « découverte » ou « multi-tâche », où

certaines pages servent de pivot à la revisite ou au passage à une tâche différente (pages de résultats de moteurs de recherche, page d'accueil d'un site, page de démarrage, etc.).

Pour appuyer cette hypothèse, on examinera la présence des actions de type *back* dans les sessions non linéaires : le recours à cette fonctionnalité est en effet particulièrement fréquent dans des configurations avec pages-pivot où l'utilisateur explore des liens à partir d'une page donnée.

Dans les deux jeux de données SensNet, les comportements sont similaires : au sein des sessions non linéaires, 85 % comportent au moins une séquence de type *back*, quelle qu'en soit la longueur. Pour les données BibUsages, seules 70 % des sessions non linéaires sont concernées. Dans les trois cas, la longueur des séquences de ce type est majoritairement de 2 pas (75 % des cas), ce qui signifie que les trois quarts des mouvements de retour arrière ne concerne qu'une page et une seule, sous forme d'aller-retour sans profondeur.

La part des pages vues au sein d'actions de type *back* dans l'ensemble des parcours reste elle-même souvent modeste : l'indicateur b_{page} qui en rend compte vaut en moyenne 0,18 (médiane : 0,11) pour les sessions BibUsages, et 0,24 pour les sessions SensNet (médiane : 0,2). Ceci signifie que les pages au sein d'actions *back* occupent en moyenne entre 20 et 25 % des pages vues dans une session non linéaire. Les distributions des valeurs de l'indicateur renforcent ce constat, et montrent une fois de plus la spécificité du panel BibUsages, pour lequel les mouvements de *back* sont non seulement moins fréquents, mais occupent moins de place dans les parcours.

Ainsi, de manière générale, la ligne brisée est loin d'être une constante dans la navigation lorsqu'on l'examine au niveau de la page. Si les sessions non linéaires s'opposent aux sessions linéaires en termes de durées et de nombre de pages vues, elles forment un groupe globalement homogène où les revisites sont plutôt de d'ordre du « pas de côté » et du petit détour. Seule une minorité de sessions donne lieu à des comportement plus particuliers, avec des pages-pivot, des revisites intensives et de nombreux *back*, relevant de contextes de navigation spécifiques que l'analyse des contenus permettra d'éclairer.

Revisites de site en site

Les comportements de revisite examinés à l'échelle du site viennent confirmer ces hypothèses et les complètent. Nous avions vu qu'à l'échelle du site, les sessions sont plus linéaires qu'à celle de la page, avec un tiers de sessions linéaires en inter-sites contre 18 à 25 % en inter-pages. Par contre, au sein des sessions non linéaires, le taux de linéarité r_{site} est plus faible qu'à l'échelle de la page, avec une moyenne autour de 0,5 pour les trois jeux de données, identique à la médiane. La distribution des valeurs de r_{site} (Figure 5.8 ci-dessous) montre en effet une concentration du taux de linéarité autour de 0,5 ainsi qu'une rupture forte avec les sessions linéaires ($r_{site} = 1$).

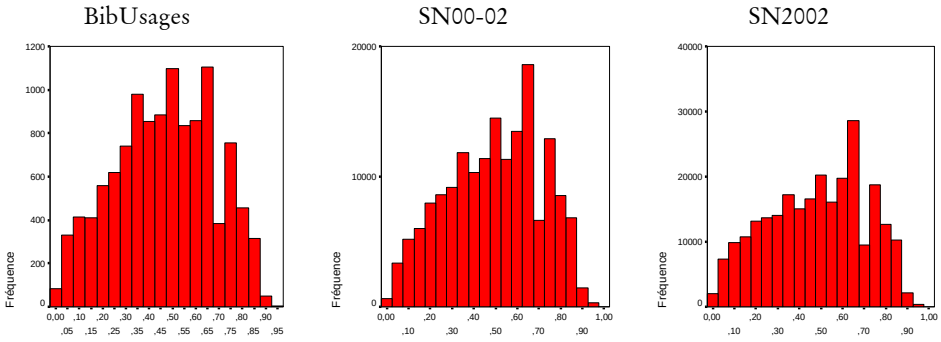


Figure 5.8. Fréquences des valeurs de r_{site} *pour les sessions inter-site non linéaires*

Le même indicateur calculé sur les durées de visite confirme l'importance des sites vus plusieurs fois dans les sessions : avec une valeur moyenne proche de 0,3 (médiane autour de 0,2), les sites vus plusieurs fois occupent en moyenne plus des deux tiers de la durée des sessions (voir Tableau 5.24).

Tableau 5.24. Valeurs de d_{site} *pour les sessions inter-sites non linéaires*

		BibUsages	SN00-02	SN2002
Moyenne		0,29	0,32	0,30
Médiane		0,19	0,24	0,21
Quartiles	25	0,04	0,06	0,04
	50	0,19	0,24	0,21
	75	0,47	0,53	0,52

La distribution des valeurs de d_{site} calculé sur la durée de visite confirme cette tendance, avec un contingent majoritaire de sessions où les sites vus une seule fois dans le parcours comptent pour moins de 10 % du temps total de la session. Le détail de l'activité sur ces sites vus plus d'une fois montre que les retours ne sont pas un simple passage rapide, mais correspondent au contraire à la majorité du temps passé sur le site revu. En effet, lorsqu'un site est vu plusieurs fois dans une session, le premier passage sur le site occupe en moyenne entre 26 à 30 % du temps total sur ce site selon le jeu de données considéré (médianes entre 18 et 24 %), avec une distribution orientée vers les faibles valeurs.

Revoir un site ne semble donc pas dû au hasard ni aux nécessités du cheminement dans les hyperliens, mais apparaît bien plus comme le signe d'un intérêt particulier pour les contenus proposés. Toutefois, revisiter un site n'amène globalement pas à le revoir plus de deux fois : le taux de concentration c_{site} reste en-deçà de 2 pour la moitié des sessions non linéaires à l'échelle du site, et ne dépasse 3,5 que pour un quart de ces sessions (voir Tableau 5.25).

Tableau 5.25. Taux de concentration c_{site}

		BibUsages	SN00-02	SN2002
Moyenne		3,22	2,81	3,31
Médiane		2	2	2
Minimum		1	1	1
Maximum		100	299	416
Quartiles	25	1,20	1	1
	50	2	2	2
	75	3,5	3,2	3,5

Les mouvements de retour arrière sont très fréquents dans la revisite de site : dans 95 à 96 % des sessions non linéaires, on observe des comportements de ce type. La part de ces retours arrière dans les parcours est bien plus importante qu'à l'échelle de la page : ils occupent en moyenne la moitié des pas de site en site dans les sessions non linéaires (voir Tableau 5.26), et la distribution des valeurs de b_{site} montre la concentration des valeurs autour cette moyenne.

Tableau 5.26. Valeurs de b_{site} *(BibUsages, SN00-02, SN2002)*

		BibUsages	SN00-02	SN2002
Moyenne		0,54	0,54	0,56
Médiane		0,53	0,54	0,56
Quartiles	25	0,40	0,40	0,40
	50	0,53	0,54	0,56
	75	0,67	0,67	0,71

Si l'on ôte à l'échelle du site les mouvements de *back* des sessions non linéaires, les sessions apparaissent comme linéaire dans 39,8 % des cas pour BibUsages, 43,9 % des cas pour SN00-02 et 46,6 % des cas pour SN2002. Cela signifie que dans plus de la moitié des cas, la non-linéarité est due à un mouvement de retour vers un ou plusieurs sites vus précédemment. On est ainsi en droit de supposer que certains sites servent de pivot à la navigation, et que les parcours s'articulent autour de ce site.

> ***Synthèse.*** *Le groupe des sessions non linéaires masque une grande diversité dans la structure des revisites. En premier lieu, à l'échelle de la page, les revisites sont plutôt de d'ordre du « pas de côté » et du petit détour, mais plus la session s'allonge et se complexifie, plus on voit apparaître des phénomènes de page-pivot sur lesquelles se concentrent les revisites. Ces éléments sont encore plus marqués à l'échelle du site : revoir un site ne semble pas dû au hasard, mais apparaît comme le signe d'un intérêt particulier pour les contenus proposés. Ces phénomènes de pages ou sites-pivot semblent être le reflet de deux modes de navigation différents : le comportement « prédateur » limite les revisites et demeure plutôt rectiligne, avec quelques digressions, tandis que les parcours de type « découverte » ou « multi-tâche » s'appuient sur certaines pages ou certains sites pour déployer une navigation en étoile plus longue et plus complexe.*

Conclusion

L'examen des éléments temporels, rythmiques et topologiques des sessions montre qu'il n'existe pas une session-type, mais des types de sessions bien différenciés, aux profils topologiques spécifiques. La plupart des éléments topologiques et rythmiques sont liés : durée, nombre de pages et de sites et revisite entretiennent des relations étroites. Si les corrélations sont fortes entre les différents indicateurs que nous avons construits, elles ont des causes structurelles : l'hétérogénéité des sessions sur le plan du nombre d'unités qui les composent rend difficile leur comparaison, et biaise le calcul des indicateurs. Il n'est alors pas surprenant que la linéarité soit directement liée au nombre de pages ou de sites visités. L'analyse temporelle permet de nuancer ces constats et d'avoir une vue plus qualitative sur les revisites : elle montre notamment le temps important qu'occupent les pages vues plusieurs fois dans les sessions, et le fait qu'un deuxième passage n'est pas un survol, mais donne lieu à une véritable lecture. L'analyse en termes de durée met également à jour des éléments de rythmique, non observables par d'autres moyens : sur ce plan, le panel BibUsages s'oppose fortement aux autres, et s'impose comme un panel « expert » où les utilisateurs ciblent leur navigation et passent peu de temps en détours.

En définitive, les sessions linéaires forment un groupe distinct des autres, pour lequel les indicateurs topologiques ont des valeurs fixes, et semblent renvoyer à des contextes précis où l'internaute visite de manière ciblée certains sites dans une courte durée. Pour les autres sessions, le panorama est plutôt diversifié : certaines sessions comportent peu de détours et les à-côté sont globalement minoritaires, tandis que d'autres se démarquent par de faibles taux de linéarité et l'emploi plus massif des *back*, et semblent dénoter des structures de navigation particulières. Ces différents comportements de navigation correspondent à des contextes d'usage différents, et une première distinction se fait jour entre des parcours ciblés, plutôt linéaires, et des parcours plus ouverts et plus longs, dont la structure complexe est caractérisée par la présence de pages-pivot et de nombreux retour-arrière. Cette opposition recouvre, dans une certaine mesure, celles observées dans la répartition horaire de l'activité : l'activité Web s'inscrit dans le contexte plus général de l'activité journalière, et certains profils de sessions sont plus présents à certaines heures qu'à d'autres. Toutefois, ces éléments formels et temporels ne sauraient expliquer seuls ces différences de comportements dans les parcours ; l'analyse des contenus visités d'une part, et de l'inscription des visites dans le corpus et la navigation de chaque panéliste de l'autre, permettront de donner sens à ces formes de parcours en les insérant au sein des pratiques en contexte.

5.3 Contenus des parcours

Parallèlement à l'étude de la topologie des sessions, on souhaite avoir un panorama des contenus visités par les internautes, et éprouver ce faisant la solidité de nos descripteurs. Dans cette optique, nous évaluons dans un premier temps précisément la capacité des annuaires à décrire les parcours seuls, puis montrons que le recours complémentaire à *CatService* permet d'améliorer à la fois la couverture et

la qualité des descriptions. Nous proposons ainsi une première segmentation des parcours sur la base des contenus visités par les internautes.

5.3.1 Étendue des descriptions de contenu

La mobilisation des ressources exogènes que représentent les annuaires Web pour décrire le contenu des parcours des internautes est à la fois riche et complexe. Elle doit être validée en termes quantitatifs et qualitatifs, au regard des différents modes de structuration et de description des objets du Web qu'ils utilisent.

Choix des annuaires exploités

Nous avons déjà vu que le taux de couverture moyen du trafic global des panels représentatifs en 2000, 2001 et 2002 est de l'ordre d'un tiers pour chaque annuaire, ce qui nous autorise à tenter d'exploiter cette source de données pour décrire le contenu des parcours. Plus en détail, le Tableau 5.27 présente, pour chaque jeu de données, le taux de couverture moyen des sessions par les annuaires aspirés en 2002.

Tableau 5.27. Couverture moyenne des sessions par les annuaires 2002

	BibUsages		SN2002		SN00-02	
	Nb. URL	Durée	Nb. URL	Durée	Nb. URL	Durée
Looksmart	35,3 %	34,8 %	31,8 %	29,6 %	36,6 %	34,8 %
Lycos	22,3 %	22,3 %	21,0 %	21,1 %	26,2 %	25,8 %
MSN	33,7 %	33,2 %	29,7 %	28,0 %	35,8 %	34,4 %
Nomade	34,0 %	33,7 %	30,2 %	28,2 %	33,7 %	31,9 %
Open Dir.	24,4 %	23,3 %	20,7 %	18,5 %	25,8 %	23,8 %
Voila	35,9 %	35,9 %	31,7 %	29,9 %	34,1 %	32,5 %
Voila PP	1,4 %	1,3 %	1,1 %	1,2 %	1,7 %	1,9 %
Yahoo	30,4 %	29,6 %	27,5 %	25,8 %	33,1 %	31,4 %

Nous avions déjà observé que les annuaires s'adaptent à l'évolution du Web, mais qu'ils ne précèdent pas le trafic : les meilleurs taux de couverture des annuaires 2002 avec le trafic global des panels NetValue étaient obtenus en 2001. On ne s'étonnera pas, dans ces conditions, que les données SN00-02 soient mieux décrites par les annuaires que celles enregistrées en 2002 uniquement. Plus surprenants sont les taux de couverture pour le panel BibUsages, systématiquement supérieurs aux deux autres, alors que ces données sont de toutes les plus tardives. On peut supposer que ce résultat est le reflet du caractère très académique du panel, qui fréquente beaucoup de sites institutionnels ou renommés, lesquels sont bien mieux représentés dans les annuaires que les autres (nouveaux sites, sites perso, etc.).

Deuxième élément notable, les taux de couverture sont relativement semblables selon qu'on les calcule en nombre d'URL ou en durée ; tout au plus remarque-t-on un écart constant d'un à deux points en faveur de la couverture en nombre d'URL vues, ce qui tend à montrer que, globalement, les internautes passent moins de temps, au sein des sessions, sur les sites indexés par les annuaires que sur les autres.

Nous nous sommes attachés à montrer au Chapitre 3, lors de la description des données issues des annuaires, que les huit annuaires étudiés diffèrent profondément

en termes de pages et de sites indexés, dans la manière dont ils décrivent ces pages et ces sites, et surtout en ce qui concerne leurs structures. Nous avons conclu qu'il est impossible de mobiliser simultanément les huit sources, et qu'il est préférable de projeter chaque annuaire séparément sur les parcours. Cette approche a en outre l'avantage d'opérer comme un système de vérification des résultats, les conclusions obtenues avec chacun des annuaires devant être cohérentes entre elles, *modulo* leurs spécificités thématiques.

La description des parcours par les catégories d'annuaires nécessite cependant d'examiner de près chaque annuaire avant de l'exploiter. En effet, certains éléments structurels se montrent gênants, voire rédhibitoires : on souhaite, pour avoir des descriptions fiables et relativement tranchées, qu'une adresse ne soit pas indexée à de multiples endroits, et que les plans de classement respectent une cohérence thématique générale qui ne fasse pas apparaître de catégorie « fourre-tout ».

De ce point de vue, tous les annuaires ne sont pas éligibles pour décrire les parcours : Looksmart duplique des pans entiers de ses catégories pour faciliter la navigation, et une URL a de grandes chances de figurer sous plusieurs catégories de premier niveau. Yahoo marque géographiquement ou économiquement les sites indexés, et inscrit quasi-systématiquement chaque URL indexée sous les catégories « Exploration géographique » ou « Commerce et économie » en même temps que sous une des autres catégories de premier niveau. Voila recourt abondamment à la multi-indexation, avec un taux de répétition moyen des URL de 3,24, ce qui est susceptible de « bruiter » l'analyse. Les taux de couverture sont également à prendre en compte : Voila Pages Perso mis à part, Lycos et Open Directory sont les plus mal lotis, avec en moyenne des taux avoisinant les 23 %, alors que leurs concurrents sont dix points au-dessus.

Au terme de ce travail d'examen des sources, nous avons choisi de retenir trois annuaires : MSN France, Nomade et Yahoo France. Ces trois annuaires ont les meilleurs taux de couverture (de l'ordre de 32 % en moyenne, en durée de session), recourent peu à la multi-indexation des URL, et présentent des structures assez équilibrées entre leurs différentes catégories de premier niveau (voir Tableau 3.26, Tableau 3.27 et Tableau 3.31, p. 125). Enfin, ils répondent chacun à des logiques de classement différentes : localisation géographique pour Yahoo, angle économique pour MSN, catégorisation orientée « monde de l'Internet » pour Nomade. De ce point de vue, ces trois annuaires se complètent bien et permettent une validation croisée des résultats.

Toutes les sessions n'étant pas décrites complètement par les annuaires, nous avons pour chaque couple données-annuaire retenu les sessions décrites à plus de 50 % par l'annuaire considéré[1]. Comme le montre le Tableau 5.28, nous perdons globalement 70 à 75 % de notre corpus de sessions dans cette sélection.

---

[1] Ce taux de couverture est calculé sur la base de la durée passée sur des pages décrites par les annuaires par rapport à la durée de la session.

Tableau 5.28. Part de l'ensemble des sessions couvertes à plus de 50 % pour chaque couple annuaire-données

	BibUsages	SN00-02	SN2002
MSN	29,9 %	31,1 %	24,3 %
Nomade	31,2 %	28,4 %	24,4 %
Yahoo	26,5 %	28,2 %	22,4 %

Nous avons bien tenté de réduire cet écart : n'existe-t-il pas, pour chaque URL visitée non décrite par l'annuaire, une page ou un site similaire qui serait, lui, décrit par l'annuaire ? Pour tester cette hypothèse audacieuse, nous avons sélectionné un sous-échantillon test de 80 000 sites comportant des URL visitées non indexées et pour chacun d'eux, envoyé sur le moteur de recherche Google une requête du type « related », qui renvoie les sites et pages similaires pour une adresse donnée. Le résultat est bien pauvre : sur les 48 400 sites pour lesquels Google proposait un site similaire, seulement 10 260 étaient indexés dans les annuaires. Si l'on ajoute à ce faible gain les distorsions que cette méthode implique (comment connaître, dans chaque cas, le degré de similarité renvoyé ?), on comprendra que nous ayons finalement laissé de côté cette piste.

Descriptions combinées

En exploitant séparément Nomade, MSN et Yahoo, sommes-nous pour autant tirés d'affaire, et disposons-nous de données exploitables ? Nous avons tenté une première description des sessions Web couvertes à plus de 50 % par les catégories de premier niveau des trois annuaires retenus. Pour chaque page décrite par l'annuaire, nous avons noté la ou les catégories de premier niveau correspondantes ; nous regardons ensuite dans la session le temps passé sur chaque catégorie, et retenons la catégorie la plus représentée dans la session. Pour les trois annuaires, le résultat s'avère très bruité, et presque écrasé par la part très importante des portails généralistes, qui font souvent office de page de démarrage en tant que fournisseurs d'accès, et drainent une forte audience en général du fait de la diversité des contenus et services qu'ils proposent. Pour MSN, c'est la catégorie « Informatique – Internet » qui ressort loin devant toutes les autres (voir Tableau 5.29 ci-dessous) ; dans le cas de Nomade, « Espace B to B » prend le dessus (38,2 % des sessions) ; et pour Yahoo, c'est la catégorie « Commerce et économie » qui écrase les autres (64 %).

Tableau 5.29. SN2002 : répartition des sessions par catégorie MSN la plus forte dans la session

Catégorie	Part des sessions
Informatique – Internet	40,2 %
Finances - Bourse – Patrimoine	9,3 %
Jeux – Consoles	7,3 %
Entreprises	7,2 %
Infos – Météo	6,6 %
Arts - Culture – Média	6,4 %
Loisirs - Passions	6,4 %
Vie quotidienne - Société	4,9 %
Shopping	3,4 %
Sports	2,2 %
Savoir - Education	2 %
Voyages - Tourisme	1,4 %
Emploi, formation	1,3 %
Santé	0,9 %
Sciences - Techniques	0,5 %

Clef de lecture : dans 6,6 % des sessions, la catégorie de premier niveau de MSN « Infos – Météo » est celle sur laquelle l'internaute a passé le plus de temps dans la session.

Il est apparu indispensable de « casser » cette catégorie dominante qui regroupe des contenus et des services très diversifiés, et de pouvoir descendre plus finement dans l'analyse des parcours sur les portails généralistes. Pour cela, nous avons mobilisé les informations fournies par *CatService*, qui décrit précisément les différents services utilisés sur ces sites. Pour chaque page vue dans une session, nous avons construit une description basée sur les types de portails et, pour les portails généralistes, les types de services, suivant les règles suivantes :

1. si le site est identifié par *CatService*,
    a. si c'est un site personnel hébergé gratuitement par un fournisseur d'accès ou un portail (Wanadoo, Free, Yahoo, etc.), on ne retient pas la description *CatService* et on passe au cas n°2 ;
    b. si la page correspond à un service de moteur de recherche, de WebMail ou de *chat*, elle est identifiée comme telle, même si le service est fourni par un portail généraliste ;
    c. si la page se situe sur un portail généraliste, on retient la description au niveau du service utilisé : « portail généraliste – informations », « portail généraliste – page d'accueil », etc.
2. si le site n'est pas identifié par *CatService* ou si c'est un site personnel, on cherche dans l'annuaire si l'URL visitée correspond à une URL indexée, et l'on retient les catégories de premier niveau de l'annuaire qui décrivent cette URL.

L'utilisation de *CatService* apporte en outre une solution à un problème de couverture qui touche essentiellement les portails généralistes. Notre méthode de projection des annuaires sur les parcours est basée sur un appariement au niveau des URL, elle implique que l'URL de l'annuaire soit un sous-ensemble de celle visitée. Les annuaires indexant la plupart du temps des points d'entrée des sites, nous manquons les sites réparties sur plusieurs sous-domaines : par exemple, si l'adresse du site de Free, `http://www.free.fr` est indexée par un annuaire, nous manquerons dans les

parcours la visite des pages sur l'offre ADSL de Free, dont les adresses commencent par `http://adsl.free.fr`, ou encore le WebMail de Free, sous `http://imp.free.fr` et quelques autres variantes. Ce biais technique s'avère d'autant plus gênant que ce sont justement les sites de taille importante, au premier rang desquels les grands portails généralistes, qui recourent à ce dispositif technique. Le recours à *CatService*, qui couvre très bien ce problème de variation du *host*, permet de pallier ce problème.

Enfin, les annuaires laissent dans l'ombre des sites qui drainent de fortes audiences mais apparaissent peut-être comme moins « présentables » : sites pornographiques et sites tournant autour de la galaxie du piratage. De la même manière qu'à la grande époque du Minitel, les services de messagerie, de dialogue et de rencontre « roses » drainaient la majorité de l'audience, le Web « rose » fait partie intégrante des pratiques sans avoir de représentation correspondante dans les services de recherche généralistes grand public que constituent les annuaires Web ici utilisés.

Pour ne pas laisser de côté les sites dits « adultes », nous avons mobilisé une information fournie par NetValue dans les données SensNet, qui identifie ces sites ; nous ajoutons ainsi, aux côtés des catégories d'annuaires, une catégorie « pornographie ». Cette description additionnelle complète bien celles des annuaires : pour les données SN2002, seules 0,8 à 1,2 % des pages vues décrites par les annuaires le sont également par la catégorie « pornographie ».

Ces trois sources de descriptions se révèlent très complémentaires (voir Figure 5.9 pour MSN, et Figure 5.10 pour Nomade) : la catégorie « pornographie » s'avère très indépendante des deux autres descripteurs, tandis que les portails généralistes et les services de recherche et de communication par *CatService* (24,1 % des URL de SN2002) et annuaires (couvrant 21 à 24 % des URL SN2002) ne se recoupent que pour 4 à 5 % des URL des données SensNet 2002.

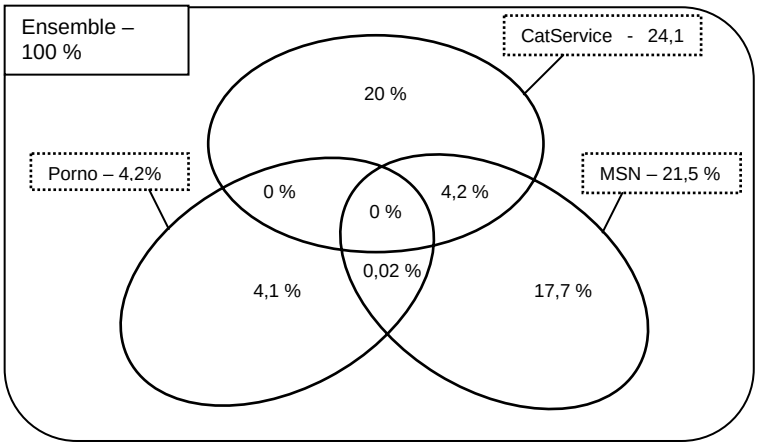


Figure 5.9. Couvertures croisées pour SN2002 (MSN, CatService, pornographique)

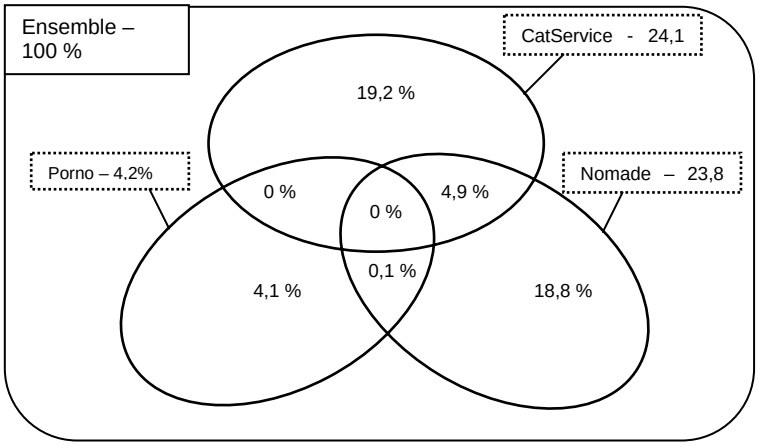


Figure 5.10. Couvertures croisées pour SN2002 (Nomade, CatService, pornographique)

Avec cette utilisation conjointe des annuaires, de *CatService* et de l'information fournie par NetValue sur les sites pornographiques, la liste des descripteurs de sessions contient une base, constituée des informations sur les quinze types de services pour les portails généralistes, des trois services de communication et de recherche et d'une étiquette « porno » (voir Tableau 5.30), à laquelle s'ajoutent les différentes catégories de premier niveau de chaque annuaire (voir Tableau 5.31).

Tableau 5.30. Catégories hors annuaires utilisées pour décrire les sessions

Services Web
Moteur WebMail WebChat
Portails généraliste : services
Achat Aide Annuaire Bourse Communication Divers Généralités Information Produit Information Service Informations Loisir En Ligne Non catégorisé Page Accueil Page Perso Personnalisation / Données Personnelles
Site pornographique

Tableau 5.31. Catégories de premier niveau pour les trois annuaires retenus

MSN (15 catégories)	Nomade (12 catégories)	Yahoo (14 catégories)
Arts - Culture - Médias	Actu, médias	Actualités et médias
Emploi, formation	Culture et loisirs	Art et culture
Entreprises	Éducation, formation	Commerce et Économie
Finances - Bourse - Patrimoine	Espace B to B	Divertissement
Informatique - Internet	Forme et Santé	Enseignement et Formation
Infos - Météo	Mes Courses	Exploration géographique
Jeux - Consoles	Nature et sciences	Informatique et Internet
Loisirs - Passions	Nouvelles technologies	Institutions et politique
Santé	Société, Vie pratique	Références et annuaires
Savoir - Éducation	Sorties, spectacles	Santé
Sciences - Techniques	Sport et détente	Sciences et technologies
Shopping	Voyage, géographie	Sciences humaines
Sports		Société
Vie quotidienne - Société		Sports et loisirs
Voyages - Tourisme		

Au terme de cette utilisation combinée des annuaires, de l'identification des services de recherche (moteurs, annuaires), de communication (WebMail, WebChat, forums) et de ceux fournis par les portails généralistes par *CatService*, et du marquage des sites pornographiques, les taux de couverture des sessions par les descriptifs sont sensiblement améliorés. Globalement, les sessions sont décrites à 48 % en termes de durée ; la part des sessions couvertes à plus de 50 % – celles que l'on pourra exploiter – augmente nettement, et oscille entre 46 et 53 % en fonction du couple annuaire-données retenu (voir Tableau 5.32).

Tableau 5.32. Part de l'ensemble des sessions couvertes à plus de 50 % pour chaque couple annuaire-données (annuaires, CatService, catégorie « porno »)

	BibUsages	SN00-02	SN2002
MSN	49,0 %	46,4 %	52,3 %
Nomade	48,4 %	48,7 %	52,4 %
Yahoo	46,3 %	49,0 %	53,5 %

Parmi ces sessions bien couvertes, on compte généralement entre 30 et 35 % de sessions complètement décrites (environ 15 % de l'ensemble des sessions). Pour les autres, le taux de couverture est également réparti entre 50 et 100 %.

> ***Synthèse.*** *L'utilisation des annuaires seuls pour décrire les contenus visités dans les parcours se heurte à un double problème : d'une part, seules 25 à 30 % des sessions sont suffisamment décrites pour être exploitées ; d'autre part, les pages vues sur les portails généralistes sont rattachées à une catégorie unique dans les annuaires, ce qui est réducteur par rapport à la diversité de leur offre, et très gênant du fait de leur forte audience. La mobilisation des descriptions issues de* CatService *permet de lever ce double biais, grâce à une description fine des différents services sur les portails ; l'adjonction à ce dispositif d'une catégorie « pornographie » permet finalement de décrire correctement la moitié des sessions des panels, ce qui constitue une base solide pour l'analyse des parcours.*

5.3.2 Contenus visités

En travaillant sur les sessions décrites par les annuaires et *CatService* pour les portails généralistes pour plus de la moitié de leur durée, nous avons pour chaque session entre trente et trente-trois descripteurs de contenu, et pouvons commencer à examiner le contenu et la diversité des sessions en termes de thématiques et de services.

Panorama des types de contenus visités

En matière de diversité des contenus visités, le nombre de catégories différentes décrivant chaque session nous fournit un bon indice de l'éclatement des sessions : pour chaque jeu de données et chaque annuaire, nous avons examiné le nombre de descripteurs distincts pour chaque session. En premier lieu, on remarque la grande homogénéité des neuf résultats obtenus : quels que soient les données et l'annuaire retenus, les chiffres relatifs au nombre de catégories sont très similaires, ce qui tend à montrer que les comportements sont assez semblables et généraux en ce qui concerne l'éclatement thématique / fonctionnel de chaque session.

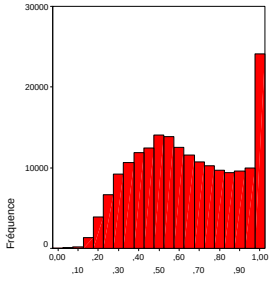
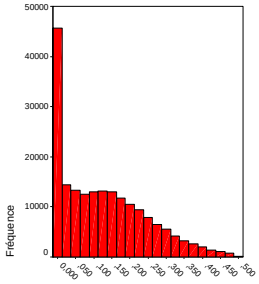
Les résultats eux-mêmes montrent une faible dispersion, avec en moyenne 3,5 catégories différentes visitées, et une médiane à 3 (voir Tableau 5.33 pour le calcul sur les données SN2002). D'autant plus que l'utilisation des descriptifs de *CatService* fait entrer dans ce compte les pages d'accueil des fournisseurs d'accès, points de passage rapides et non significatifs dans la session.

Tableau 5.33. SN2002, nombre de catégories descriptives par session

		MSN	Nomade	Yahoo
Moyenne		3,5	3,5	3,4
Médiane		3	3	3
Quintiles	20	1	1	1
	40	2	2	2
	60	4	4	4
	80	5	5	5

Dans la majorité des cas, la catégorie qui occupe le plus de temps dans la session représente une part importante de la durée totale de la session. Ici encore, les résultats sont très similaires entre les trois jeux de données et les trois annuaires (voir Tableau 5.34 pour le calcul sur les données SN2002) : dans l'ensemble, la catégorie la plus vue occupe presque les deux tiers de la durée de la session en moyenne, et ce quelle que soit la catégorie, tandis que la deuxième catégorie occupe un peu plus de 10 % de la session seulement. La troisième catégorie a un taux de couverture plus faible encore, proche de 3,5 % dans la plupart des cas.

Tableau 5.34. Part des catégories 1 et 2 dans la durée des sessions bien décrites –SN2002

	Catégorie la plus vue			Catégorie n° 2		
	MSN	Nomade	Yahoo	MSN	Nomade	Yahoo
Moyenne	63,7 %	63,8 %	64,1 %	12,7 %	12,6 %	12,3 %
Médiane	62,2 %	62,4 %	62,8 %	10,7 %	10,6 %	10,1 %
Distribution des valeurs						

Sur cette base, une première approche des contenus des sessions consiste à examiner la catégorie descriptive qui occupe la durée la plus importante dans la session. Nous pratiquons en premier lieu cet examen sur les données SensNet en 2002, les plus représentatives des usages généraux des internautes.

Tableau 5.35. Répartition des sessions par catégorie la plus forte dans la session, MSN et Nomade – SN2002[1]

MSN et CatService	%
TS - WebMail	16,6 %
PG - Page Accueil	12,2 %
Informatique - Internet	9,9 %
Finances - Bourse - Patrimoine	5,6 %
TS - WebChat	5,2 %
Entreprises	4,7 %
Jeux - Consoles	4,5 %
PORNO	4,5 %
Infos - Météo	4,2 %
Arts - Culture - Médias	4,1 %
Loisirs - Passions	3,8 %
TS - Moteur	3,2 %
Vie quotidienne - Société	3,1 %
PG - Informations	3,0 %
PG - Personnalisation	2,3 %
Shopping	2,1 %
PG - Non catégorisé	2,0 %
Sports	1,4 %
Savoir - Éducation	1,3 %
PG - Communication	1,3 %
Voyages - Tourisme	1,0 %

Nomade et CatService	%
TS - WebMail	16,5 %
PG - Page Accueil	12,1 %
Espace B to B	9,3 %
Société, Vie pratique	8,1 %
Sport et détente	7,9 %
Mes Courses	7,6 %
TS - WebChat	5,2 %
Actu, médias	4,5 %
PORNO	4,5 %
TS - Moteur	3,2 %
PG - Informations	3,0 %
Culture et loisirs	2,9 %
Nouvelles technologies	2,4 %
PG - Personnalisation	2,3 %
PG - Non catégorisé	2,0 %
Voyage, géographie	1,9 %
Éducation, formation	1,5 %
PG - Communication	1,3 %
Forme et Santé	0,8 %
PG - Divers	0,8 %
Nature et sciences	0,5 %

[1] Convention de nommage : les descripteurs issus de *CatService* sont préfixés par « PG » pour les portail généralistes, et « TS » pour les types de services (moteur, WebMail, WebChat, Forum).

Emploi, formation	0,9 %
PG - Divers	0,8 %
Santé	0,6 %
Sciences - Techniques	0,3 %
PG - Loisir En Ligne	0,3 %
PG - Annuaire	0,3 %
PG - Achat	0,3 %
PG - Page Perso	0,1 %
PG - Information Service	0,1 %
PG - Aide	0,1 %
PG - Information Produit	0,1 %
PG - Généralités	0,0 %
PG - Bourse	0,0 %

PG - Loisir En Ligne	0,3 %
PG - Annuaire	0,3 %
PG - Achat	0,3 %
Sorties, spectacles	0,2 %
PG - Page Perso	0,1 %
PG - Information Service	0,1 %
PG - Aide	0,1 %
PG - Information Produit	0,1 %
PG - Généralités	0,0 %
PG - Bourse	0,0 %

Clef de lecture : dans 16,6 % des sessions SN2002, c'est sur la catégorie « WebMail » que l'internaute passe le plus de temps dans la session.

Les résultats obtenus sur les trois bases descriptives sont relativement concordantes (voir Tableau 5.35 pour MSN et Nomade) : ils montrent en premier lieu l'importance des portails généralistes et des outils de communication dans les sessions Web, en particulier le WebMail. Ils attestent surtout la grande diversité des thématiques des parcours : les taux de représentation des catégories d'annuaires sont assez équilibrés, et correspondent globalement au nombre de sites présentés par l'annuaire dans la catégorie correspondante.

En revanche, on se gardera bien de donner à partir de ces chiffres des conclusions sur les usages d'Internet en général : les sessions sont ici considérées globalement, indépendamment de l'utilisateur, et rien ne nous renseigne ici sur le nombre de panélistes concernés. Par exemple, la catégorie « Finance – Bourse – Patrimoine » de MSN est la plus vue dans 5,6 % des sessions, mais ce comportement touche plus d'un tiers des utilisateurs du panel.

Nous verrons par la suite dans quelle mesure les différents services et thèmes des parcours sont répartis chez les utilisateurs. Nous avons déjà pu observer dans les pratiques générales d'Internet de grandes disparités entre internautes, tant dans l'intensité des pratiques que dans les différents protocoles et services utilisés, et l'on imagine bien que tout le monde ne s'intéresse pas à tout. Pour l'heure, on se contentera de remarquer, au niveau des sessions, que la diversité est de mise et que les pratiques reflètent la diversité de l'offre de contenus et de services.

Homogénéité à l'intérieur des sessions ?

On conclurait volontiers, sur la base de ces chiffres, que les sessions Web sont globalement mono-thématiques ; ce serait sans compter sur la diversité que masquent ces taux de couverture calculés sur l'ensemble des sessions. Comme nous avons déjà pu le remarquer, les sessions sont très différentes quant au nombre de sites différents visités : dans la mesure où les annuaires indexent principalement des sites plus que des pages, les descriptions de contenu sont directement influencées par le nombre de sites vus dans la session.

Ces disparités ont une première influence sur les taux de couverture des sessions par les descriptifs. Si l'on discrétise le nombre de sessions en quatre modalités aux effectifs similaires, les taux de couverture sont globalement semblables et oscillent

entre 42 et 44 % en durée pour SN2002 (résultats similaires pour les autres jeux de données) ; la part des sessions « bien » décrites est elle aussi stable, entre 39 et 44 %. Par contre, la distribution du taux de couverture est très variable (voir Figure 5.11) : lorsqu'un ou deux sites sont vus, dans la grande majorité des cas, la session est soit complètement décrite, soit pas du tout. À mesure qu'augmente le nombre de sites différents visités dans la session, la description quitte ce comportement binaire : au-delà de cinq sites différents visités dans la session, on a bien peu de chances d'avoir une description complète du parcours de l'internaute.

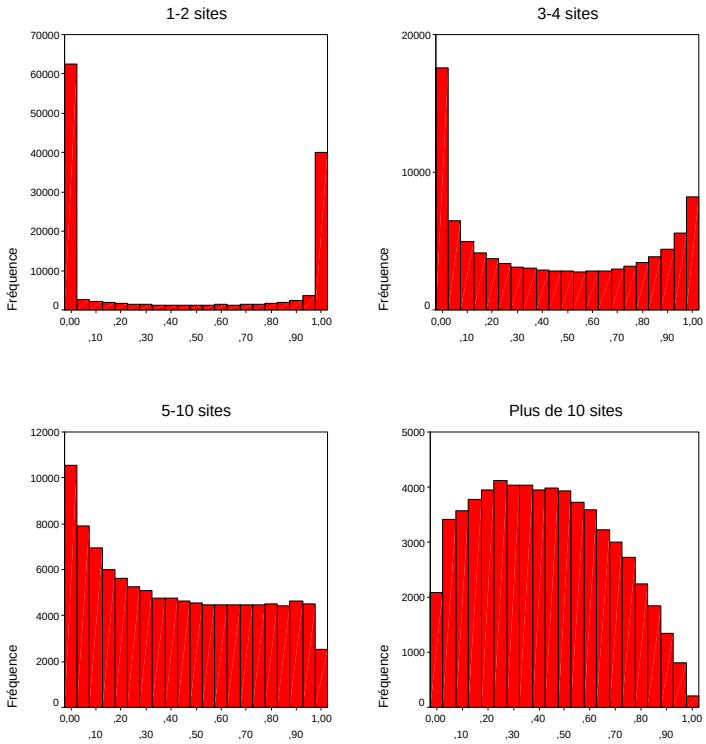


Figure 5.11. Nombre de sites distincts et taux de couverture par les descriptifs de sessions – SN2002, Yahoo-CatService-Porno

Ces éléments affectent directement le nombre de descriptifs différents représentés dans les sessions (voir Tableau 5.36). Malgré le recours à *CatService* qui descend à l'échelle des services pour les portails généralistes et multiplie le nombre possible de catégories descriptives pour ce type de sites, une session impliquant la visite de moins de quatre sites ne relève que de deux ou trois descriptifs en moyenne, contre six lorsque plus de dix sites différents sont vus.

Tableau 5.36. Nombre de sites et moyenne / médiane du nombre de catégories descriptives par session – SN2002

	MSN	Nomade	Yahoo
1-2 site	1,99 / 2	2,00 / 2	1,99 / 2
3-4 sites	3,00 / 3	2,99 / 3	2,89 / 3
5-10 sites	4,22 / 4	4,17 / 4	3,99 / 4
Plus de 10 sites	6,05 / 6	6,00 / 6	5,63 / 5

Corrélativement, la catégorie descriptive la plus représentée dans les sessions occupe de moins en moins de place à mesure que le nombre de sites et de descripteurs augmente (voir Tableau 5.37).

Tableau 5.37. Part de la catégorie la plus vue dans la durée des sessions bien décrites – SN2002

	MSN	Nomade	Yahoo
1-2 site	82 %	82 %	82 %
3-4 sites	64 %	64 %	65 %
5-10 sites	53 %	53 %	54 %
Plus de 10 sites	44 %	44 %	45 %

De 82 % de la durée totale de la session en moyenne lorsqu'un ou deux sites sont vus, elle n'en représente que 44 % lorsque le nombre de sites est supérieur à dix. La part de la deuxième catégorie la plus représentée dans les sessions reste quant à elle relativement stable : elle représente en moyenne 10 % de la durée des sessions lorsque peu de sites sont vus, et 13 % pour les plus riches. L'allongement des sessions s'accompagne donc d'une diversification de leurs contenus. Si l'homogénéité est de mise pour les sessions courtes, où peu de sites sont vus, elle est mise à mal dès lors que plus de cinq sites différents sont vus. Ce constat nous interdit de considérer les contenus visités de manière globale, et invite à les examiner séparément pour chaque groupe de sessions en fonction du nombre de sites vus dans la session.

Tableau 5.38. Répartition des sessions bien décrites par catégorie la plus vue dans la session – SN2002, 1-2 sites distincts visités

MSN	**%**
PG - Page Accueil	22,9
TS - WebMail	17,5
Informatique - Internet	13,1
Finances - Bourse	5,5
PG - Informations	4,1
Jeux - Consoles	3,7
Entreprises	3,7
PG - Personnalisation	3,5
Infos - Météo	3
Arts - Culture - Médias	2,7
Autres	20,3

Nomade	**%**
PG - Page Accueil	23,1
TS - WebMail	17,7
Espace B to B	13
Société, Vie pratique	7,2
Sport et détente	5,4
Mes Courses	5,3
PG - Informations	4,2
PG - Personnalisation	3,5
Actu, médias	3
TS - WebChat	2,5
Autres	15,3

Yahoo	**%**
PG - Page Accueil	24
Commerce et Economie	22,5
TS - WebMail	18,3
PG - Informations	4,3
PG - Personnalisation	3,6
Sports et loisirs	3,5
Exploration géographique	2,8
TS - WebChat	2,5
PG - Non catégorisé	2,5
Actualités et médias	2,4
Autres	13,6

Clef de lecture : si l'on décrit les sessions par les catégories MSN et *CatService*, la catégorie de *CatService* « WebMail » est la plus représentée dans 17,5 % des sessions.

Deux exemples illustreront le bien-fondé de cette démarche : si l'on compare la catégorie la plus représentée dans les sessions d'un ou deux sites d'une part, et de plus de dix sites d'autre part, les résultats sont sensiblement différents. Dans le premier cas, la page d'accueil des portails généralistes domine, suivie du WebMail (voir Tableau 5.38 ci-dessus). On peut faire l'hypothèse, pour ces sessions, d'une visite rapide et très ciblée de services d'information ou de communication, l'utilisateur sachant très bien ce qu'il cherche.

Pour les sessions de plus de dix sites, au contraire, c'est la catégorie des sites pornographiques qui est la plus représentée dans 20 % des sessions (voir Tableau 5.39), contre 4,5 % toutes sessions confondues. Les moteurs de recherche entrent également en jeu, témoignant sans doute de comportements de recherche longue où l'utilisateur est amené à visiter beaucoup de pages de résultats sur des sites différents.

Tableau 5.39. Répartition des sessions bien décrites par catégorie la plus vue dans la session – SN2002, plus de 10 sites distincts visités

MSN	%
PORNO	18,9
TS - Moteur	8,1
Informatique - Internet	7,7
TS - WebMail	7
TS - WebChat	6,5
Arts - Culture - Médias	5,5
Finances - Bourse	5,5
Jeux - Consoles	5,2
Loisirs - Passions	4,4
Entreprises	4,4
Infos - Météo	4,1
PG - Page Accueil	3,5
Autres	19,2

Nomade	%
PORNO	18,2
Sport et détente	9,1
Mes Courses	8,9
TS - Moteur	7,9
Société, Vie pratique	7,5
TS - WebMail	6,6
Espace B to B	6,6
TS - WebChat	6,3
Actu, médias	5,3
Culture et loisirs	4,1
Nouvelles technologies	3,4
PG - Page Accueil	3,2
Autres	12,9

Yahoo	%
Commerce et Economie	25
PORNO	18
TS - Moteur	7,8
TS - WebMail	6,7
TS - WebChat	6,5
Sports et loisirs	4,8
Exploration géographique	4,8
Actualités et médias	3,6
PG - Page Accueil	3,3
Divertissement	2,7
Société	2,6
Art et culture	2,1
Autres	11

Ici encore, le panel BibUsages marque sa différence : pour ces sessions longues, les catégories MSN les plus présentes sont « Arts – Culture – Médias » (14,6 % des sessions), « TS – Moteur » (13,8 %) et « Loisirs – Passions » (12,7 %) ; les sites pornographiques ne sont pas en reste pour autant (11,1 %), mais sont en retrait par rapport au panel général des internautes à la même période.

Plus généralement, les différences de thèmes et de services au regard du nombre de sites visités et de la durée des sessions confirment la nécessité de subordonner l'étude des contenus à celle de la topologie des parcours. Forme, durée, rythme des sessions sont la trace de chaînes opératoires qui sont étroitement liées aux thèmes et aux services visés par l'utilisateur, et témoignent de contextes d'usage différenciés.

> ***Synthèse.*** *L'examen des contenus visités dans les sessions montre l'importance des portails généralistes et des outils de communication dans les usages du Web, en particulier le WebMail. Considérés de manière globale, ils attestent la grande diversité des thématiques des parcours, mais à l'échelle de la session, cette diversité est beaucoup plus restreinte. Si les sessions courtes sont fortement mono-thématiques ou mono-fonctionnelles, et renvoient à un cours d'action unique, l'allongement des sessions, en durée comme en nombre de sites, s'accompagne d'une diversification de*

leurs contenus. Elle est également corrélée aux types de contenus eux-mêmes : dans les sessions courtes, les services d'information ou de communication dominent, tandis que les sessions longues mettent en avant les sites pornographiques et les moteurs de recherche. Ces éléments attestent le lien fort entre topologie et contenu des parcours sur le Web.

5.4 Profils de sessions

De l'analyse des usages généraux d'Internet au détail de l'activité au sein des sessions, on a pu observer une grande diversité dans les comportements : l'intensité d'usage du Web variable selon les utilisateurs fait écho aux disparités dans la durée, l'éclatement, le rythme et les contenus de parcours. Une approche typologique des parcours doit permettre de rendre compte de cette diversité et des régularités qui s'y jouent.

5.4.1 Classification

Sur la base de l'observation précise des éléments rythmiques, temporels et topologiques des sessions, nous sommes maintenant en mesure de construire une classification des sessions. Il s'agit ici d'affiner les oppositions qui ont pu être mises à jour dans l'analyse des composantes topologiques prises séparément.

Variables retenues

L'examen des variables temporelles, du nombre de sites et des indicateurs nous permet de les sélectionner et de les organiser de manière pertinente. Nous avons en effet vu que les variables continues ici manipulées masquent des distributions et des réalités qui nous obligent à les discrétiser : le nombre de sites différents vus dans une session ainsi que sa durée peuvent par exemple avoir des valeurs soit très faibles, soit extrêmes, mais il existe dans les faits une différence forte entre sessions très courtes avec peu de sites, et sessions plus longues et plus complexes.

De la même manière, le taux de linéarité, qu'il soit calculé à l'échelle du site ou de la page, peut avoir toutes les valeurs comprises entre 0 et 1, mais la valeur 1 correspond à une session linéaire, classe à part qui implique mécaniquement certaines valeurs pour les autres taux (taux de concentration, nombre de *back*, etc.). Dans tous les cas, les chiffres figurent une continuité là où, dans les pratiques, on observe des réalités et des comportements bien distincts.

Partant, nous avons discrétisé l'ensemble des variables pertinentes pour l'analyse en 3 à 5 classes de manière à obtenir des groupes homogènes en termes d'effectifs et surtout en termes de comportements sous-jacents ; le Tableau 5.40 présente la liste de ces variables et des différentes modalités construites pour chacune.

Tableau 5.40. Discrétisation des variables temporelles et topologiques retenues

Variables	Échelle d'analyse	Discrétisation	Notation	% des sess.
Nombre de sites distincts	-	1-2	1-2 sites	33,6 %
		3-4	3-4 sites	23,6 %
		5-10	5-10 sites	27,1 %
		>10	Plus de 10 sites	15,8 %
Durée de la session	-	1-3	1-3 min.	24,6 %
		4-13	4-13 min.	25,0 %
		14-34	14-34 min.	24,1 %
		>35	Plus de 35 min.	26,3 %
Durée médiane par page	-	0-2	DPage – 0-2 sec.	22,2 %
		3-4	DPage – 3-4 sec.	12,7 %
		5-9	DPage – 5-9 sec.	19,8 %
		10-15	DPage – 10-15 sec.	17,0 %
		>16	DPage – 16 sec. et plus	28,4 %
Durée médiane par site	-	0-9	DSite – 0-9 sec.	7,6 %
		10-19	DSite – 10-19 sec.	11,5 %
		20-29	DSite – 20-29 sec.	11,2 %
		30-1'09	DSite – 30 sec-1 min. 09	36,8 %
		>1'10	DSite – 1 min. 10 et plus	32,9 %
Taux de linéarité	Page	0-0,29	Pages – peu linéaire	2,8 %
		0,3-0,59	Pages – moy linéaire	24,6 %
		0,6-0,99	Pages – très linéaire	54,0 %
		1	Pages – linéaire	18,6 %
Taux de linéarité	Site	0-0,29	Sites – peu linéaire	13,0 %
		0,3-0,59	Sites – moy linéaire	25,9 %
		0,6-0,99	Sites – très linéaire	25,1 %
		1	Sites – linéaire	36,0 %
Taux de linéarité calculé sur la durée	Page	0-0,49	Pages – dur. Peu linéaire	35,5 %
		0,5-0,75	Pages – dur. Moy. linéaire	25,1 %
		0,76-0,99	Pages – dur. Très linéaire	20,6 %
			Pages – dur. Linéaire	18,8 %
Taux de linéarité calculé sur la durée	Site	0-0,29	Sites – dur. Peu linéaire	37,5 %
		0,3-0,59	Sites – dur. Moy. linéaire	13,8 %
		0,6-0,99	Sites – dur. Très linéaire	12,7 %
		1	Sites – dur. Linéaire	36,0 %
Concentration des revisites	Page	1	P – Conc. nulle	20,6 %
		1-1,9	P – Conc. faible	30,1 %
		2-2,9	P – Conc. moyenne	17,7 %
		3 et plus	P – Conc. forte	13,1 %
		-	P – Conc. undef (pas de revisite)	18,6 %
Concentration des revisites	Site	1	S – Conc. nulle	18,1 %
		1-1,9	S – Conc. faible	10,9 %
		2-3,49	S – Conc. moyenne	18,4 %
		3,5 et plus	S – Conc. forte	16,5 %
		-	S – Conc. undef (pas de revisite)	36,0 %
Nombre d'actions back	Page	0	P back – aucune	31,3 %
		1-3	P back – moyen	33,1 %
		4 et plus	P back – beaucoup	35,6 %

Nombre d'actions back	Site	0 1-4 5 et plus	S back – aucune S back – moyen S back – forte	39,7 % 36,4 % 23,9 %

Classification sur les données SN2002

Nous avons ainsi douze variables, pour un total de cinquante modalités distinctes. Afin d'avoir une vue la plus représentative possible des différents types de sessions que l'on peut rencontrer dans les pratiques, nous avons pratiqué préférentiellement une classification sur les données SN2002. Celle-ci servira de base de référence ensuite pour l'analyse des sessions observées dans les données SN00-02 et BibUsages.

En premier lieu, les résultats de l'analyse en composantes multiples pratiquée sur les 400 000 sessions de SensNet en 2002 sont encourageants (voir Tableau 5.41) : les dix premiers axes factoriels représentent près de 60 % de la variance de l'échantillon, et les cinq premiers axes résument à eux seuls 44 % de l'information.

Tableau 5.41. Classification sur la topologie des sessions, SN2002 – valeurs propres

Numéro d'axe	Valeur propre	Pourcent	Pourcent cumulé
1	0,6028	19,0 %	19,0 %
2	0,3054	9,6 %	28,7 %
3	0,1898	6,0 %	34,7 %
4	0,1518	4,8 %	39,5 %
5	0,1321	4,2 %	43,6 %
6	0,1120	3,5 %	47,2 %
7	0,1040	3,3 %	50,5 %
8	0,0981	3,1 %	53,6 %
9	0,0940	3,0 %	56,5 %
10	0,0908	2,9 %	59,4 %

On est donc bien fondé à effectuer une classification ascendante hiérarchique sur les résultats de l'ACM, en prenant en compte les dix premiers axes factoriels. Trois coupures de l'arbre sont proposées à l'issue de la classification, partitionnant l'échantillon en cinq, huit ou dix classes. Nous avons retenu la partition en cinq classes, qui présente le saut le plus important et les variabilités intra et inter-classes les plus significatives.

Les résultats de cette classification donnent des groupes bien distincts. La Figure 5.12 p. 219 représente la projection des individus sur les axes 1 et 2 (29 % de l'information représentée) : elle oppose très nettement les classes 1 et 2 aux trois autres. La projection des modalités des variables sur le graphique montre une opposition forte sur l'axe 1 entre sessions linéaires et non linéaires, tandis que l'axe 2 semble distinguer, au sein des sessions non linéaires, celles qui le sont peu et celles qui le sont beaucoup.

Dans une vue basée sur les axes 3 et 4 (Figure 5.13, p. 220), on retrouve cette homogénéité des classes 3, 4 et 5 qui forment un noyau central, autour duquel s'opposent les classes 1 et 2. Ici, sont essentiellement distinguées la linéarité inter-sites ainsi que le taux de concentration inter-site (axe 3), et le temps médian passé par visite de site (axe 4).

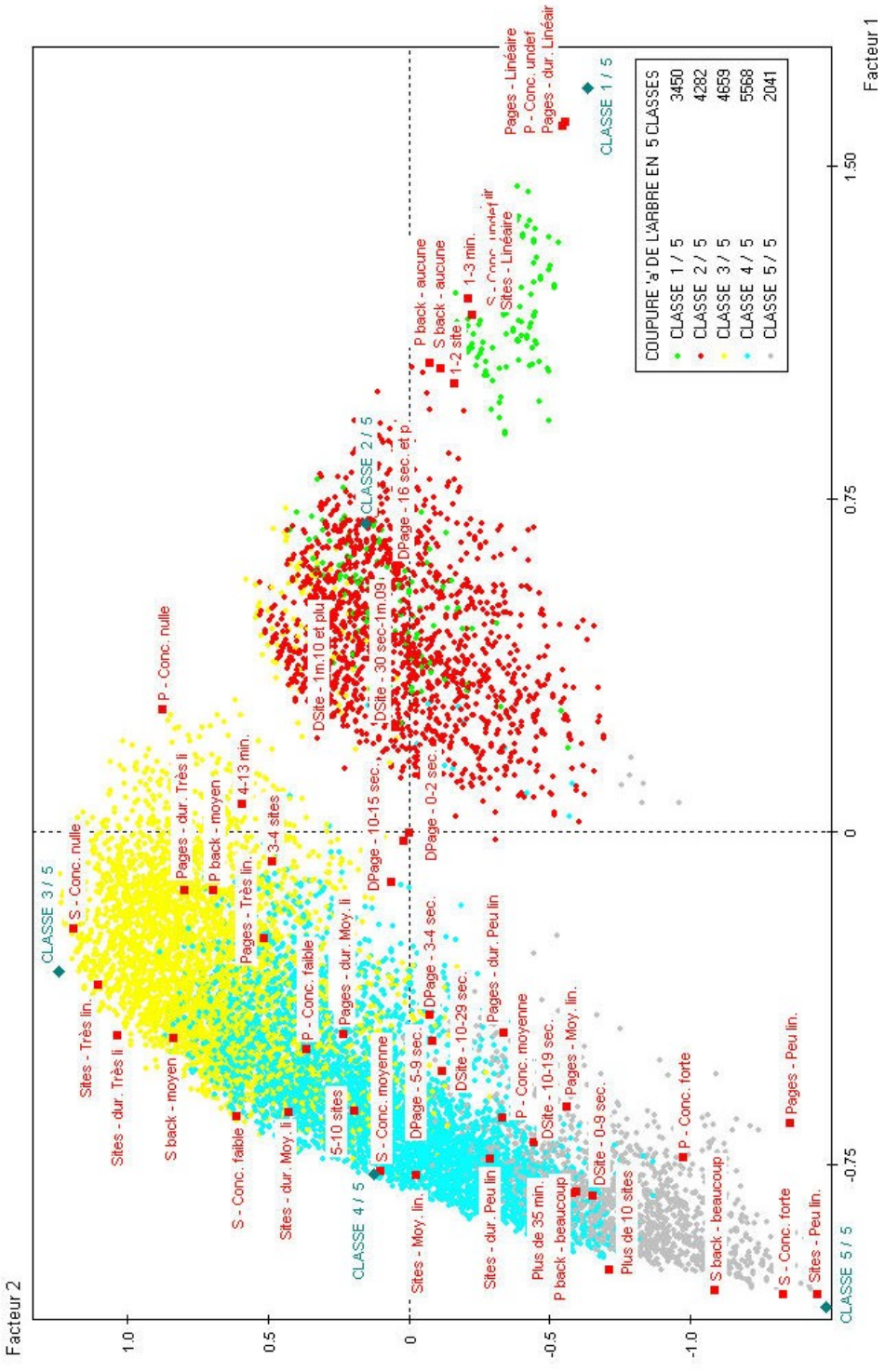


Figure 5.12. Classification sur les indicateurs topologiques, SN2002 - axes 1 et 2

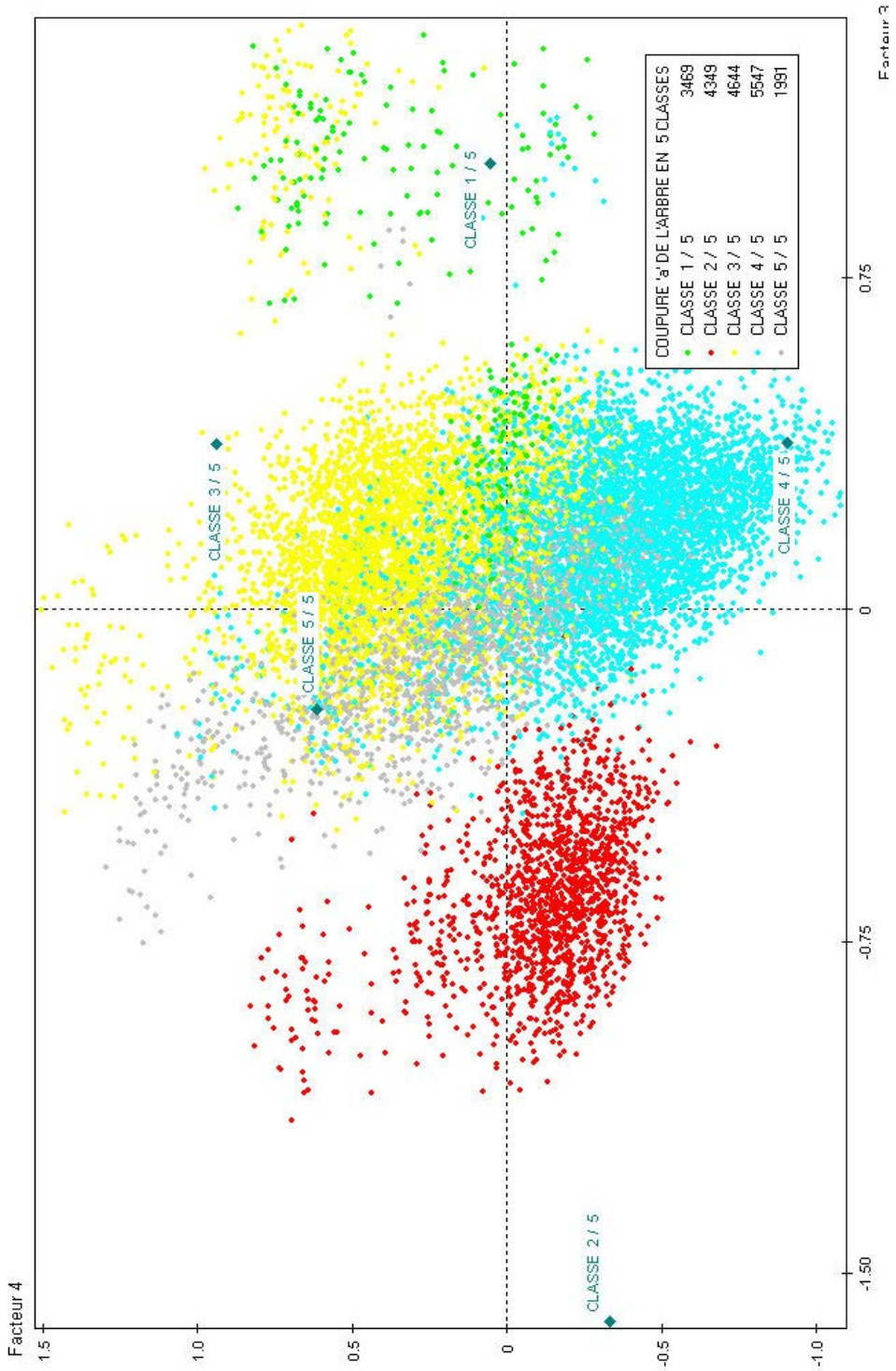


Figure 5.13. Classification sur les indicateurs topologiques, SN2002 - axes 3 et 4

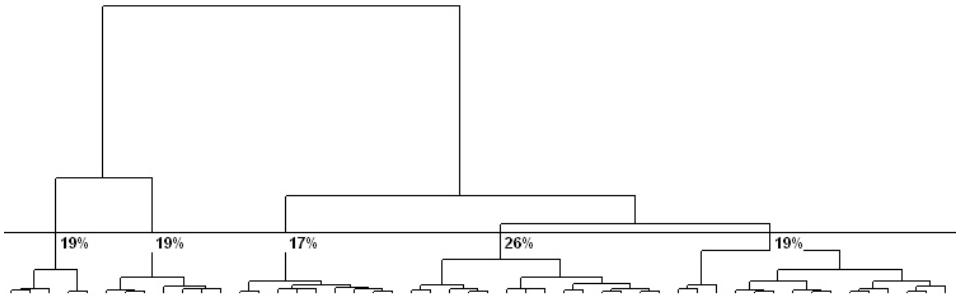


Figure 5.14. Classification sur la topologie des sessions, SN2002 – dendrogramme

Les classes 1 et 2 forment ainsi un groupe homogène de sessions qui s'opposent aux trois autres groupes, ce qu'illustre le dendrogramme représentant la classification (voir Figure 5.14). Les sessions de ce groupe sont courtes, linéaires ou quasi-linéaires, peu de sites sont vus et l'utilisateur passe plutôt du temps sur chaque site visité ; on semble être ici dans le contexte de visites ciblées où l'internaute accomplit des actions bien précises, qui ne l'amènent pas à sortir de son champ d'action.

Le deuxième grand groupe de sessions comprend trois classes distinctes, qui sont globalement caractérisées par une linéarité moyenne à faible, un nombre plus élevé de sites et de pages, et un allongement dans la durée. Le corpus voit ainsi s'opposer des sessions plutôt courtes et directes aux sessions allongées et plus complexes.

> ***Synthèse.*** *La structure particulière des indicateurs topologiques nous oblige à y opérer une discrétisation manuelle tenant compte de la spécificité des comportement qu'ils représentent. L'analyse en composantes multiples et la classification des sessions menées sur cette base traduisent la nécessité de subordonner l'analyse des contenus à celle des modes d'activité et des structures de navigation. Cinq groupes sont distingués, soumis à une opposition globale entre sessions courtes et linéaires et sessions allongées et complexes.*

5.4.2 Profils de sessions

Groupe des sessions courtes et directes

La première classe (17,4 % des sessions) est constituée de sessions courtes (1 à 3 minutes) et linéaires, et agrège des variables qui se retrouvent étroitement liées pour certaines valeurs particulières (voir Tableau 5.42 ci-dessous) : linéarité à l'échelle de la page et du site, absence de mouvements de *back*, faible nombre de sites visités (un ou deux).

On nommera cette classe « parcours éclairs » : avec des durées courtes (une à trois minutes, contre 35 minutes en moyenne pour l'ensemble des sessions), deux sites au plus, moins d'une minute par site, ces sessions Web sont fondamentalement caractérisées par leur brièveté. La Figure 5.15 en donne deux illustrations parmi les individus les plus représentatifs de la classe.

Tableau 5.42. SN2002, classification sur les indicateurs topologiques – Parcours éclairs

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
100 %	18,6 %	Linéarité	Page	Linéaire
100 %	18,6 %	Concentration des revisites	Page	Conc. undef
100 %	18,8 %	Linéarité sur la durée	Page	Linéaire
100 %	31,3 %	Nb action back	Page	Aucune
92,4 %	36,0 %	Linéarité sur la durée	Site	Linéaire
92,4 %	36,0 %	Linéarité	Site	Linéaire
92,4 %	36,0 %	Concentration	Site	Conc. undef
79,5 %	24,6 %	Durée de session	-	1-3 min.
93,4 %	39,7 %	Nb actions back	Site	Aucune
81,9 %	33,6 %	Nb sites distincts	-	1-2 site
61,0 %	28,4 %	Durée médiane par visite	Page	16 sec. et plus
60,4 %	36,8 %	Durée médiane par visite	Site	30 sec-1m.09

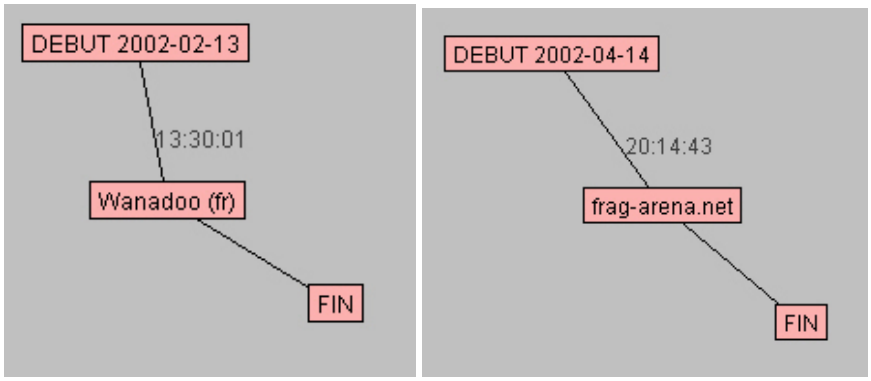


Figure 5.15. SN2002 - exemples typiques de la classe « parcours éclairs » (inter-site)

La classe 2 des « parcours ciblés » (19,9 % de l'ensemble) se rapproche de la classe 1 par le faible nombre de sites (1 à 2 sites) et une forte présence de sessions linéaires en inter-sites (voir Tableau 5.43).

Tableau 5.43. SN2002, classification sur les indicateurs topologiques – Parcours ciblés

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
100 %	36,0 %	Concentration	Site	Conc. undef
100 %	36,0 %	Linéarité	Site	Linéaire
100 %	36,0 %	Linéarité sur la durée	Site	Linéaire
100 %	39,7 %	Nb actions back	Site	Aucune
71,9 %	32,9 %	Durée médiane par visite	Site	1m.10 et plus
69,5 %	33,6 %	Nombre de sites distincts	-	1-2 site
49,7 %	20,6 %	Concentration	Page	Conc. nulle
55,5 %	33,1 %	Nb actions back	Page	Moyen
73,5 %	54,0 %	Linéarité	Page	Quasi-linéaire

Par contre, à l'échelle de la page, les sessions ne sont pas linéaires, mais seulement « quasi-linéaires » (taux entre 0,6 et 1) : à l'intérieur d'un site, l'utilisateur est amené à revoir modérément certaines pages. Cette revisite semble due principalement aux mouvements de type *back*, dont le nombre est relativement important en regard de la durée générale des sessions du groupe. L'exemple de session donné ci-dessous illustre ces éléments : à l'échelle du site (Figure 5.16), la session est linéaire, avec deux sites visités ; à l'échelle de la page (Figure 5.17), la linéarité se trouve brisée par quelques « écarts » ponctuels.

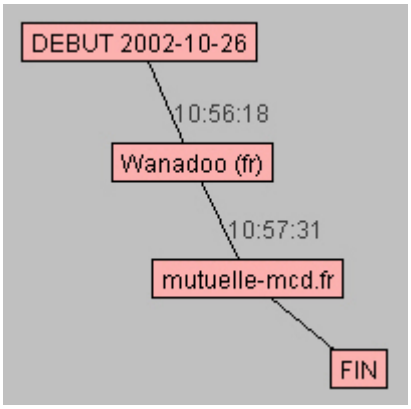


Figure 5.16. SN2002 - exemple typique de la classe « parcours ciblés »

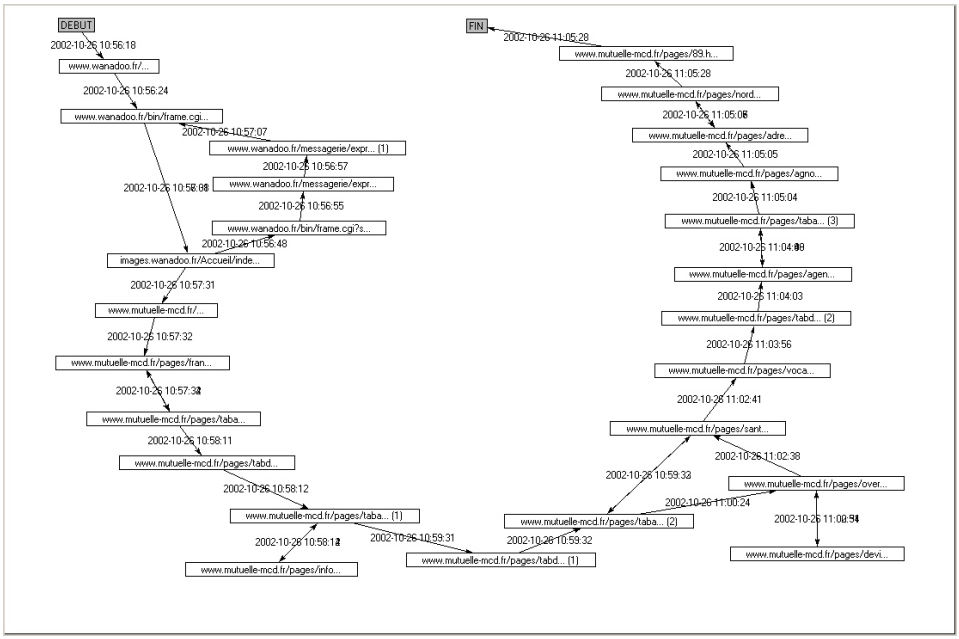


Figure 5.17. SN2002 - exemple de la classe « parcours ciblés » (inter-page)

Enfin, en termes de durée, ces sessions sont moins déterminées que celles du premier groupe, la durée n'intervenant pas comme modalité caractéristique de la

classe : il semble ici s'agir plutôt de « parcours ciblés », où les sites visés sont bien identifiés, mais la navigation au sein de ces sites s'allonge et se complexifie.

Sur le plan du contenu, parcours éclairs et parcours ciblés sont assez similaires : la projection des catégories d'annuaires et de *CatService* sur ces deux groupes montre une prédominance de la page d'accueil des portails généralistes pour le premier, et du WebMail pour le second, ce qui va de pair avec l'hypothèse de sessions très ciblées. À l'inverse, les services de recherche sur le Web sont complètement absents de ces sessions dont le faible nombre de sites implique qu'elles correspondent à des sites connus et mémorisés (page de démarrage, enregistrement dans des favoris) ou qu'il s'agisse de l'ouverture d'un lien à partir d'une source hors Web, comme un mail par exemple.

Groupe des sessions longues et complexes

La troisième classe, les « parcours à détours », regroupe 21,3 % des sessions et constitue la charnière entre les deux grands groupes : on a pu voir dans la Figure 5.12 (p. 219) que certains des individus de cette classe se confondent avec ceux de la classe 2. Ici, la linéarité à l'échelle du site n'est plus assurée, mais elle reste importante (voir Tableau 5.44), et la revisite de sites est principalement imputables à des mouvements de type *back*.

Tableau 5.44. SN2002, classification sur les indicateurs topologiques – Parcours à détours

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
95,5 %	25,1 %	Linéarité	Site	Quasi-linéaire
78,0 %	18,1 %	Concentration	Site	Conc. nulle
85,1 %	36,4 %	Nb actions back	Site	Moyen
44,1 %	12,7 %	Linéarité sur la durée	Site	Quasi-linéaire
76,3 %	54,0 %	Linéarité	Page	Quasi-linéaire
36,8 %	20,6 %	Concentration	Page	Conc. nulle
36,8 %	20,6 %	Linéarité sur la durée	Page	Quasi-linéaire
49,8 %	33,1 %	Nb actions back	Page	Moyen
37,3 %	23,6 %	Nb sites distincts	-	3-4 sites
38,6 %	25,0 %	Durée de session	-	4-13 min.
39,2 %	27,1 %	Nb sites distincts	-	5-10 sites
23,0 %	13,8 %	Linéarité sur la durée	Site	Moyennement linéaire
31,3 %	24,1 %	Durée de session	-	14-34 min.

Ces mouvements de *back* ne sont pas articulés autour d'une page pivot, les taux de concentration restant nuls : on est ici dans le cas de parcours simples avec quelques digressions. La linéarité à l'échelle du site calculée sur la durée restant importante, ce qui indique que le temps passé sur des sites revus demeure faible. Dans ces sessions, on observe une véritable « colonne vertébrale » qui dirige le parcours, et les quelques détours ou boucles qui le jalonnent demeurent marginaux. Les exemples ci-dessous de sessions typiques de la classe le confirment (voir Figure

5.18) : dans cette classe des « parcours à détours », on observe des « pas de côté » et des boucles, mais aucune page ni site autour desquels s'articule la navigation.

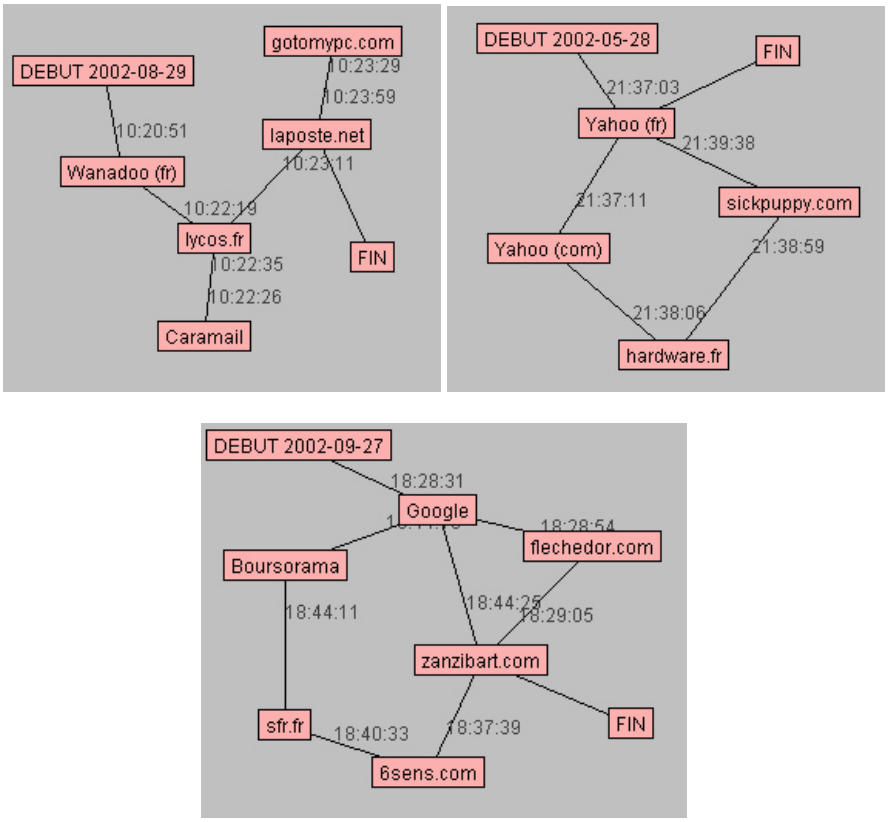


Figure 5.18. SN2002 - exemple typique de la classe « parcours à détours »

En termes de durée, cette classe est plutôt orientée vers des temporalités moyennes (moins d'un quart d'heure), mais avec une palette de sites différents visités pouvant aller jusqu'à dix. Ceci explique que les variables rythmiques (temps moyen par page et par site visité) n'entrent pas dans les variables les plus spécifiques de la classe.

Dans cette classe, on retrouve encore beaucoup de sessions où le WebMail occupe la première place en termes de durée, mais on observe une très nette diversification des thèmes des sessions. Les différentes catégories d'annuaires y sont représentées, avec un accent particulier pour celles relatives à la « vie pratique » : « Infos – météo » (MSN), « Société, vie pratique » (Nomade), « Exploration géographique » (catégorie de Yahoo qui renvoie plus particulièrement à des sites géolocalisés, correspondant à des services accessibles « hors Web » comme les associations, les institutions, etc.). Il semble que ces sessions soient plus orientées vers une pratique du type « pages jaunes » du Web : les moteurs de recherche y sont peu employés, ce qui montre que dans ce contexte, les internautes savent plutôt s'orienter et visitent des sites connus ou suivent des liens à partir de sites de confiance.

C'est dans la classe 4, que l'on va trouver une sur-représentation particulièrement forte des moteurs de recherche. Cette classe des « parcours à pivots » représente 26,4 % de l'ensemble des sessions. Les caractéristiques de la classe montrent un allongement et une complexification des sessions (voir Tableau 5.45) : la linéarité à l'échelle du site est moyenne, et certains sites-pivot apparaissent (concentration moyenne et faible au niveau du site, faible au niveau de la page). La faible linéarité à l'échelle du site en termes de durée montre que ces sites revisités occupent une part importante de la durée des sessions ; cette revisite peut se faire *via* les actions de type *back* (forte présence au niveau des pages comme des sites), mais également par des boucles.

Tableau 5.45. SN2002, classification sur les indicateurs topologiques – Parcours à pivots

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
83,2 %	25,9 %	Linéarité	Site	Moyennement linéaire
57,9 %	18,4 %	Concentration	Site	Conc. moyenne
32,5 %	10,9 %	Concentration	Site	Conc. faible
62,8 %	36,4 %	Nb actions back	Site	Moyen
63,1 %	37,5 %	Linéarité	Site	Peu linéaire
52,7 %	30,1 %	Concentration	Page	Conc. faible
55,2 %	35,6 %	Nb actions back	Page	Beaucoup
30,7 %	15,8 %	Nb sites distincts	-	Plus de 10 sites
27,1 %	13,8 %	Concentration sur la durée	Site	Moyennement linéaire
41,5 %	26,3 %	Durée de la session	-	Plus de 35 min.
40,8 %	27,1 %	Nb sites distincts	-	5-10 sites
35,8 %	23,9 %	Nb actions back	Site	Beaucoup
37,1 %	25,1 %	Linéarité sur la durée	Page	Moyennement linéaire
65,4 %	54,0 %	Linéarité sur la durée	Page	Quasi-linéaire
33,5 %	24,1 %	Durée de la session	-	14-34 min.
25,4 %	17,7 %	Concentration	Page	Conc. moyenne
17,6 %	11,2 %	Durée médiane par visite	Site	10-29 sec.
32,4 %	24,6 %	Linéarité	Page	Moyennement linéaire
16,7 %	11,5 %	Durée médiane par visite	Site	20-19 sec.

Les exemples de sessions témoignent de cette diversité : dans la Figure 5.19, on observe une double boucle accolée, avec un seul mouvement de *back* à l'échelle du site. Les Figure 5.20 et Figure 5.21 montrent des mouvements plus complexes, mêlant boucles et retours-arrière, avec dans les deux cas deux sites qui servent de pivot à la visite des autres sites.

Dans ces retours et ces détours, l'utilisateur est amené à voir beaucoup de sites (plus de 10), et corrélativement la session s'allonge et dépasse la demi-heure. Cet allongement fait entrer dans les variables caractéristiques de la classe les éléments de rythme, absents de la classe des « parcours à détours » : la durée médiane par visite de site comprise entre 10 et 20 secondes – entre 5 et 10 secondes par visite de page – atteste une accélération de la navigation. Une part importante des sites et des pages ne semble être que traversée rapidement, ce qui rejoint l'utilisation forte de la fonction *back*.

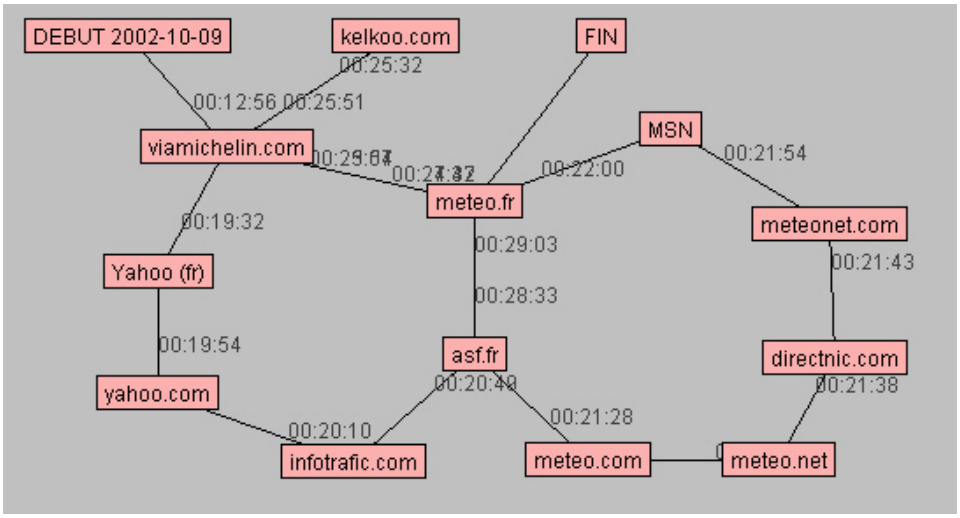


Figure 5.19. SN2002 - exemple typique de la classe « parcours à pivots »

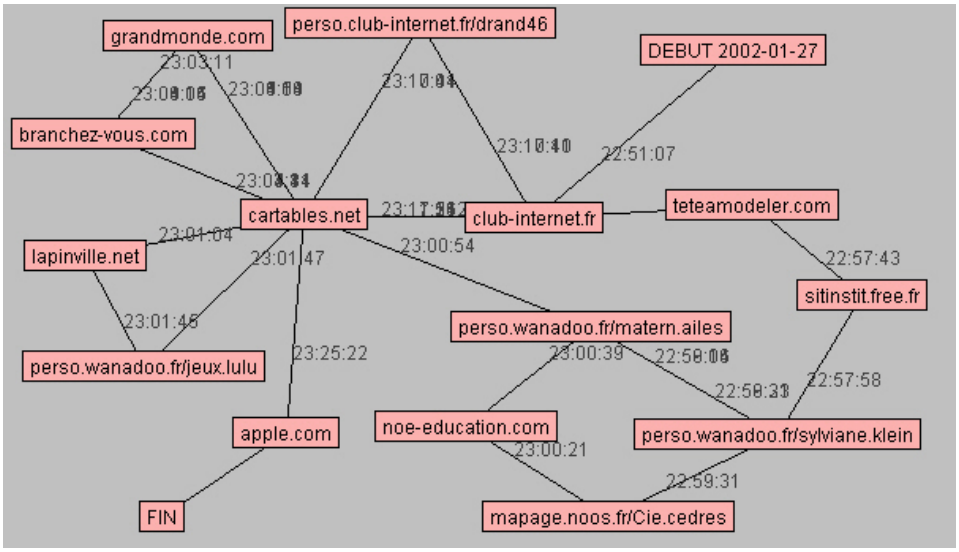


Figure 5.20. SN2002 - exemple typique de la classe « parcours à pivots »

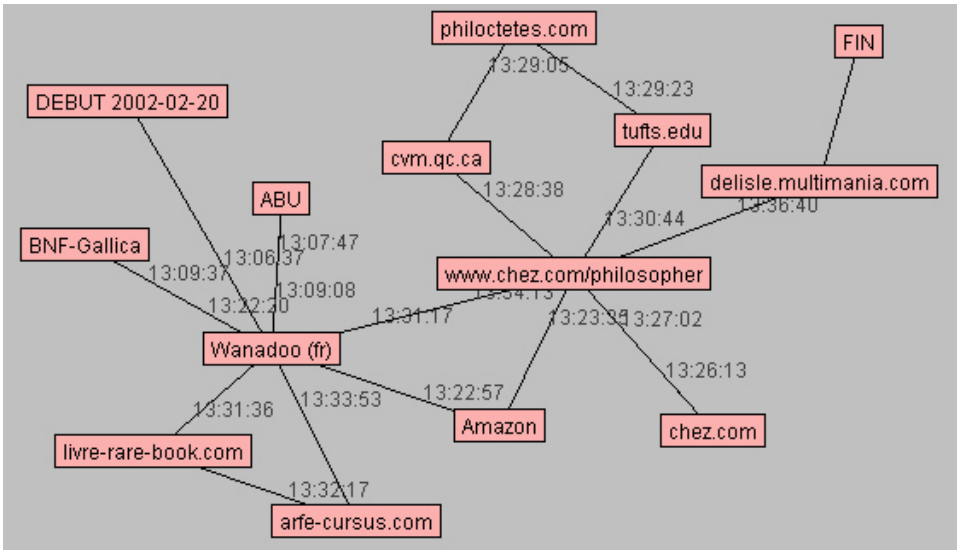


Figure 5.21. SN2002 – exemple typique de la classe « parcours à pivots »

Ces éléments morphologiques sont cohérents avec le recours important aux moteurs de recherche : l'internaute balaye des pages de résultats renvoyés par les moteurs, ou explore de nouveaux sites à partir de liens dans des pages-ressources. Sur le plan des thématiques de ces sessions, on observe un certaine hétérogénéité générale, mais une importante cohérence interne : chaque session semble axée autour d'un thème particulier autour duquel s'articule la recherche de l'internaute. Ce peut être la préparation d'un voyage comme on le voit dans l'exemple de la Figure 5.19, où l'internaute visite des sites d'information météorologique et de préparation d'itinéraire ; ou bien une recherche de textes philosophiques comme dans l'exemple de la Figure 5.21, qui amène l'utilisateur à visiter des bibliothèques numériques (Gallica, ABU), des sites de vente de biens culturels (Amazon, `livre-rare-book.com`) et des sites traitant de la philosophie. Sous-représentés dans les trois autres classes de parcours, les sites pornographiques sont ici très présents, ce qui rejoint le recours important aux moteurs de recherche sur lesquels on sait que les requêtes les plus fréquentes sont liées à ce thème.

La dernière classe, qui concerne 15 % des sessions, s'oppose autant aux « parcours éclairs » qu'aux « parcours à détours » : les caractéristiques principales de cette classe des « parcours éclatés » sont une forte concentration au niveau des sites, et une très faible linéarité des parcours (voir Tableau 5.46). On retrouve ici les sessions les plus longues des données, tant en nombre de sites qu'en durée. La fonction *back* des navigateurs est très utilisée tant au niveau des pages que des sites, et le temps moyen pour chaque passage sur un site est dans les moyennes basses.

Tableau 5.46. SN2002, classification sur les indicateurs topologiques – Parcours éclatés

% dans la classe	% global	Variable	Échelle d'analyse	Modalités caractéristiques
95,2 %	16,5 %	Concentration	Site	Conc. forte
81,6 %	13,0 %	Linéarité	Site	Peu linéaire
96,0 %	23,9 %	Nb actions back	Site	Beaucoup
86,4 %	37,5 %	Linéarité	Site	Peu linéaire
84,1 %	35,6 %	Nb actions back	Page	Beaucoup
63,7 %	26,3 %	Durée de la session	-	Plus de 35 min.
41,2 %	13,1 %	Concentration	Page	Conc. forte
40,1 %	15,8 %	Nb sites distincts	-	Plus de 10 sites
51,4 %	24,6 %	Linéarité	Page	Moyennement linéaire
25,1 %	7,6 %	Durée médiane par visite	Site	0-9 sec.
61,9 %	35,5 %	Linéarité	Page	Peu linéaire
29,0 %	11,5 %	Durée médiane par visite	Site	10-19 sec.
34,3 %	17,7 %	Concentration	Page	Conc. moyenne
9,3 %	2,8 %	Linéarité	Page	Peu linéaire
39,9 %	27,1 %	Nb sites distincts	-	5-10 sites
31,2 %	19,8 %	Durée médiane par visite	Page	5-9 sec.
19,1 %	11,2 %	Durée médiane par visite	Site	20-29 sec.
19,5 %	12,7 %	Durée médiane par visite	Page	3-4 sec.

Si l'on examine de plus près les sessions correspondantes, on constate que ces chiffres rendent compte de deux types de parcours distincts. D'une part, des sessions très éparpillées, où beaucoup de sites sont revus, et certaines boucles sont parcourues plusieurs fois, ce qu'illustrent les Figure 5.22 et Figure 5.23.

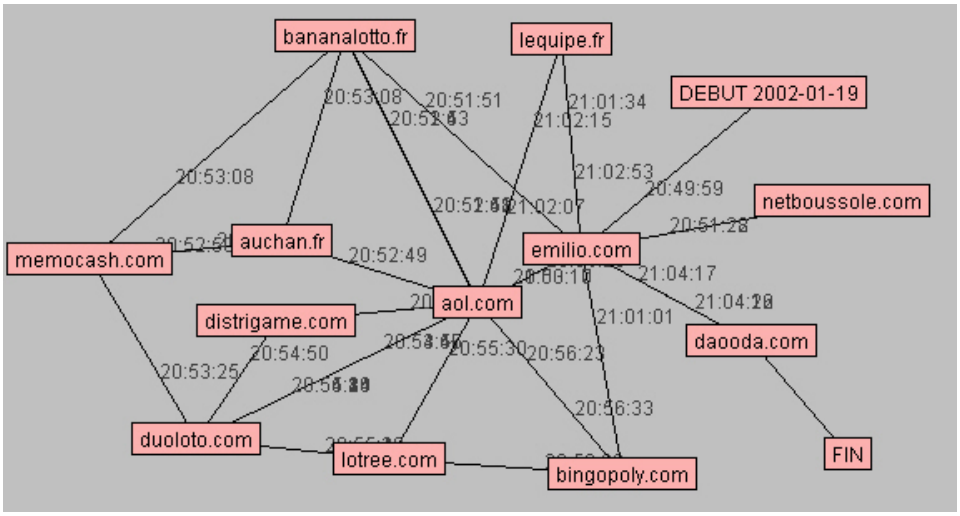


Figure 5.22. SN2002 – exemple typique de la classe « parcours éclatés »

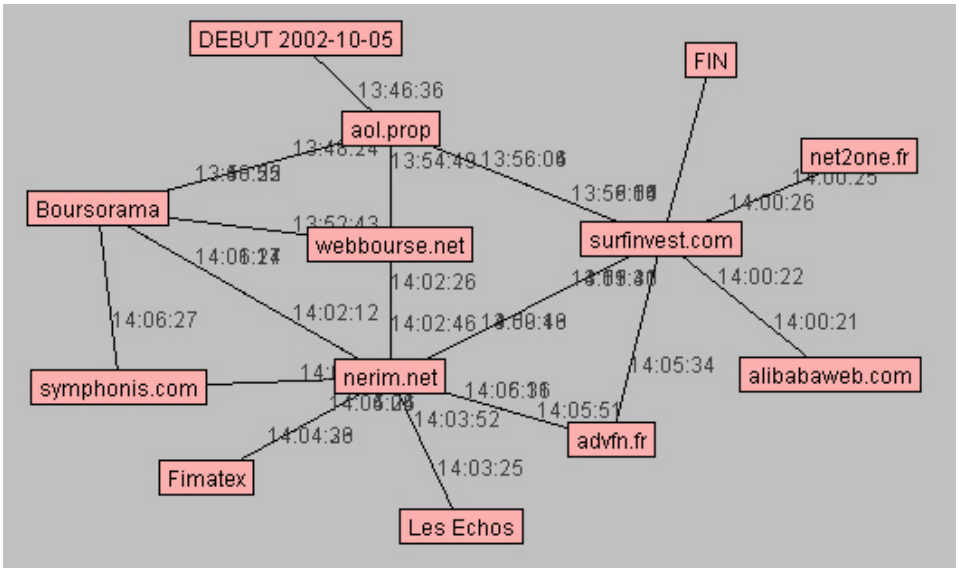


Figure 5.23. SN2002 – exemple typique de la classe « parcours éclatés »

D'autre part, des sessions où certains sites « complémentaires » ou certaines pages à taux de rafraîchissement élevé (les services de *chat*, par exemple) figurent dans les données une redondance et des échanges rapides entre deux sites ou deux pages. Ces éléments ont pour conséquence de figurer des taux de concentration très élevés, mais il s'agit dans la plupart des cas d'un biais des données. Ainsi, dans la session présentée Figure 5.24, les aller-retour entre les sites `locatorserver.net` et `cyberain.net` sont un artéfact par rapport au point de vue de l'utilisateur, `locatorserver.net` étant un site de bannières publicitaires mal identifié dans les données.

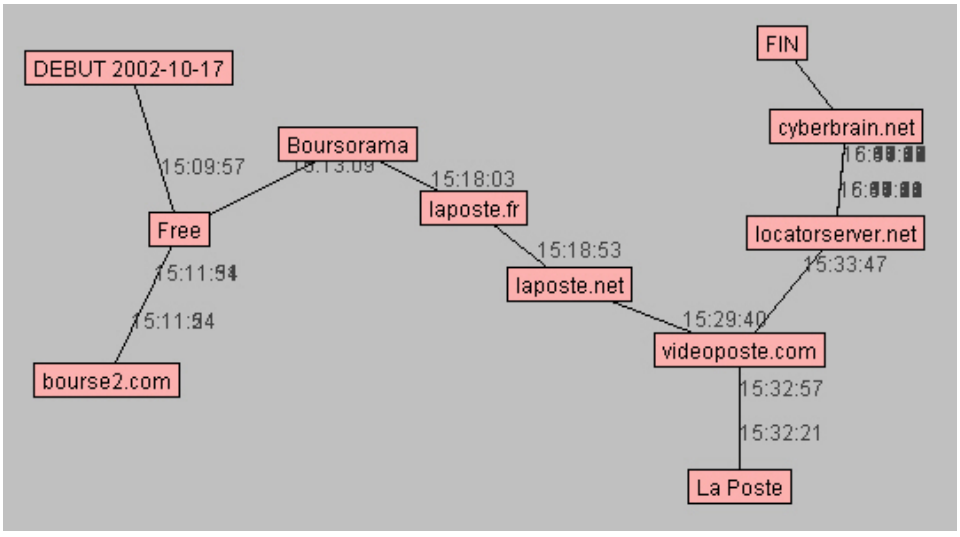


Figure 5.24. SN2002 – exemple typique de la classe « parcours éclatés »

Sur le plan des contenus visités, on retrouve dans ces « parcours éclatés » des contenus très diversifiés, et une diminution de la cohérence thématique globale : comme on a déjà pu l'observer en examinant les contenus des sessions, l'accroissement du nombre de sites visités s'accompagne d'une diversification des thématiques des sessions. En conséquence, aucun thème particulier ne ressort au sein de cette classe, sinon la pratique du WebChat qui implique de faibles taux de linéarité liés au rafraîchissement des pages et à l'allongement de la durée de sessions dans le cadre des échanges interpersonnels. Tout au plus observe-t-on une présence plus marquée des contenus orientés vers les activités ludiques (« Jeux – Consoles » dans MSN, « Sports et loisirs » pour Yahoo, « Sport et détente » chez Nomade, catégorie « pornographie ») d'une part et culturelles (« Culture et loisir » dans Nomade, « Arts – Culture » chez MSN) d'autre part, devançant de peu les activités orientées « vie pratique » impliquant des navigations plus longues, en particulier les catégories liées à la bourse et aux services bancaires.

Comparaison avec les autres jeux de données

Le même travail de classification pratiqué sur les deux autres jeux de données fournit globalement les mêmes résultats, et met en évidence les cinq groupes-types de navigation que nous venons de décrire (voir Figure 5.26 et Figure 5.27 ci-dessous). Ceci étant posé, pour comparer plus finement le positionnement de sessions de BibUsages et de SN00-02 par rapport au panel représentatif de 2002, nous devions travailler sur le même référentiel ; pour cela, nous avons inclus les sessions des panels BibUsages et SN00-02 en tant qu'individus illustratifs dans la classification pratiquée sur les données SN2002, et examiné leur position par rapport aux sessions du panel représentatif. Un premier élément de différenciation concerne la présence des différentes catégories de parcours dans les deux panels (voir Figure 5.25).

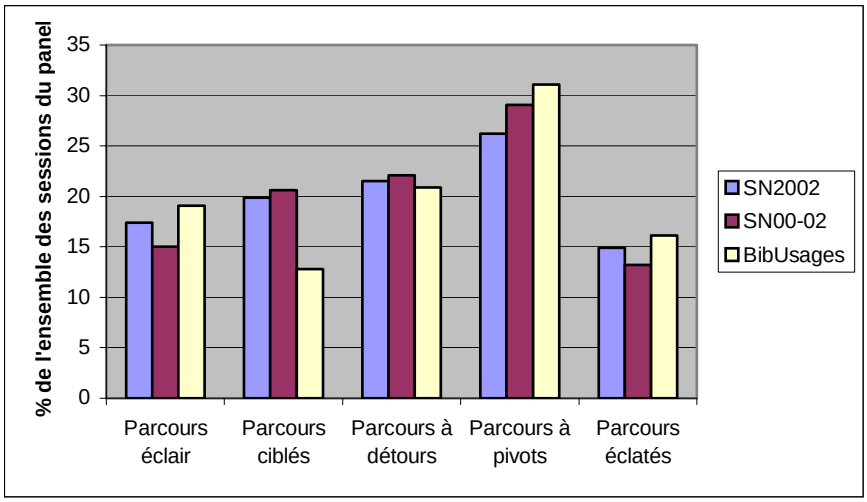


Figure 5.25. Répartition par groupe de sessions pour chaque jeu de données

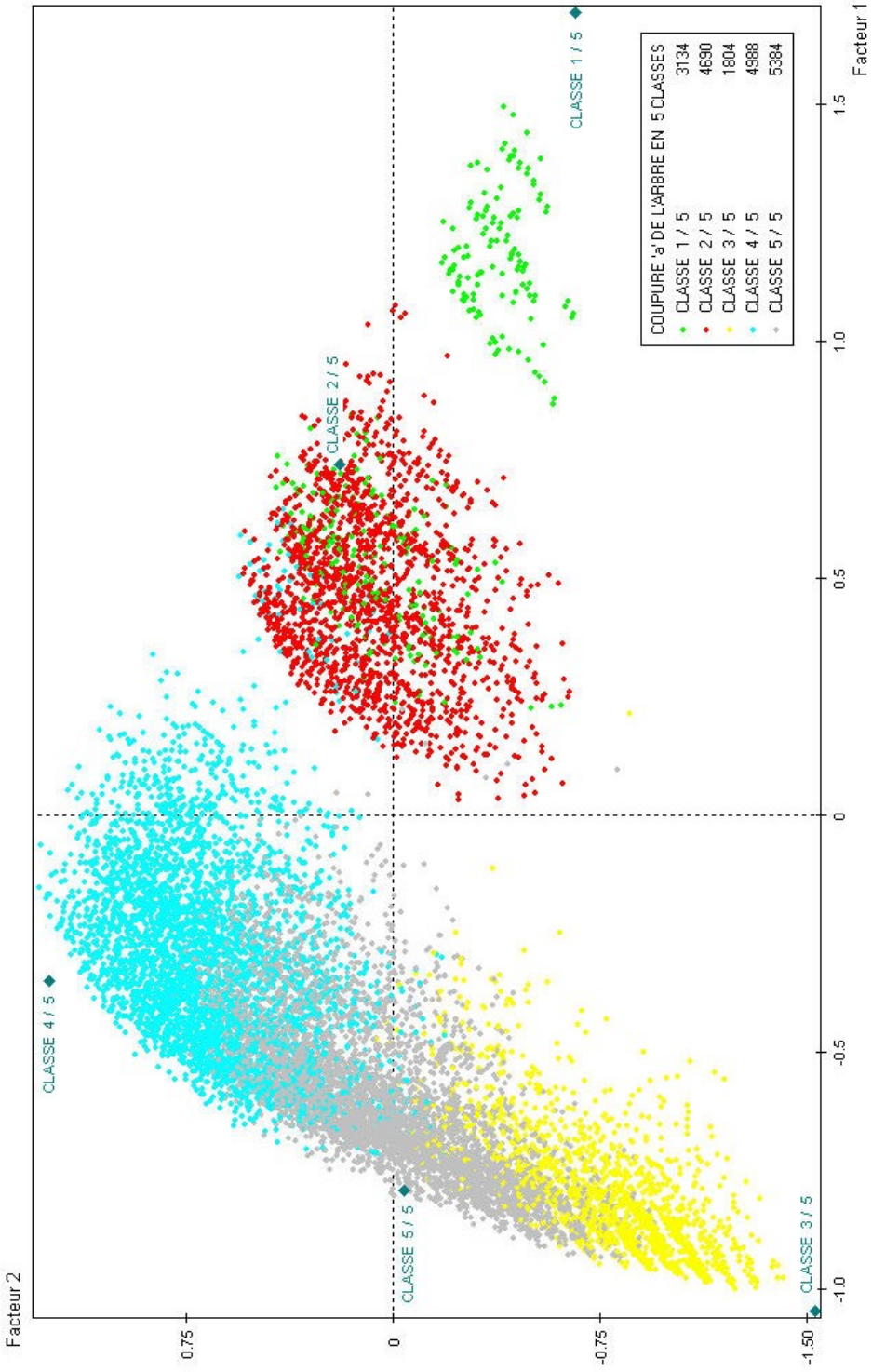


Figure 5.26. Classification sur les indicateurs topologiques, SN00-02 – axes 1 et 2

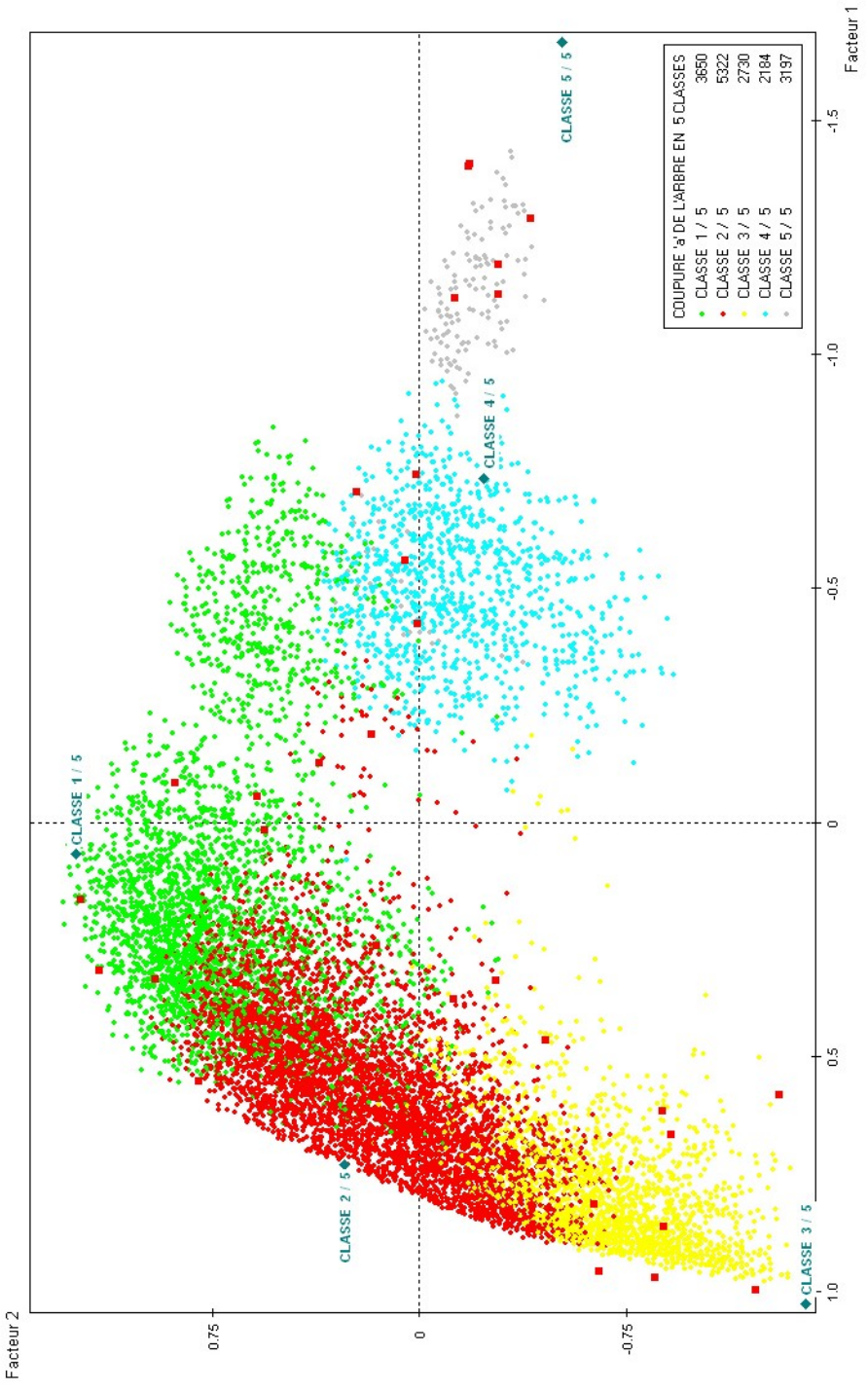


Figure 5.27. Classification sur les indicateurs topologiques, BibUsages – axes 1 et 2

Pour le premier groupe des sessions courtes et rapides, les panélistes de BibUsages paraissent faire preuve de plus d'efficacité que ceux des panels SensNet, avec une part plus importante de parcours éclairs, tandis que les parcours ciblés sont au contraire très peu représentés chez eux (12,8 % des sessions). Il semble que ces utilisateurs avertis vont plus rapidement à l'essentiel lorsqu'ils savent précisément ce qu'ils cherchent. Ceci est corroboré par les entretiens : la plupart des interviewés ont déclaré utiliser les favoris pour classer les sites jugés intéressants, la taille des favoris pouvant varier d'une cinquantaine d'adresses à plus d'un millier. Du côté du panel SN00-02, c'est un mouvement inverse qui se produit, les parcours ciblés étant plus fréquents que les parcours éclairs : la différence avec le panel BibUsages semble indiquer que ce n'est pas l'ancienneté de la pratique qui explique l'efficacité, mais plutôt l'intensité : dans le panel 2000-2002, nous avons pu voir qu'une part non négligeable des internautes sont de faibles utilisateurs, tandis que tous les panélistes de BibUsages sont des usagers réguliers et intensifs du Web.

L'effet d'ancienneté de la pratique semble plutôt paraître au niveau des sessions plus longues : au sein des trois groupes relevant de parcours complexes, c'est dans les parcours à pivot qu'on note les différences les plus notables. Ce type de navigation touche 26,2 % des sessions du panel généraliste 2002, contre 29,1 % des sessions de SN00-02, et 31,1 % de celles de BibUsages. Ce mode de navigation, fortement lié à l'activité de recherche sur le Web à l'aide de moteurs ou de pages-ressources semble attester une maîtrise des outils de recherche : l'utilisateur feuillette les différentes pages dans une logique d'épuisement et de tri de l'offre pour trouver la plus pertinente dans le contexte de sa recherche.

Au cours des entretiens, on a pu voir que cette navigation efficace est opposée par les interviewés à une logique de parcours plus exploratoire et éparse, et qu'elle lui est souvent privilégiée :

> *Je suis un picoreur mais un picoreur qui sait ce qu'il veut. C'est-à-dire que j'essaie de ne pas me, comment dire, de ne pas trop me disperser. Parce que moi, je suis pris par le côté utilitaire ; je suis pas un vagabond, j'aimerais bien mais je n'ai pas le temps. Je me sers d'Internet, en fait comme d'un outil. C'est un outil, c'est un outil, bon ça peut être des fois un outil culturel, donc c'est pour s'amuser, et c'est surtout un outil pour faire des choses, pour trouver de la documentation. (Utilisateur F)*

> *De temps en temps, c'est vrai que je papillonne sur le Web, et c'est vrai que je le fais de moins en moins souvent parce que j'ai moins de temps. […] Je peux être par exemple soit sur le site, par exemple Figaro ou Le Monde ou Libé et en fonction, par exemple, sur le site de Libé, il y a les, les fameux portraits qui sont en dernière page du quotidien et c'est vrai que de temps en temps y a un nom de personnage du portrait qui peut m'intéresser, je peux aller voir son portrait. Ça me renvoie sur une idée et c'est vrai qu'alors soit je note le concept, ou un thème associé et je peux renvoyer sur Google. C'est vraiment de l'hypertexte. (Utilisateur K)*

Le butinage n'est pas rejeté en tant que tel, mais il nécessite un investissement plus important dans la durée, que la plupart ne souhaitent pas assumer : « Je suis pas surfeur, si vous voulez, j'ai pas le temps. » (Utilisateur J).

Pour confirmer ces hypothèses, il est nécessaire de se placer au niveau de l'utilisateur : nous avons envisagé pour l'instant les sessions de manière globale pour chaque panel, de sorte que les utilisateurs intensifs, qui font plus de sessions, pèsent plus lourd dans la représentation finale. En définitive, ces modes de navigation sont-ils liés à l'intensité de la pratique et à l'expertise qui en découle, ou sont-ils partagés par l'ensemble des utilisateurs et déterminés par le contexte local de la tâche ? C'est en examinant la place de chaque type de session pour chaque utilisateur que l'on peut replacer la navigation dans le contexte d'actions normées et examiner la place des déterminations locales et globales dans la morphologie des parcours sur le Web.

> ***Synthèse****. La classification des sessions SN2002 sur la base des indicateurs topologiques fait ressortir cinq parcours-types bien différenciés. D'un côté, parcours éclairs et parcours ciblés forment un groupe homogène de sessions courtes, linéaires ou quasi-linéaires, essentiellement tournées vers les portails généralistes et le WebMail. De l'autre, parcours à détours, parcours à pivots et parcours éclatés suivent une gradation dans la complexité de la navigation, et renvoient à trois contextes d'usage différenciés. Les premiers sont liés aux contenus orientés « vie pratique » et vie hors du Web, et leur linéarité essentiellement rompue par des mouvements courts de* back *; les seconds sont plus apparentés à l'usage de moteurs pour des recherches ouvertes, où certaines pages servent de pivot à la navigation et l'exploration de la Toile ; les derniers, les plus longs et les plus complexes structurellement, sont liés notamment à certains contenus orientés vers les jeux et la communication (WebChat notamment). La projection des sessions des panélistes BibUsages sur ce panorama représentatif des sessions montre la spécificité de ces internautes, plus orientés vers les parcours de recherche et les parcours ciblés.*

Conclusion

Le travail mené sur des données de trafic montre en premier lieu la validité des outils et méthodes de description du contenu et de la forme des parcours que nous avons élaborés. Les premières analyses statistiques sur des données volumineuses et représentatives des usages résidentiels d'Internet en 2002 montrent la grande diversité des comportements de navigation : variété dans les durées, le nombre de sites visités, la complexité des formes de parcours, les thèmes et services accédés, etc. Il en résulte un objet statistique dont la complexité est le reflet de l'inscription des parcours dans des contextes et des pratiques très diversifiés ; ceci justifie une approche par le « geste » plutôt que par le contenu, c'est-à-dire une segmentation des parcours sur la base de leur topologie en première approche. Pour mener à bien cette segmentation, il a été nécessaire de travailler au plus près des indicateurs statistiques et d'opérer des discrétisations *ad hoc*, dans la mesure où les indicateurs que nous

avons construits prennent des significations particulières et entretiennent des corrélations spécifiques pour certaines valeurs.

Ce travail statistique fin nous amène à construire une typologie des parcours Web en cinq classes sur la base de leur forme et de leur temporalité qui rend compte de la diversité des modes de navigation, et des contextes d'usage : les sessions courtes et rapides s'apparentent à des pratiques ciblées où l'utilisateur sait où il va, et s'opposent à des parcours plus diversifiés et plus complexes. Ces éléments morphologiques ne déterminent pas complètement les contenus, mais y sont fortement liés : un comportement exploratoire le restera quel que soit le thème de recherche, mais implique le recours aux outils de recherche et à certains types de pages-ressources qui influencent la forme générale du parcours. Dans les navigations routinières, au contraire, on retrouve des services comme le WebMail ou les informations, qui impliquent une activité répétée et régulière au sein d'un hypertexte connu et maîtrisé. La mise en relation de ces comportements avec le contexte de l'utilisateur en termes d'intensité de pratique et de « territoires personnels » sur le Web doit permettre d'aller plus loin et d'expliquer ces comportements en regard des pratiques individuelles.

Chapitre 6

Navigation en contexte

Nous avons identifié cinq profils-type de sessions sur la base de leur forme et de leur temporalité, et constaté que chacun d'entre eux est corrélé à certains types de contenus ; pour comprendre ces différences, nous devons maintenant replonger les parcours au sein des pratiques individuelles. Cette mise en contexte de la visite de sites et de pages éclaire les parcours sous l'angle des territoires personnels des internautes : elle permet de mettre à jour les activités normées au niveau individuel, et les modes d'appréhension prototypiques au niveau collectif. Elle nous amène également à voir comment les contenus sont appréhendés en situation et leur sens est construit au sein des parcours.

6.1 La session à l'aune de l'utilisateur

Les études sur les comportements de navigation en situation de recherche d'information ont montré la double primauté de l'expertise de l'utilisateur dans la manipulation des outils de recherche, et de la connaissance du domaine de recherche dans la topologie des parcours sur le Web (par exemple : [Jenkins *et al.* 2003]). Même si notre étude dépasse le cadre de la recherche sur le Web pour considérer les activités de navigation dans leur ensemble, on cherchera légitimement à expliquer les différences constatées entre les sessions en les rattachant à l'individu et à ses pratiques.

6.1.1 Profils d'usages et profils de sessions

Types de panélistes, types de sessions ?

Dans [Catledge & Pitkow 1995], Catledge et Pitkow scindent leur panel de cent sept utilisateurs en trois classes, basées sur la longueur et la fréquence de séquences de navigation : « serendipitous browsers », « general purpose browsers » et « searchers ». [Ghitalla *et al.* 2003] observent également parmi les vingt-trois internautes qu'ils étudient des stratégies de navigation bien distinctes, qui impliquent

des différences dans la manipulation des interfaces autant que dans les sites accédés et les ressources mobilisées au sein de recherches documentaires.

On s'attendra à bon droit à trouver dans nos données des éléments similaires. Un exemple sur sept panélistes tirés au hasard dans les données SN2002 va d'ailleurs dans ce sens : la part, dans l'ensemble des sessions de chaque panéliste, de chacun des cinq types de parcours que nous avons identifiés montre des différences notables entre les individus (voir Figure 6.1). Pour le panéliste A, la répartition est assez équilibrée à l'exception des parcours éclatés qui sont très minoritaires ; pour l'utilisateur C, au contraire, parcours à pivots et parcours éclatés représentent les trois quarts des sessions.

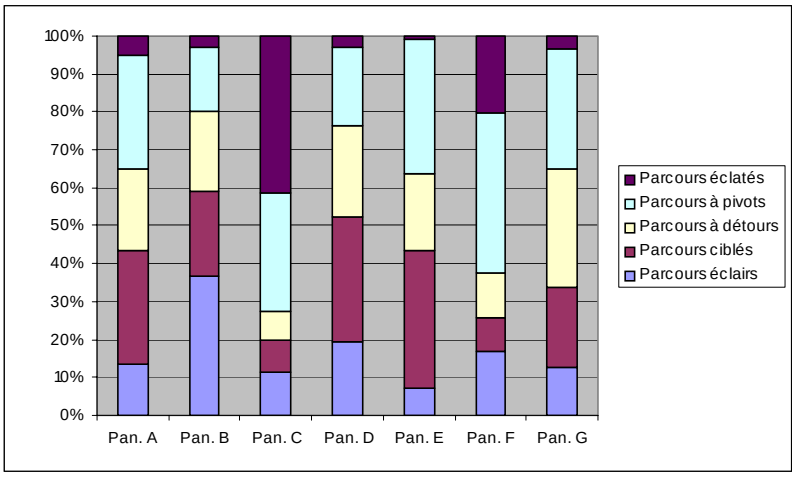


Figure 6.1. SN2002 – Part des différents types de sessions pour sept panélistes

De manière générale, à l'exception des très faibles utilisateurs qui font moins d'une dizaine de sessions sur la période d'observation, aucun des panélistes observés dans les trois jeux de données n'a de comportement exclusivement dédié à tel ou tel type de parcours. Cependant, certains types de comportements de navigation sont comparativement plus présents chez certains : pour le panel SN2002, par exemple, les parcours éclatés représentent en moyenne 13,3 % des sessions d'un internaute, mais pour un quart des individus, ils forment moins de 2 % des sessions (voir Tableau 6.1).

Tableau 6.1. Répartition moyenne des types de parcours pour les panélistes SN2002

	Part des parcours éclairs	Part des parcours ciblés	Part des parcours à détours	Part des parcours à pivots	Part des parcours éclatés
Moyenne	18,0 %	20,4 %	21,8 %	26,5 %	13,3 %
Quartile 1/2	7,3 %	8,9 %	13,5 %	16,7 %	1,9 %
Quartile 2/3	15,0 %	17,5 %	20,8 %	25,3 %	8,0 %
Quartile 3/4	24,6 %	28,1 %	28,6 %	34,0 %	18,7 %

Clef de lecture : les panélistes de SN2002 ont en moyenne 18 % de parcours éclairs ; pour un quart d'entre eux ces parcours moins de 7,3 % de leurs sessions.

Nous avons donc pratiqué une classification des internautes sur la base de la part de chaque type de parcours dans l'ensemble de leur corpus de sessions, afin de voir si des profils d'internautes peuvent être dégagés à partir des profils de sessions. Les résultats de la classification montrent des profils relativement différenciés (voir Figure 6.2 ci-dessous). On ne s'étonnera pas de ce que la partition en six classes ainsi obtenue recoupe globalement les types de sessions ; deux éléments intéressants sont ici à noter. En premier lieu, alors qu'on avait vu s'opposer dans l'examen global des sessions un groupe de sessions courtes et simples (parcours éclairs et ciblés) à un ensemble plus complexe formé des trois autres types de sessions, on constate, en ramenant les sessions aux individus, que trois pôles principaux émergent :

- d'un côté, un ensemble « parcours éclairs / ciblés »
- directement opposé à celui-ci, un groupe « parcours éclatés / à pivots »
- le groupe des parcours à détours, qui se pose comme spécifique par rapport aux deux autres.

D'autre part, l'examen des effectifs relatifs à chaque classe et à chacun de ces trois groupes montre que s'il existe certains panélistes « spécialisés » dans des comportements de navigation particuliers, un peu moins de la moitié des individus se rattache à une classe centrale et affecte un comportement médian.

Ce dernier constat relativise l'affirmation selon laquelle les internautes se comportent chacun de manière très particulière dans leurs modes de navigation : si pour une moitié d'entre eux une certaine forme de spécialisation dans un type de navigation est observable, pour l'autre moitié, les comportements de navigation sur le Web se répartissent globalement en sessions longues et courtes, simples et complexes, structurées ou décousues.

Types de sessions et types d'usages d'Internet

Pour éclairer cette relative concordance entre profils de sessions et profils d'internautes, on tentera de croiser la répartition des types de sessions de chaque panéliste avec des variables relatives à l'utilisateur. On laissera de côté les éléments socio-démographiques : on imagine difficilement dire qu'un cadre fait plus de parcours ciblés qu'un artisan – tout du moins une telle différence, si elle devait être observée, n'est pas explicative. Les variables relatives aux usages d'Internet sont plus pertinentes, en ce qu'elles positionnent les situations de navigations dans un contexte global d'usage de l'outil : l'ancienneté de la pratique renvoie à l'appropriation des TIC et pourrait expliquer l'apparition de routines et de navigations plus ciblées ou plus structurées autour de ressources identifiées ; le profil général d'usage d'Internet que nous avons présenté au Chapitre 5, qui croise intensité d'utilisation et diversité des services utilisés, peut également éclairer les comportements de navigation ; enfin, le type d'abonnement, notamment en termes de débit, peut s'avérer très structurant pour les types de sessions pratiqués.

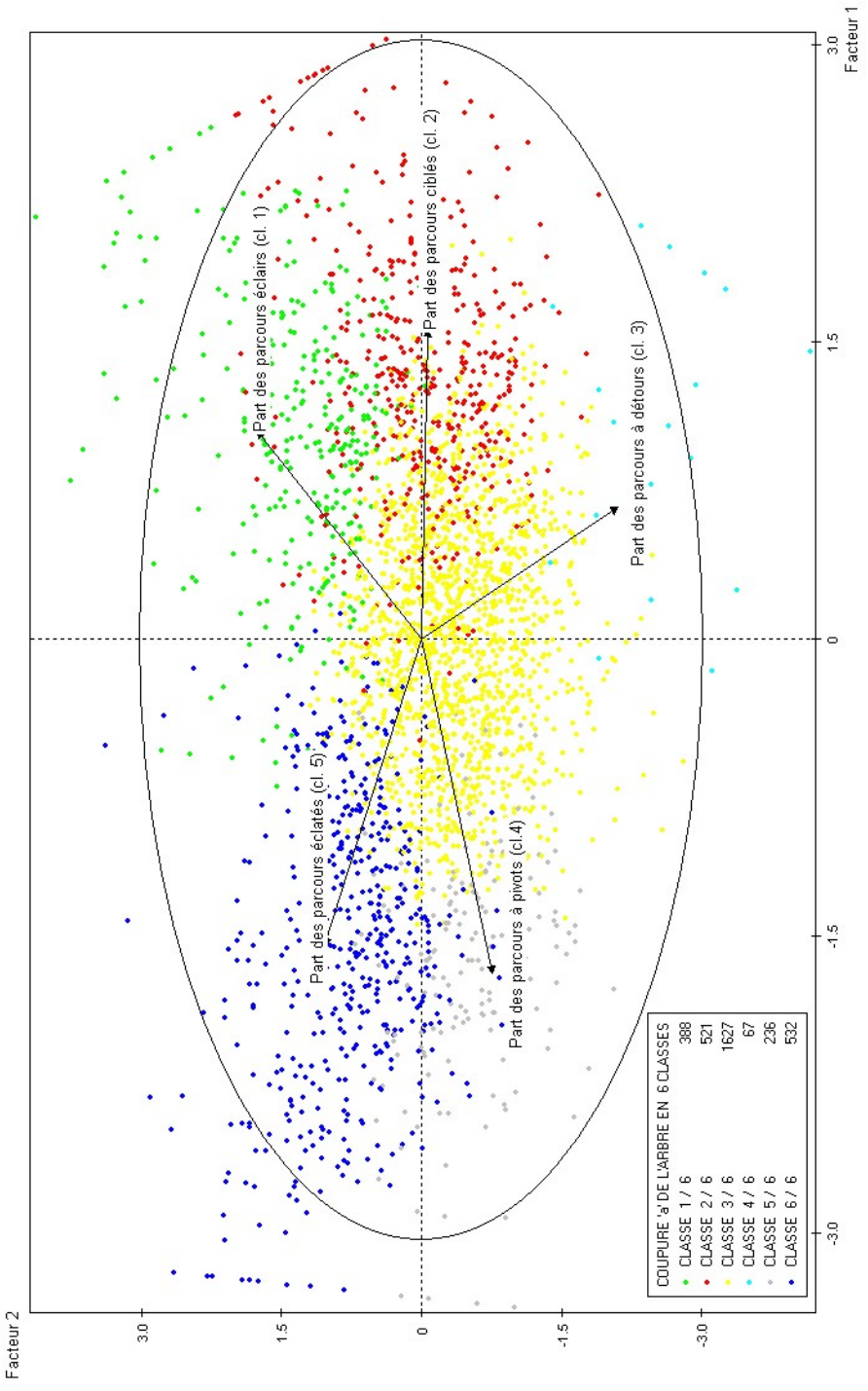


Figure 6.2. SN2002 – Classification des panélistes en fonction des types de sessions réalisés

La confrontation de ces variables et de la typologie des sessions vient pourtant démonter ces hypothèses. L'ancienneté de la pratique, tout d'abord, ne semble que très marginalement liée aux types de sessions (voir Tableau 6.2 ci-dessous) : les différents groupes d'internautes du panel SN2002, anciens ou récents, ont en moyenne la même répartition des types de parcours sur les dix mois d'observation. On notera tout au plus une présence décroissante des parcours éclatés à mesure que la pratique est ancienne (11 % chez les plus anciens, contre 15 % chez les plus récents), tandis que les parcours éclairs sont au contraire sont plus faibles chez les nouveaux internautes (16 % des sessions en moyenne) que chez les plus anciens (20 % en moyenne). Ce léger effet d'apprentissage ne suffit pourtant pas à expliquer la diversité des types de sessions, et l'on conclura à l'indépendance globale des deux variables.

Tableau 6.2. SN2002 – Ancienneté de la pratique et types de parcours

Première connexion	Part des parcours éclairs	Part des parcours ciblés	Part des parcours à détours	Part des parcours à pivots	Part des parcours éclatés	*Total*
Avant 1997	20%	22%	22%	26%	11%	100 %
1997	19%	22%	23%	25%	12%	100 %
1998	18%	20%	23%	26%	13%	100 %
1999	18%	21%	22%	26%	13%	100 %
2000	18%	20%	21%	27%	15%	100 %
2001	16%	20%	22%	27%	15%	100 %

Clef de lecture : chez les panélistes connectés pour la première fois avant 1997, 20 % des sessions en moyenne sont du type « parcours éclairs ».

Il en va de même pour le type de connexion dont est équipé l'internaute (voir Tableau 6.3) : entre ceux équipés d'un modem RTC classique et ceux disposant d'un accès à haut débit (ADSL, Câble), on ne note pas de différence particulière dans la répartition des cinq classes de sessions.

Tableau 6.3. SN2002 – Type de parcours et type de connexion (répartition moyenne)

Type de connexion	Parcours éclairs	Parcours ciblés	Parcours à détours	Parcours à pivots	Parcours éclatés	Total
Bas débit (3035 ind.)	17,7 %	20,8 %	21,9 %	26,4 %	13,2 %	100 %
Haut débit (240 ind.)	20,6 %	17,2 %	20,5 %	27,7 %	14,0 %	100 %
NSP (96 ind.)	21,8 %	16,3 %	20,7 %	25,6 %	15,7 %	100 %

Le croisement entre types de sessions et typologie d'usages d'Internet amène à une conclusion similaire : ici encore, les différents parcours-type sont répartis de manière équivalente à travers les quatre groupes d'usagers que nous avons identifiés (voir Tableau 6.4). On notera, certes, que les usagers « ludiques » ont un profil quelque peu différent des trois autres groupes, avec une part plus importante de parcours éclatés et à pivot, au détriment des parcours ciblés : chez ces internautes intensifs, l'augmentation du nombre de sessions se fait au profit de parcours longs et complexes.

Tableau 6.4. SN2002 – Usages d'Internet et types de parcours

Types d'usages d'Internet	Part des parcours éclairs	Part des parcours ciblés	Part des parcours à détours	Part des parcours à pivots	Part des parcours éclatés	*Total*
Usages occasionnels	18 %	22 %	23 %	26 %	11 %	100 %
Web ordinaire	20 %	21 %	21 %	26 %	12 %	100 %
Comm. interpersonnelle	15 %	20 %	22 %	27 %	16 %	100 %
Usages ludiques	18 %	15 %	20 %	29 %	18 %	100 %

Clef de lecture : dans le groupe des internautes tournés vers les « usages ludiques », 18 % des sessions en moyenne sont du type « parcours éclairs ».

Ces éléments nous amènent à conclure à des déterminations beaucoup plus locales pour expliquer les différents profils de sessions : ce n'est pas tant le profil général de l'utilisateur, son intensité d'utilisation, sa maîtrise des interfaces ou sa connaissance des contenus et services disponibles sur le Web qui déterminent son comportement de navigation, mais plutôt les types d'activités dont la navigation est le support, et la connaissance locale des sites visités. C'est ainsi que nous interprétons les différences observables entre nos trois jeux de données (voir Tableau 6.5 ci-dessous) : les « gallicanautes » se différencient fortement des panels généralistes, avec une sur-représentation des parcours à pivot (33 % en moyenne, contre 26 % pour SN2002) au détriment des parcours ciblés (11 % pour BibUsages, contre 20 % pour SN2002).

Tableau 6.5. Part moyenne des types de sessions chez les panélistes pour chaque panel

	Part des parcours éclairs	Part des parcours ciblés	Part des parcours à détours	Part des parcours à pivots	Part des parcours éclatés
BibUsages	19 %	11 %	21 %	33 %	17 %
SN00-02	16 %	21 %	22 %	29 %	12 %
SN2002	18 %	20 %	22 %	26 %	13 %

Cette opposition est à mettre en parallèle avec les différences que nous avons déjà soulignées au cours du Chapitre 5 : les sessions plus courtes, plus denses et plus rythmées observées pour le panel BibUsages nous amenaient à conclure à une plus grande efficacité dans la navigation, que l'on était tenté de relier à un effet de l'ancienneté de la pratique. Sur ce dernier point, l'examen de l'ancienneté dans les données SensNet en 2002 montre que ce n'est pas la bonne hypothèse. On expliquera plutôt par les tâches effectuées localement la spécificité du panel BibUsages : panel mixte résidentiel/professionnel, composé essentiellement de chercheurs, amateurs ou professionnels, qui utilisent intensivement – mais pas exclusivement – le Web comme une ressource documentaire. Les correspondances observées entre forme de parcours et contenus visités vont également dans ce sens : si l'on ne peut attacher une thématique particulière à un type de parcours, on constate que certains types de services et d'outils sont sur-représentés dans certains contextes de navigation.

Dès lors, la cohérence des sessions d'un même individu tient plus à la variété des situations d'usage dans lesquelles il se trouve : l'un va consulter sa messagerie à

chaque session, l'autre n'utilisera le Web que pour rechercher des documents sur un thème précis, un troisième sera majoritairement tourné vers le WebChat. La notion de territoire revêt alors une importance particulière pour comprendre les parcours sur la Toile ; rattaché à un utilisateur donné, le territoire s'inscrit dans l'histoire de ses pratiques et de ses savoir-faire, et c'est dans ce double contexte que se déploie le sens des activités de navigation sur le Web.

> ***Synthèse.*** *À l'exception des très faibles utilisateurs, aucun des panélistes observés dans les trois jeux de données n'a de comportement exclusivement dédié à tel ou tel parcours-type. Certains types de comportements de navigation sont comparativement plus présents chez certains internautes, mais ni la segmentation des utilisateurs sur la base de l'usage général d'Internet, ni les variables d'ancienneté ou de type de connexion ne permettent d'expliquer ces différences.*

6.1.2 Territoires sur le Web

Les données de trafic ne nous permettent pas de reconstruire véritablement les tâches des utilisateurs : on peut tout au plus supposer, en examinant manuellement les données, qu'un utilisateur est dans telle ou telle configuration, mais son intention, et corrélativement la grille interprétative qui guide son appréhension des contenus visités, nous échappent. Par contre, ces données centrées-utilisateur nous permettent de connaître les sites visités par chaque panéliste sur l'ensemble de la période et le contexte dans lequel chacun apparaît. On peut ainsi dresser pour chaque internaute une cartographie des ressources qu'il mobilise sur le Web, et observer sur la durée l'envergure et la structure de ce territoire personnel. Pour cela, le panel SN00-02 nous apporte des données inédites de par la taille de l'échantillon autant que la durée d'observation, et c'est principalement sur ces données que nous allons nous baser pour approfondir cette problématique.

Intensités variables, habitudes constantes

L'ensemble des 597 personnes du panel SN00-02 a consulté au cours des 34 mois d'observation 192 000 sites/portails différents. Pour autant, tous ces sites n'ont pas été vus par l'ensemble du panel, loin s'en faut (voir Tableau 6.6) : les deux tiers des sites n'ont été visités que par un seul internaute du panel, tandis qu'un autre quart l'a été par 2 à 5 panélistes différents.

Tableau 6.6. SN00-02, nombre de panélistes différents par site/portail

	Nombre de sites	Part des sites
1 visiteur	127 357	66,3 %
2 à 5	49 446	25,8 %
6 à 99	14 818	7,7 %
Plus de 100	378	0,2 %

Au total, seule une trentaine de sites ont été vus par plus de la moitié du panel, présentés au Tableau 6.7 ci-dessous : on retrouve dans cette liste les grands portails généralistes et les fournisseurs d'accès (MSN, Wanadoo, Yahoo, Club-internet) qui

s'imposent comme des carrefours et des points de passage pour aller « ailleurs » sur le Web (*via* les moteurs de recherche en particulier), et qui proposent en outre des services de communication qui attirent un grand nombre d'utilisateurs et les fidélisent. Par ailleurs, deux autres catégories de sites sont ici représentés : d'une part les sites proposant des outils nécessaires à la navigation (`www.real.com`, et `www.macromedia.com`, qui proposent des *plugins* multimédia souvent indispensables à la visite de certains sites ou de certains contenus), et d'autre part des sites de vente ou de services en ligne (Fnac, Alapage, SNCF, Pages Jaunes).

Tableau 6.7. SN00-02 – sites/portail vus par plus de la moitié du panel de 597 individus au cours des 34 mois

Site	Nb pan.
Wanadoo (fr)	553
Yahoo (fr)	544
microsoft.com	543
MSN	539
multimania.fr	528
Voila (fr)	522
Yahoo (com)	521
Lycos	515
ifrance.com	480
real.com	465
Club-internet	464
macromedia.com	456
Google	443
Nomade	407
Amazon	406
pagesjaunes.fr	400
Altavista (com)	396

Site	Nb pan.
SNCF	385
ibazar.fr	361
chez.com	357
Fnac	356
passport.com	356
online.fr	344
libertysurf.fr	336
wedoo.com	324
Altavista (fr)	313
tiscali.fr	313
aol.com	311
windowsmedia.com	309
swisstools.net	308
aol.fr	305
Alapage	303
TF1	300

L'audience des sites Web apparaît ainsi très éclatée : hormis les quelques points de passage communs à la majorité des internautes, chacun semble aller chercher sur le Web les contenus qui l'intéressent spécifiquement. Ceci ne doit pas nous surprendre outre mesure : en n'observant que 597 personnes, fût-ce durant 34 mois, nous avons peu de chances que l'hétérogénéité du panel en termes de centres d'intérêt amènent ses individus à visiter les mêmes contenus.

Le même calcul appliqué aux données BibUsages le confirme : ici, si la part des sites vus par un seul visiteur est de 81 % (contre 67 % pour SN00-02), ce sont 66 sites qui ont été vus par plus de la moitié du panel. Ce résultat est d'autant plus important qu'avec ses 76 utilisateurs, le panel BibUsages est bien plus petit, et que la durée d'observation n'est que de six mois. Par contre, il est beaucoup plus homogène : constitué par des internautes intéressés par les textes et les contenus culturels en ligne, il montre qu'une communauté d'intérêt se retrouve autour de certains sites particuliers et investit des territoires similaires.

À l'échelle de l'utilisateur, le nombre de sites et de portails différents visités par chaque individu du panel SN00-02 au cours des 34 mois d'observation est très variable, allant de 11 pour le plus faible utilisateur à 8 900 pour le plus intensif, avec une moyenne globale de 943 sites (médiane : 569). Pour autant, tous les sites d'un

utilisateur ne sont pas vus de manière égale : l'examen pour chaque panéliste de la répétition des sites dans ses différentes sessions (voir Tableau 6.8) montre qu'en moyenne, les trois quarts des sites visités par les panélistes SN00-02 au cours des 34 mois d'observation n'apparaissent que dans une seule session.

Tableau 6.8. SN00-02 – Répartition des sites visités en fonction du nombre de sessions où ils apparaissent chez un panéliste

	1 session	2-4 sessions	5-6 sessions	10-99 sessions	Plus de 100 sess.
Moyenne	74,4 %	18,1 %	3,9 %	3,3 %	0,3 %
Médiane	74,5 %	18,3 %	3,8 %	3,1 %	0,2 %
Minimum	50,8 %	0,0 %	0,0 %	0,0 %	0,0 %
Maximum	100,0 %	40,0 %	12,1 %	12,3 %	6,2 %

Clef de lecture : en moyenne, 74,4 % des sites vus par un panéliste donné ne sont vus que dans une seule session, 18,1 % dans 2 à 4 sessions.

Cette dispersion des sites visités est-elle liée à l'intensité d'usage de chaque panéliste ? En d'autres termes, les utilisateurs intensifs du Web ont-ils tendance à voir plus de nouveaux sites, à avoir des pratiques de « surf » plus fréquentes par rapport aux utilisateurs modérés, cantonnés à quelques sites bien définis ? Le Tableau 6.9 scinde, par quartile du nombre total de sessions, la fréquence d'apparition des sites de chaque panéliste dans ses sessions sur les 34 mois, et permet de voir si les utilisateurs intensifs ont plus d'hapax dans leur corpus de sites que les autres.

Tableau 6.9. SN00-02 – Intensité d'usage et présence des sites visités dans les sessions

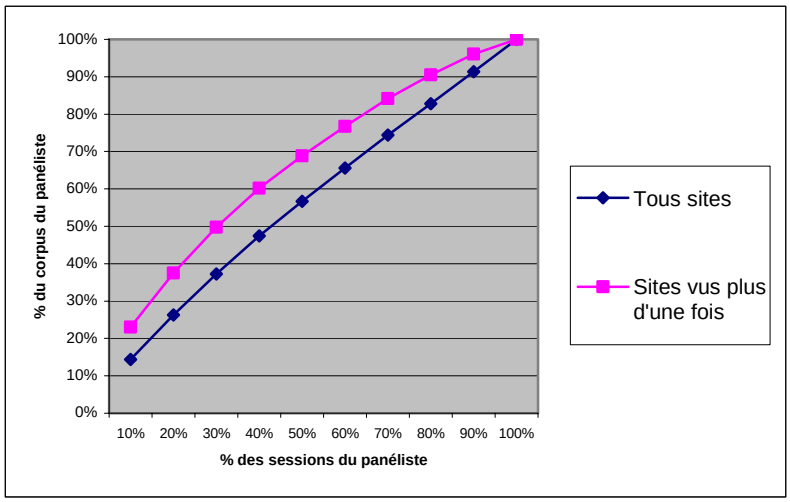
Quartile du nombre de sessions total	Sites vus dans 1 session du pan.	Sites vus dans plusieurs sessions du pan.			*Total*
		0 à 19 % de ses sessions	20 à 39 % de ses sessions	40 à 100 % de ses sessions	
1er quartile	78,0 %	18,8 %	1,7 %	1,5 %	100 %
2e quartile	75,7 %	23,3 %	0,5 %	0,5 %	100 %
3e quartile	73,7 %	25,8 %	0,3 %	0,2 %	100 %
4e quartile	71,3 %	28,4 %	0,8 %	0,1 %	100 %

Clef de lecture : chez les faibles utilisateurs (1er quartile), 78 % des sites d'un utilisateur n'apparaissent que dans une session en moyenne, et 1,5 % de sites sont présents dans plus de 40 % de ses sessions sur les 34 mois.

Malgré une légère sur-représentation des sites présents dans une seule session chez les faibles utilisateurs (78 %), et l'inverse chez les intensifs (71 %), les différences nous semblent trop faibles d'un groupe à l'autre pour que l'on puisse conclure à un lien direct entre intensité d'usage et concentration des sites visités sur le Web. Tout au plus pourra-t-on remarquer que les internautes ayant effectué le plus grand nombre de sessions semblent, paradoxalement, être ceux qui revisitent le plus de sites, alors que l'on pouvait supposer une fréquentation plus « explosée » du Web de leur part. Aussi suggèrera-t-on que les utilisateurs intensifs connaissent mieux les ressources sur le Web, utilisent mieux les outils de recherche et ciblent mieux les contenus qui les intéressent, ce qui les amène à moins d'errements et moins de sites vus une seule fois.

Un corpus en constante expansion

Sur cette base, on pourrait faire l'hypothèse que chacun a sur le Web des territoires bien déterminés, et qu'une fois passée une phase de découverte des ressources intéressantes pour lui, son « corpus » de sites se stabilise autour de quelques-uns. Dans les faits, le corpus de sites de chaque panéliste est en constante augmentation au cours des 34 mois d'observation : en moyenne, lorsqu'il a effectué la moitié des sessions observées sur la période, ce sont seulement 57 % de son corpus de sites qu'il a exploré (voir Figure 6.3), et à 90 % des sessions observées, il a vu 91 % seulement de son corpus de sites.



Clef de lecture : un panéliste ayant effectué 20 % des sessions observées sur les 34 mois a visité en moyenne 26 % de l'ensemble des différents sites qu'il voit sur la période.

Figure 6.3. SN00-02 – Évolution du nombre de sites différents visités au fil des sessions

Ce lien quasi linéaire entre nombre de sessions et nombre de sites différents visités indique qu'au fil des sessions, l'internaute continue à explorer de nouveaux sites, au premier comme au dernier mois d'observation : la part importante des sites vus dans une seule session ne peut donc pas être imputée à une période initiale de découverte des ressources Web, mais semble faire partie intégrante des pratiques de navigation au quotidien.

Si l'on restreint l'analyse aux sites vus dans deux sessions différentes au moins, on cerne des sites qui sont d'un certain intérêt pour l'internaute, et qui peuvent entrer dans le cadre de son espace familier et connu : on s'attendra plutôt à ce que ces sites soient tous vus dans une période plus ou moins courte, et qu'au début des 34 mois d'observation, on en ait balayé une bonne partie. Ici encore, les données viennent démentir cette approche intuitive : comme le montre la Figure 6.3, au fil des sessions de l'internaute, celui-ci découvre de nouveaux sites sur lesquels il va revenir par la suite, et à la moitié des sessions observées sur la période, ce sont seulement 70 % des sites vus plus d'une fois qui ont été visités.

En définitive, ce n'est pour chaque internaute qu'une poignée de sites qui suscitent une réelle fidélité. Cette notion ne doit pas être confondue avec la

régularité et la présence d'un site dans toutes les sessions du panéliste : il est nécessaire de relier la session à une activité bien précise, et dans ce cadre, la régularité peut être très différente de la fréquence. Un panéliste peut visiter systématiquement le même site pour une tâche précise, cette tâche n'intervenant qu'occasionnellement : la déclaration d'impôts en ligne fournit un exemple typique de cette situation, où l'internaute ne visite le site du Ministère de finances qu'une fois par an, mais peut le faire tous les ans à la même époque. Pour analyser la fidélité aux sites, on examinera alors l'empan d'un site, c'est-à-dire le nombre de jours qui séparent la première et la dernière visite d'un site vu plusieurs fois par le panéliste au cours des 34 mois d'observation : dans un tiers de cas, cette durée ne dépasse pas trente jours, et elle n'est supérieure à un an que pour un quart des observations. Le nombre moyen de sites dont l'empan dépasse une année est de 72 en moyenne par panéliste, avec moins de 16 sites pour les internautes les moins actifs, et plus d'une centaine pour les plus intensifs. L'élément le plus remarquable ici est que, malgré cette diversité dans l'intensité d'usage, le nombre de sites vus plus d'un an est toujours proportionnel au nombre total de sites visités par l'internaute : pour 90 % des internautes du panel SN00-02, ces sites représentent entre 25 et 35 % de l'ensemble des sites visités plusieurs fois sur la période, et 7,5 % de l'ensemble des sites qu'il voit.

En définitive, rares sont les sites qui sont à la fois vus fréquemment et sur une longue période : pour l'ensemble des sites vus plus d'une fois sur les trois ans par un utilisateur donné, on compte en moyenne 83 jours entre deux visites (médiane : 34 jours), et pour un quart des sites, cet écart est de moins d'une semaine. Fréquence et fidélité ne se confondent donc pas, et le territoire sur le Web apparaît délimité en trois zones :

1. les sites qui ne sont vus que dans une seule session, et qui représentent les trois quarts des sites vus par un internaute en moyenne ;
2. les sites vus plus d'une fois, mais qui ne fidélisent pas l'internaute. Ces sites sont actifs sur des courtes et moyennes périodes, 80 jours en moyenne (médiane : 34 jours), et pour la moitié des cas apparaissent dans deux sessions seulement. Ils forment en moyenne 17 % du corpus de sites d'un internaute, et semblent correspondre à des activités passagères pour l'individu, ou à une forme de « rejet après essai ».
3. les sites familiers : vus sur plus d'un an, ils sont visités préférentiellement par l'internaute dans un contexte donné. On retrouve bien évidemment dans cette catégorie les portails généralistes et les moteurs de recherche, ainsi que les services qui fidélisent les utilisateurs comme le WebMail, mais également des sites spécialisés sur tel ou tel sujet, pour lequel l'internaute a un intérêt particulier : `tomshardware.com` (informations sur le matériel informatique), `finances.gouv.fr` (site du Ministère des finances), `allocine.fr` (films et séances de cinéma), etc. Ils n'apparaissent pas systématiquement dans beaucoup de sessions – moins de 5 pour la moitié des cas, mais semblent avoir la préférence de l'utilisateur pour un domaine ou un service donné.

Territoires thématiques

La notion de thématiques et de services préférentiels sur le Web fait écho à la définition de territoires et de lieux habituels en termes de sites visités. Sur la base des

descriptions de contenu que nous avons rattachées aux pages visitées, nous pouvons établir pour chaque utilisateur une cartographie des thèmes et des services auxquels il accède, et examiner la fréquence et la régularité de l'accès à ces contenus dans les sessions.

Pour cela, nous décrivons chaque session par la catégorie *CatService* ou la catégorie d'annuaire de premier niveau sur laquelle l'internaute passe le plus de temps dans la session, comme nous l'avons déjà fait précédemment (voir Chapitre 5). On a pu voir que dans la plupart des cas, cette catégorie majoritaire couvre l'essentiel de la durée des sessions, celles-ci étant plutôt monothématiques ou monofonctionnelles. On dresse ainsi un profil des contenus visités par chaque internaute, correspondant à la part de ses sessions relatives à tel ou tel descripteur de contenu.

L'examen des profils nous révèle que les internautes ont chacun des centres d'intérêt bien délimités sur le Web, qui occupent la majorité de leur vie sur la Toile : l'exemple donné au Tableau 6.10 pour deux internautes du panel SN00-02 montre chez le premier un intérêt très prononcé pour les sites pornographiques et un fort usage des moteurs de recherche ainsi que, plus modérément, des sites de supermarchés en ligne, tandis que le second panéliste est plus orienté vers les sites consacrés au sport et aux loisirs, ainsi qu'au service de WebMail.

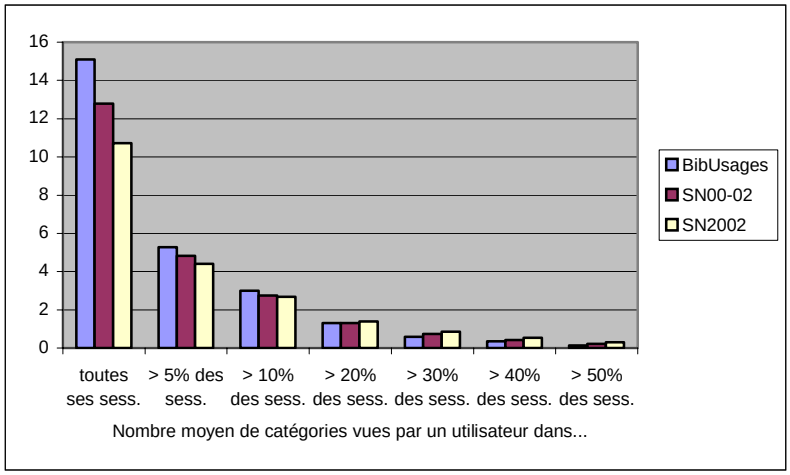
Tableau 6.10. Profil thématique de panélistes – exemples pour SN00-02 avec Nomade[1]

Profil du panéliste 21166	
Catégorie	% sess.
PORNO	39,3 %
Aucune description	19,5 %
Moteur	9,5 %
Mes Courses	6,3 %
Nouvelles technologies	5,8 %
PG - Page Accueil	4,8 %
Espace B to B	4,5 %
WebMail	2,5 %
Société, Vie pratique	2,3 %
Culture et loisirs	1,5 %
Actu, médias	1,3 %
Sport et détente	1,0 %
Voyage, géographie	0,8 %
Éducation, formation	0,3 %
Nature et sciences	0,3 %
PG - Achat	0,3 %
PG - Non catégorisé	0,3 %
PG - Personnalisation	0,3 %
Total	100 %

Profil du panéliste 296	
Catégorie	%sess.
Sport et détente	22,9 %
Aucune description	20,0 %
WebMail	10,0 %
Culture et loisirs	7,1 %
Espace B to B	7,1 %
Mes Courses	7,1 %
Moteur	7,1 %
Actu, médias	4,3 %
Éducation, formation	4,3 %
PG – Personnalisation	2,9 %
Forme et Santé	1,4 %
Nature et sciences	1,4 %
PG - Non catégorisé	1,4 %
Société, Vie pratique	1,4 %
WebChat	1,4 %
Total	100 %

[1] Nous avons appliqué ce calcul à l'ensemble des sessions, d'où une catégorie « Aucune description » forte, correspondant aux sessions mal décrites par les annuaires. Cela étant, les résultats sont similaires si l'on restreint le calcul aux sessions bien décrites.

Un fort effet de concentration est observable pour l'ensemble des internautes des trois panels, et joue à deux niveaux : d'une part, alors que le nombre de catégories descriptives va de 31 à 34 selon l'annuaire considéré, on en compte beaucoup moins pour chaque panéliste pris individuellement. Pour le panel SN2002, ce sont en moyenne 10,7 catégories différentes qui décrivent l'ensemble des sessions de chaque panéliste, 12,8 pour l'échantillon SN00-02 et 15 pour les internautes de BibUsages. D'autre part, sur cette dizaine de catégories auxquelles le panéliste a consacré au moins une session, les deux tiers font l'objet d'une visite très occasionnelle et figurent dans moins de 5 % des sessions, tandis qu'en moyenne une seule catégorie décrit plus de 35 % des sessions du panéliste (voir Figure 6.4).



Clef de lecture : pour chaque internaute de BibUsages, 15 catégories différentes décrivent l'ensemble de ses sessions en moyenne, mais 5 catégories seulement sont représentées dans plus de 5 % des sessions.

Figure 6.4. Nombre de catégories décrivant les sessions des panélistes

Le nombre de catégories différentes sur lesquelles un panéliste porte son intérêt est, assez logiquement, fonction de l'intensité d'usage du panéliste : plus celui-ci réalise de sessions, plus il est amené à voir des sites sur des contenus variables. Pour le panel SN00-02, le quart le moins actif voit au total 5 types de contenus différents en moyenne sur les 34 mois d'observation, tandis que le quart le plus intensif en consulte plus de 18. Pourtant, fait remarquable, le nombre de catégories représentatives de plus de 5 % des sessions est indépendant de l'intensité d'usage et demeure autour de 5 pour l'ensemble du panel (voir Tableau 6.11).

Tableau 6.11. SN00-02 – Intensité d'usage et variété des contenus visités (moyennes)

Quartile du nombre de sessions	Nb total de catégories vues par le panéliste	Nb cats vues dans plus de 5% des sessions	Nb cats vues dans plus de 10% des sessions	Nb cats vues dans plus de 30% des sessions	Nb cats vues dans plus de 50% des sessions
1er quartile	5,5	4,4	2,9	1,0	0,3
2e quartile	11,7	5,2	2,8	0,6	0,2
3e quartile	15,3	5,0	2,7	0,6	0,2
4e quartile	18,6	4,7	2,6	0,7	0,2

Clef de lecture : pour les panélistes les moins actifs, 5,5 catégories décrivent l'ensemble de leurs sessions, mais une seule catégorie en moyenne décrit plus de 30 % des sessions.

Ces chiffres confortent l'idée de la constitution d'un espace personnel sur le Web par les internautes : ce qui était vrai au niveau des sites visités l'est également sous l'angle des contenus. Si un individu peut être amené à visiter des sites sur des thèmes très variés, la majorité de ses pratiques reste centrée sur un nombre limité de services et de thématiques. Ce constat invite à relativiser, ou du moins à préciser, la notion de « surf » telle qu'elle peut être employée couramment : si le Web peut permettre d'accéder à toutes sortes de services, de types de documents, de thématiques, dans les pratiques, cette possibilité n'est pas utilisée à l'échelle macroscopique, chacun se créant les niches relatives à ses centres d'intérêt.

Ces informations sont importantes pour la compréhension des comportements de navigation : ramenées à l'échelle de la session, elles doivent permettre d'éclairer les comportements de navigation à l'aune des pratiques et des territoires individuels.

> ***Synthèse.*** *L'audience des sites Web apparaît ainsi très éclatée : hormis les quelques points de passage communs à la majorité des internautes, chacun se constitue sur le Web un territoire spécifique. La structure de ces territoires personnels peut être scindé en trois zones : les sites vus une fois seulement, qui constituent la grande majorité du corpus de sites de chaque panéliste, et qui ne cessent de l'alimenter au fil du temps ; les sites vus plusieurs fois mais de faible empan, correspondant à un besoin borné dans les temps ; et les sites réguliers, de fort empan, qui constituent le cœur des parcours de l'internaute. Dans cette dernière catégorie, on distingue une poignée de sites qui constitue le territoire familier, et les sites visités moins fréquemment correspondant à des contextes particuliers apparaissant peu souvent. Une structuration similaire est observée sur le plan des contenus : une grande diversité globale se lit dans les pages visitées par chaque panéliste, mais seuls quatre à cinq thèmes et services font l'objet de visites régulières et fréquentes.*

6.2 Sessions en contexte

Pour chaque parcours, on peut maintenant replacer les sites visités dans le corpus général du panéliste d'une part, et les thèmes et services accédés dans la cartographie des contenus habituels ou occasionnels de l'internaute d'autre part. Cette double mise en contexte des parcours éclaire et explique les comportements observés.

6.2.1 Types de parcours et territoires personnels

Sessions fermées *vs.* sessions ouvertes

Les sites vus par un panéliste dans une seule session au cours des 34 mois d'observation constituent en moyenne les trois quarts du corpus de sites du panéliste. En revanche, ils ne figurent pas de manière égale dans toutes les sessions : seuls 35 % des sessions du panel SN00-02 contiennent des sites de ce type, et leur présence ou non dans un parcours est significative. En effet, les cinq parcours-type que nous avons identifiés ne sont pas représentés de la même manière dans les sessions comprenant ces sites à visite unique et dans celles qui n'en contiennent pas (voir Tableau 6.12).

Tableau 6.12. SN00-02 – Présence de sites vus une seule fois par le panéliste et types de parcours

Classe de session	Sessions ne contenant pas de sites vus une seule fois	Sessions contenant un ou plusieurs sites vus une seule fois	Toutes sessions
Parcours éclairs	21,5 %	3,4 %	15,0 %
Parcours ciblés	28,5 %	6,3 %	20,6 %
Parcours à détours	22,1 %	22 %	22,1 %
Parcours à pivots	19,0 %	47,4 %	29,1 %
Parcours éclatés	8,9 %	20,9 %	13,2 %
Total	100 %	100 %	100 %

Dans le premier cas, parcours à pivots et parcours éclatés sont sur-représentés, et constituent les deux tiers des sessions de la classe, au détriment des parcours éclairs et ciblés ; si la session ne contient que des sites vus plus d'une fois, au contraire, les parcours rapides et linéaires sont plus présents que dans l'ensemble des sessions (la moitié des sessions, contre 35 % dans l'ensemble). Enfin, les parcours à détours échappent à ce comportements, leur part étant la même dans les deux cas (21 %), ce qui confirme le profil mixte de ce type de parcours, à cheval entre la classe des parcours à pivots et celle des parcours ciblés.

Ces éléments se retrouvent également sous l'angle des thèmes et des services accédés : pour croiser ces deux variables, nous avons évalué pour chaque panéliste la fréquence des thèmes et services qu'il visite sur l'ensemble de la période, que nous avons recodée en cinq catégories allant de « contenu très rare » à « contenu habituel » (voir Tableau 6.13).

Tableau 6.13. Recodage de la fréquence des thèmes/services de chaque panéliste

Fréquence codée	Présence dans les sessions du panéliste
thème / service très rare	majoritaire dans aucune session
thème / service rares	majoritaire dans moins de 5 % des sessions
thème / service occasionnels	majoritaire dans 5 % à 20 % des sessions
thème / service fréquents	majoritaire dans 20 % à 40 % des sessions
thème / service habituels	majoritaire dans plus de 40 % des sessions

Les sessions contenant des thèmes peu fréquents chez le panéliste sont plus longues, plus complexes, et s'apparentent à des parcours de découverte (voir Tableau 6.14). Les sessions comportant des contenus rares et très rares pour un panéliste sont ainsi quasiment absentes des parcours éclairs et ciblés (1,5 % et 6,1 %), alors qu'elles sont très présentes dans les parcours à pivots (50 %) ; au fur et à mesure que l'on sélectionne des sessions avec des contenus plus fréquemment visités, ces sessions sont de plus en plus tournées vers les parcours éclairs et ciblés, et de moins en moins vers les parcours à pivots et éclatés.

Tableau 6.14. SN00-02 – Types de parcours et fréquence des contenus pour le panéliste (sessions bien décrites)

Classe de session	Sessions avec des contenus très rares	Sessions avec des contenus rares	Sessions avec des contenus occasionnels	Sessions avec des contenus fréquents	Sessions avec des contenus habituels	Toutes sessions
Parcours éclairs	1,5 %	6,1 %	8,7 %	10,3 %	14,1 %	13,8 %
Parcours ciblés	10,7 %	20,7 %	22,2 %	25,9 %	26,3 %	24,9 %
Parcours à détours	19,3 %	24,3 %	23,6 %	22,4 %	21,2 %	22,2 %
Parcours à pivots	49,2 %	35,9 %	33,2 %	30,2 %	25,6 %	28,3 %
Parcours éclatés	19,4 %	13,0 %	12,2 %	11,1 %	12,9 %	10,8 %
Total	100 %	100 %	100 %	100 %	100 %	100 %

La visite de nouveaux sites et l'appréhension de contenus inhabituels sont donc résolument liées à des parcours longs, plus complexes, plutôt orientés vers la recherche, en particulier avec les parcours à pivots. Une opposition nette se dégage entre deux contextes particuliers de navigation : d'un côté, les sessions routinières, plutôt fermées, qui amènent l'internaute à visiter des sites connus, et de l'autre des sessions ouvertes et exploratoires, où l'utilisateur découvre de nouveaux sites et de nouveaux contenus.

Du familier vers l'inconnu

Si la catégorie des sites habituels est très minoritaire dans l'ensemble des sites visités, elle constitue pourtant le noyau central des territoires personnels sur le Web. En effet, bien qu'aucun site de cette catégorie pris isolément n'apparaisse dans toutes les sessions, ces sites considérés dans leur ensemble sont présents dans 91 % des sessions de chaque panéliste en moyenne (voir Tableau 6.15).

Tableau 6.15. SN00-02 – Empan des sites et présence dans les sessions

% de sessions contenant au moins un site d'empan..	
Empan nul (site vu dans une seule journée)	43%
1 à 9 jours	13%
10 à 30 jours	12%
Un à cinq mois	35%
Six à 12 mois	36%
Plus d'un an	91%

Clef de lecture : pour un panéliste donné, les sites de son corpus dont l'empan est supérieur à un an sont présents en moyenne dans 91 % de ses sessions sur les trois ans.

À l'inverse, les sites qui ne sont vus qu'au cours d'une seule journée (au sein d'une ou plusieurs sessions dans la journée) sont présents dans moins de la moitié des sessions d'un panéliste. Le croisement de ces deux variables amène à penser que la découverte de nouveaux sites, qu'ils soient vus une fois ou dans une période allant jusqu'à douze mois, se fait quasi-systématiquement à partir d'un terrain connu : sites d'information proposant des liens, outils de recherches préférentiels. Ceci se voit confirmé lorsque l'on examine le temps passé dans chaque session sur ces sites habituels : dans 70 % des sessions, c'est sur ces sites que l'internaute passe le plus de temps dans la session (voir Tableau 6.16). Les autres catégories de sites n'entrent en première ligne que dans moins d'un tiers des sessions, les sites vus dans l'espace d'une journée n'étant dominants que dans 10 % des cas.

Tableau 6.16. SN00-02 – Empan des sites et importance dans les sessions

Catégorie de sites la plus présente dans la session en durée	% sessions
Sites vus dans une journée (empan nul)	9,5 %
Sites occasionnels, très faible durée de vie (empan de 1 à 9 jours)	2,0 %
Sites occasionnels, faible durée de vie (empan de 10 à 29 jours)	1,6 %
Sites occasionnels, durée de vie moyenne (empan de 1 à 5 mois)	8,2 %
Sites occasionnels, durée de vie importante (empan de 6 à 12 mois)	8,7 %
Sites habituels (empan de plus d'un an)	70,1 %
Total	100 %

Clef de lecture : dans 9,5 % des sessions, c'est sur les sites à visite unique que l'internaute passe le plus de temps dans la session.

Les sites habituels sont donc le support privilégié de la navigation à double titre : non seulement ils sont présents dans la grande majorité des sessions, mais en plus, ils occupent souvent la part la plus importante du temps de navigation. Cela est compréhensible pour les sessions routinières où un seul site est visité, mais pour les sessions plus longues, où l'internaute découvre de nouveaux sites, cela signifie que le territoire nouveau et occasionnel est toujours structuré autour du territoire connu et maîtrisé, et que les écarts de parcours restent limités, voire timides. Ce constat rejette définitivement l'idée de surf débridé au gré des hypertextes au rang des mythes : une bonne partie des sessions n'amène l'internaute qu'à des endroits connus et fréquents, et même lors de sessions ouvertes, il ne s'éloigne que peu de temps de l'espace personnel qu'il s'est forgé.

Typologie

Pour confirmer ces éléments, les observer de manière systématique et analyser leur combinaison, nous avons pratiqué une classification des sessions sur la base des indices relatifs à la place des sites de chaque session dans l'ensemble des sessions de l'individu. En travaillant toujours sur les données SN00-02, nous décrivons chaque session par trois familles de variables : fréquence des thèmes ou des service accédés, fréquence des sites visités et empan des sites visités. Chaque famille de variable est décomposée en valeurs significatives, dont on évalue la place dans la session ; nous ramenons ces trois éléments au temps passé dans la session, et obtenons un total de quatorze variables présentées au Tableau 6.17. Du fait de l'inclusion de variables de

contenu, nous ne travaillons que sur les sessions bien décrites par le duo *CatService* - Nomade, l'annuaire retenu ici pour caractériser les contenus visités.

Tableau 6.17 . Variables retenues pour décrire les territoires dans les sessions

Famille de variable	% de la durée de la session sur des…
Fréquence des sites	sites n'apparaissant que dans cette session sites apparaissant dans 0 à 9 % des sessions sites apparaissant dans 10 à 39 % des sessions sites apparaissant dans 40 à 100 % des sessions
Empan des sites	sites n'apparaissant que dans une journée (empan nul) sites d'empan entre 1 et 29 jours sites d'empan entre 1 et 6 mois sites d'empan entre 7 et 12 mois sites d'empan supérieur à un an
Fréquence des thèmes et services	thèmes / services très rares thèmes / services rares thèmes / services occasionnels thèmes / services fréquents thèmes / services habituels

À l'issue de l'analyse en composantes principales des sessions sur ces variables, nous avons pratiqué une classification sur les six premiers facteurs (74 % de l'information), et retenu une partition en trois classes (voir Figure 6.5 ci-dessous). Comme le laisse entrevoir la projection sur les axes 1 et 2, le détail de la composition des classes oppose assez nettement un groupe (classe 1) aux deux autres.

Ce groupe rassemble 51,6 % des sessions, et s'apparente aux parcours exclusivement routiniers : 93 % de la durée de ces sessions se fait sur des sites dont l'empan dépasse l'année (global : 73 %), 58 % sur des sites figurant parmi les plus fréquents, et les deux tiers sur des thèmes et services fréquents ou habituels (voir Tableau 6.18).

Tableau 6.18. SN00-02 – Variables discriminantes des sessions routinières (cl. 1)

	Moy. pour la classe	Moy. globale	Variable
Corrélation positive	95 %	73 %	% durée sur sites d'empan de 350 jours et plus
	58 %	36 %	% durée sur sites vus dans 40 à 100% des sessions
	29 %	20 %	% de durée sur cats Nomade/CS fréquentes
	31 %	24 %	% durée sur sites vus dans 10 à 39% des sessions
	20 %	13 %	% de durée sur cats Nomade/CS habituelles
	27 %	29 %	% de durée sur cats Nomade/CS occasionnelles
Corrélation négative	0 %	3 %	% durée sur sites d'empan entre 1 et 29 jours
	12 %	20 %	% de durée sur cats Nomade/CS rares
	2 %	8 %	% durée sur sites d'empan entre 180 et 349 jours
	1 %	7 %	% durée sur sites d'empan entre 30 et 179 jours
	1 %	9 %	% durée sur sites vus dans une seule session
	2 %	9 %	% durée sur sites d'empan 0
	10%	32%	% durée sur sites vus dans 0 à 9% des sessions

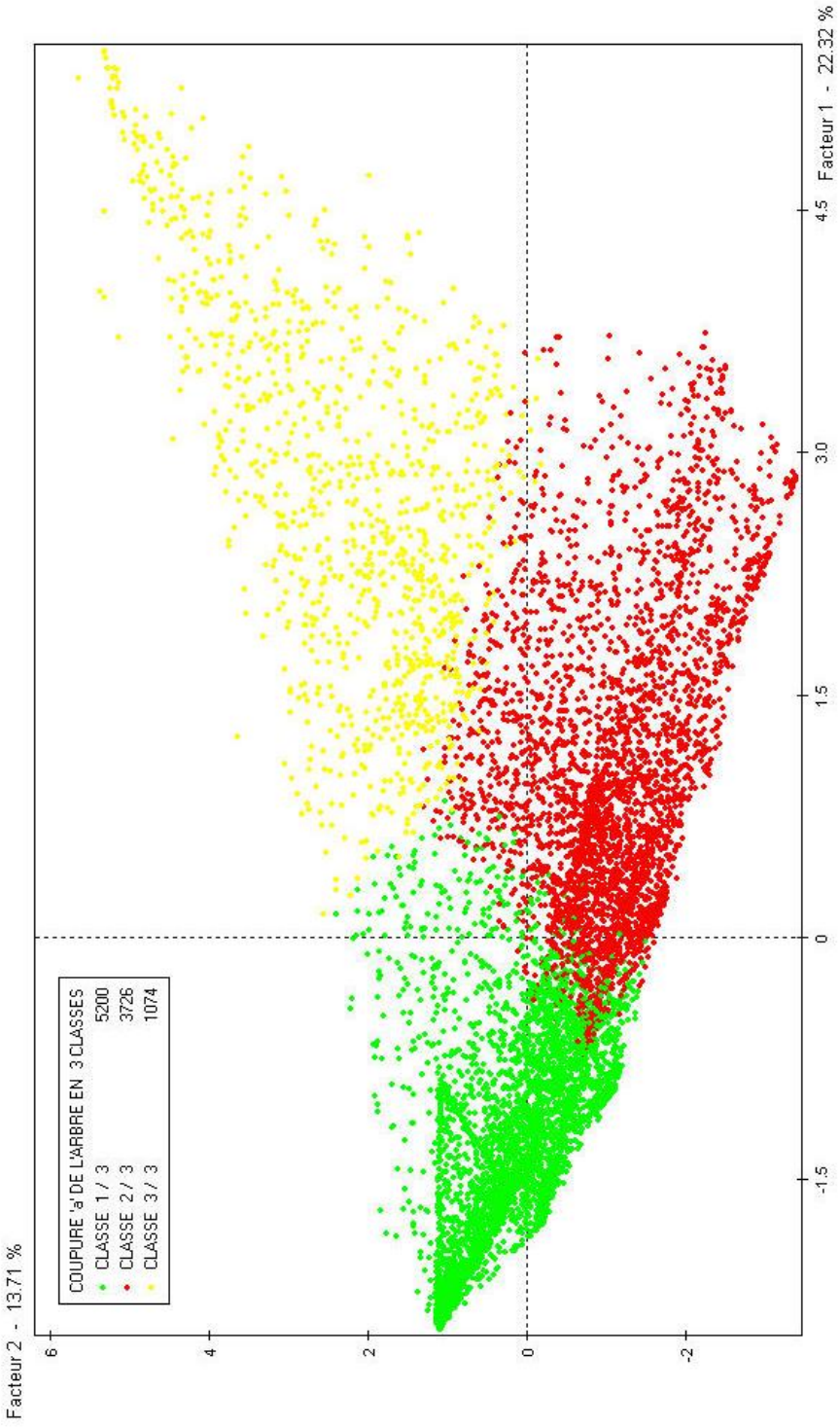


Figure 6.5. SN00-02 – Classification des sessions relative aux territoires sur le Web

À ces sessions routinières, s'opposent deux groupes de sessions plus ouvertes, orientées vers des contenus moins habituels. Le premier groupe, qui représente 36,5 % des parcours, est centré sur des contenus occasionnels : les sites présents dans peu de sessions et ceux d'empan moyen (entre un mois et un an) y sont sur-représentés, ainsi que les catégories Nomade / *CatService* rares (voir Tableau 6.19). Pour autant, ces sessions ne sont pas orientées vers la découverte, et semblent plutôt correspondre à des contenus ou des services que le panéliste visite relativement régulièrement sur une période donnée.

Tableau 6.19. SN00-02 – Variables discriminantes des sessions « contenus occasionnels » (cl. 2)

	Moy. pour la classe	Moy. globale	Variable
Corrélation positive	67 %	32 %	% durée session sur sites vus dans 0 à 9% des sessions
	20 %	8 %	% durée session sur sites d'empan entre 180 et 349 jours
	16 %	7 %	% durée session sur sites d'empan entre 30 et 179 jours
	28 %	20 %	% de durée session sur cats Nomade/CS rares
	7 %	3 %	% durée session sur sites d'empan entre 1 et 29 jours
	33 %	29 %	% de durée session sur cats Nomade/CS occasionnelles
Corrélation négative	17 %	24 %	% durée session sur sites vus dans 10 à 39% des sessions
	7 %	13 %	% de durée session sur cats Nomade/CS habituelles
	5 %	9 %	% durée session sur sites d'empan nul
	5 %	9 %	% durée session sur sites vus dans une seule session
	11 %	20 %	% de durée session sur cats Nomade/CS fréquentes
	52 %	73 %	% durée session sur sites d'empan de 350 jours et plus
	11 %	36 %	% durée session sur sites vus dans 40 à 100% des sessions

Tableau 6.20. SN00-02 – Variables discriminantes des sessions « découverte » (cl. 3)

	Moy. pour la classe	Moy. globale	Variable
Corrélation positive	56 %	9 %	% durée session sur sites d'empan nul
	53 %	9 %	% durée session sur sites vus dans une seule session
	35 %	20 %	% de durée session sur cats Nomade/CS rares
	1 %	0 %	% de durée session sur cats Nomade/CS très rares
Corrélation négative	5 %	7 %	% durée session sur sites d'empan entre 30 et 179 jours
	5 %	8 %	% durée session sur sites d'empan entre 180 et 349 jours
	23 %	32 %	% durée session sur sites vus dans 0 à 9% des sessions
	3 %	13 %	% de durée session sur cats Nomade/CS habituelles
	7 %	20 %	% de durée session sur cats Nomade/CS fréquentes
	10 %	24 %	% durée session sur sites vus dans 10 à 39% des sessions
	14 %	36 %	% durée session sur sites vus dans 40 à 100% des sessions
	32 %	73 %	% durée session sur sites d'empan de 350 jours et plus

Le dernier groupe de sessions, représentant 11,9 % de l'ensemble, est au contraire orienté vers la découverte et l'exploration de nouveaux territoires (voir Tableau 6.20). Les sites apparaissant dans une seule session, sous-représentés dans les autres groupes, sont ici majoritaires dans la session, dont ils représentent en moyenne 56 %

de la durée. Corrélativement, les thèmes et services rares et très rares sont fréquents dans ce groupe, au détriment des thèmes fréquents et habituels.

L'activité de navigation se partage donc globalement entre deux grands types de sessions : d'un côté, les sessions routinières où l'internaute accède à des pages qu'il voit très fréquemment, et qui font partie de son territoire restreint. De l'autre, on trouve des parcours plus ouverts, décomposés en sessions de semi-routine, aux contenus occasionnels que l'utilisateur sera néanmoins amené à revisiter, et en sessions de découverte au sein desquelles l'internaute visite beaucoup de sites qu'il ne reverra jamais par la suite.

> ***Synthèse.*** *Sur le plan des thèmes et services accédés, une opposition nette se dégage entre deux contextes particuliers de navigation : d'un côté, les sessions routinières, plutôt fermées, qui amènent l'internaute à visiter des sites connus, et de l'autre des sessions ouvertes et exploratoires, où l'utilisateur découvre de nouveaux sites et de nouveaux contenus. Cette découverte suit un mouvement allant du connu vers l'inconnu : les sites habituels constituent le noyau central des territoires personnels sur le Web, et forment le support privilégié de la navigation à double titre. Non seulement ils sont présents dans la grande majorité des sessions, mais en plus, ils occupent souvent la part la plus importante du temps de navigation. La classification des sessions sur la base de la fréquence et de l'empan des sites et des contenus affine cette description et met en évidence deux grands types de sessions : d'un côté, les sessions routinières renvoyant à un territoire restreint fréquemment balayé ; de l'autre, des parcours plus ouverts, décomposés en sessions de semi-routine, aux sites et contenus occasionnels mais réguliers en contexte, et sessions de découverte dont la plupart des sites ne seront jamais revus par la suite.*

6.2.2 Navigation routinière et parcours exploratoires

Cette description tripartite en sessions routinières, à contenus occasionnels et tournées vers la découverte recouvre sur le fond l'opposition constatée sur la base de la topologie des parcours entre sessions courtes et linéaires (parcours éclairs et ciblés), et sessions longues et complexes (parcours à détours, à pivots et éclatés). Le croisement de ces deux typologies montre le lien qui existe entre elles, et permet de l'affiner (voir Tableau 6.21 ci-dessous). Au sein des sessions routinières, parcours éclairs et parcours ciblés représentent plus de la moitié des sessions, contre 28 % pour les sessions « contenus occasionnels » et 16 % pour les sessions « découverte » : il y a bien un lien fort entre la visite de contenus nouveaux et l'allongement et la complexification des parcours. Pour autant, même au sein des sessions routinières, les parcours complexes demeurent importants, tandis que dans les sessions « découverte », les parcours à pivots sont majoritaires : ces deux exemples appellent à considérer la combinaison des comportements de navigation et des territoires comme le signe de contextes de navigation différenciés, où plusieurs activités peuvent être imbriquées.

Tableau 6.21. SN00-02 – Topologie des sessions et territoires personnels

	Répartition globale	Territoires sur le Web		
		Sessions routinières	Sessions « contenus occasionnels »	Sessions « découverte »
Parcours éclairs	13,8 %	18,2 %	10,4 %	5,2 %
Parcours ciblés	24,9 % } 38,7 %	33,3 % } 51,5 %	17,6 % } 28 %	10,7 % } 15,9 %
Parcours à détours	22,2 %	20,2 %	23,7 %	26,2 %
Parcours à pivots	28,3 % } 61,3 %	19,4 % } 48,5 %	34,9 % } 72 %	47,0 % } 84,1 %
Parcours éclatés	10,8 %	8,9 %	13,4 %	10,9 %
Total	100 %	100 %	100 %	100 %

L'examen détaillé des différents groupes va nous permettre d'approfondir ces éléments et de voir comment sont articulés et valorisés au sein des sessions le familier et l'inconnu au sein de cours d'action particuliers.

Navigation routinière

Les deux tiers des parcours ciblés et des parcours éclairs relèvent de sessions routinières, principalement occupées par des contenus visités fréquemment et sur de longues périodes ; réciproquement, les sessions routinières sont composées pour la moitié d'entre elles de parcours ciblés et de parcours éclairs. Au sein des sessions routinières, on distinguera donc deux sous-classes : d'un côté, les parcours routiniers directs, linéaires et rapides, et de l'autre des parcours routiniers complexes amenant l'utilisateur à revoir des pages et des sites au sein de la session. Entre ces deux extrêmes, c'est la pluri-activité qui semble être déterminante : au sein des parcours routiniers, on compte en moyenne 1,5 descriptifs de contenu différents (catégories Nomade et *CatService*) pour les parcours éclairs, contre 4,2 pour les parcours à pivot ou les parcours éclatés. L'examen manuel des sessions le confirme : les sessions routinières directes sont dominées par des contenus à fort renouvellement, ou qui amènent l'utilisateur à y revenir fréquemment. Au premier rang de ceux-ci, on trouve les portails généralistes, souvent ceux de fournisseurs d'accès, les services de messagerie, ainsi que des sites d'informations généraux (services dédiés sur les portails généralistes, ou journaux en ligne comme le Parisien, Libération, le Monde) ou boursière (Boursorama, sites de banques, etc.).

Lorsque la session s'allonge et se complexifie, ces contenus sont soit juxtaposés, soit entrelacés. Dans le premier cas, l'internaute balise son territoire familier sur le Web en passant sur chacun des sites familiers plus ou moins de temps mais toujours en une seule séquence, et n'y revient pas par la suite. La session présentée Figure 6.6 en est un exemple : les sites Club Internet, Le Monde et `education.gouv.fr` apparaissent dans une grande partie des sessions du panéliste, et sont ici visités séquentiellement.

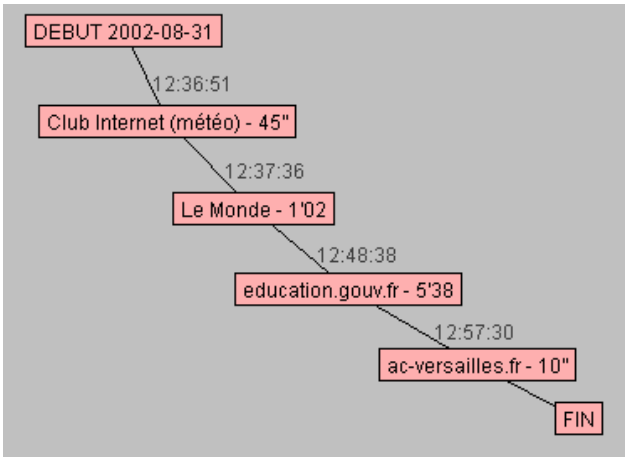


Figure 6.6. SN00-02 – session 107972 (pan 16632)

Ceci n'exclut pas une cohérence entre les différents sites de la session : dans cet exemple, si la visite initiale du service d'informations météorologiques sur Club-Internet semble indépendante du reste de la session, par la suite l'internaute consulte sur le site du Monde un article relatif à l'augmentation du salaire des ministres, puis la page d'accueil de la rubrique « Éducation » où il y suit un lien vers le site de l'Éducation Nationale pour y consulter des informations sur les avantages sociaux réservés aux enseignants.

Dans le cas de l'entrelacement, la navigation peut être alternée et s'apparenter à du multifenêtrage. Toutefois, ce cas reste minoritaire : on observe bien souvent le recours à un site-pivot autour duquel s'organisent et se distribuent les différents contenus visités, qu'ils aient ou non un rapport entre eux ou avec ce site central. Celui-ci est dans la plupart des cas un portail généraliste, dont l'utilisateur exploite le moteur de recherche ou suit les liens pour accéder aux sites qui l'intéressent. Un exemple de cette organisation particulière de la navigation est donné Figure 6.7 et Tableau 6.22 : la rubrique « Informations » de Yahoo fait office de nœud central, même si l'utilisateur n'y passe que peu de temps à chaque fois.

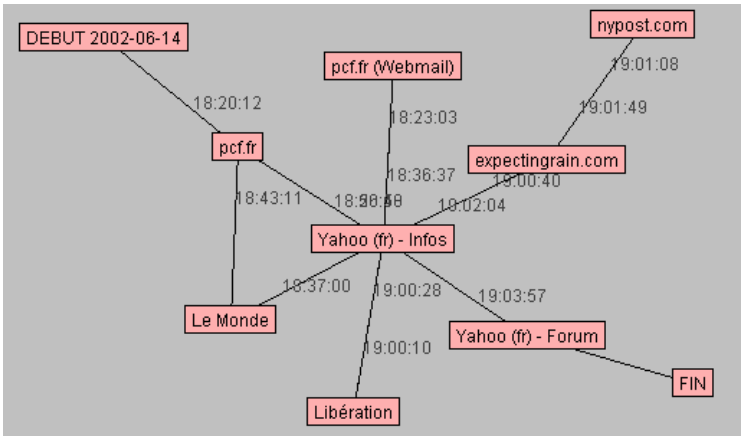


Figure 6.7. SN00-02 – session 48806 (pan 1731)

Tableau 6.22. SN00-02 – session 48806, vue séquentielle

Site / Portail	Date	Nb pages	Durée	Empan	Présence dans les sess. du pan.
www.pcf.fr	18:20:12	1	36"	930	331 (17 %)
Yahoo (fr) - Informations	18:20:48	7	2'15	1031	840 (43 %)
pcf.fr - WebMail	18:23:03	36	13'34	900	404 (21 %)
Yahoo (fr) - Informations	18:36:37	1	23"	1031	840 (43 %)
Le Monde	18:37:00	5	6'11	641	191 (10 %)
www.pcf.fr	18:43:11	28	13'39	930	331 (17 %)
Yahoo (fr) - Informations	18:56:50	11	3'20	1031	840 (43 %)
Libération	19:00:10	2	18"	641	319 (16 %)
Yahoo (fr) - Informations	19:00:28	1	12"	1031	840 (43 %)
expectingrain.com	19:00:40	1	28"	921	172 (9 %)
nypost.com	19:01:08	3	41"	0	1 (0 %)
expectingrain.com	19:01:49	1	15"	921	172 (9 %)
Yahoo (fr) - Informations	19:02:04	5	1'53	1031	840 (43 %)
Yahoo (fr) - Forum	19:03:57	20	1'01	488	102 (5 %)

Dans ce type de configuration, il n'est pas rare que l'internaute « déborde » de ses contenus habituels, mais cet écart demeure ponctuel et borné à un seul site, et occupe peu de temps dans la session. Ces écarts correspondent bien souvent au suivi de liens proposés dans les sites habituels : leur contenu est déjà décrit dans la page source, et ces éléments de découverte de nouvelles pages et de nouveaux sites sont motivés et maîtrisés, et ne donnent pas lieu à de longues digressions. Dans l'exemple ci-dessus, le site du journal nord-américain New York Post (nypost.com) est vu ici pour la première et unique fois par le panéliste, et correspond à un lien présent dans une page du site expectingrain.com.

En définitive, il apparaît que lors de ces balayages récurrents des terrains connus, les internautes se connectent à des flux. Ce peut être des flux d'information – informations générales, boursières, sites spécialisés – ou des flux de communication – WebMail, forums, *chat* ; dans les deux cas, le mode de réception des contenus ne semble pas si éloigné de celui des médias traditionnels : une fois que l'internaute a identifié les canaux qui l'intéressent, c'est par ce biais qu'il « consomme » les contenus, comme il le ferait pour des émissions de télévision ou de radio, ou des quotidiens et des magazines. La différence réside ici d'une part dans la disponibilité des contenus Web et la possibilité de les mobiliser au moment qui convient le mieux à l'utilisateur, et d'autre part dans l'ouverture que ces contenus routiniers offrent vers l'inhabituel *via* l'hypertexte.

L'occasionnel : régularité en contexte

L'appréhension de contenus occasionnels est différente, et implique des parcours plus complexes : on a pu constater que les parcours à pivots et à détours sont plus présents dans ce cas (voir Tableau 6.21, p. 258) que pour les contenus habituels. Cependant, pour accéder à ces contenus peu fréquents, l'internaute a peu recours aux moteurs de recherche : seul un tiers des sessions « contenus occasionnels » comportent un accès à un moteur, contre 12 % pour les sessions routinières, et 60 % pour les sessions « découverte » (voir Tableau 6.23 ci-dessous).

Tableau 6.23. SN00-02 – Part des sessions avec moteur par type de session

	Sessions routinières	Sessions « contenus occasionnels »	Sessions « découverte »
Parcours éclairs	3,5 %	8,5 %	20,6 %
Parcours ciblés	5,3 %	11,2 %	31,6 %
Parcours à détours	13,8 %	27,8 %	56,4 %
Parcours à pivots	23,5 %	42,3 %	72,6 %
Parcours éclatés	22,9 %	35,8 %	62,5 %
Ensemble	11,8 %	29 %	60,2 %

Il s'agit ici de sites connus et visités plusieurs fois par le panéliste, mais dans des contextes qui se produisent rarement. Le contenu de ces sites est variable selon les intérêts de chaque internaute ; toutefois, de manière générale, on retrouve fréquemment dans ce genre de configurations des activités d'achat en ligne ou d'organisation de voyages. En outre, lorsqu'un moteur de recherche est mobilisé dans ce contexte, il s'agit bien souvent de requêtes ponctuelles correspondant à la recherche de l'adresse d'un site connu, par exemple « fnac », « alapage », « nouvelles frontières », forme d'utilisation des moteurs comme pseudo-*bookmarks*.

L'exemple de session présenté Figure 6.8 et Tableau 6.24 ci-dessous illustre ces éléments : l'internaute cherche visiblement à connaître les films qui passent dans les cinémas près de chez lui. Pour cela, il va dans un premier temps sur le service dédié du portail Yahoo ; après cette recherche infructueuse, il visite les catégories « Art et culture ➔ Cinéma » et « Art et culture ➔ Cinéma ➔ Annuaires et guides » dans Yahoo, puis interroge le moteur de recherche avec la requête « allocine ». Cet usage du moteur s'apparente à un pseudo-signet : l'internaute connaît l'existence du site et souhaite y aller, mais il n'en a pas l'adresse. Après cela, une longue séquence de plus de 7 minutes sur `allocine.fr` conduit l'utilisateur à trouver les horaires des films pour les salles près de son domicile, et il poursuit sa recherche en visitant le site dédié à un film particulier figurant au programme.

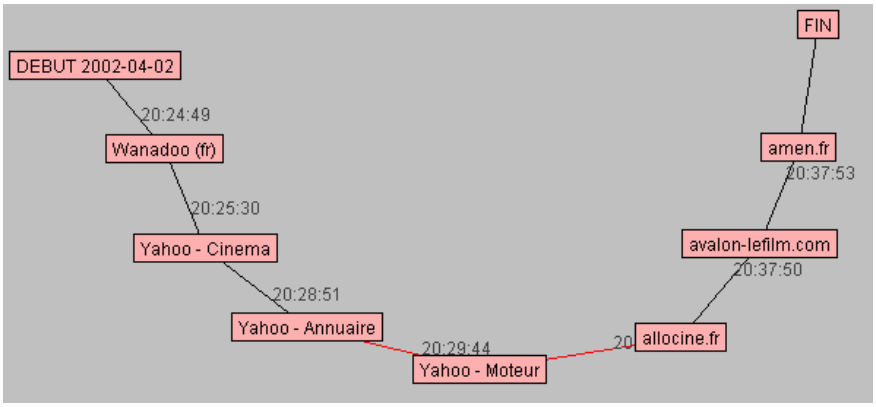


Figure 6.8. SN00-02 – session 328 (pan. 4)

Tableau 6.24. SN00-02 – session 328 (pan. 4), vue séquentielle

Site / portail	Date	Nb pages	Durée	Empan	Présence dans les sess. du pan.
Wanadoo (fr)	20:24:49	6	41”	818	109 (50 %)
Yahoo - cinéma	20:25:30	15	3’21	106	5 (2 %)
Yahoo - annuaire	20:28:51	2	55”	816	14 (6 %)
Yahoo - moteur	20:29:44	1	31”	791	11 (5 %)
allocine.fr	20:30:15	21	7’35	865	7 (3 %)
avalon-lefilm.com	20:37:50	1	3”	0	1 (0 %)
webserveur91.amen.fr/avalon	20:37:53	11	11’57	0	1 (0 %)

Dans un autre registre, la session présentée ci-dessous (Figure 6.9 et Tableau 6.25) est structurée autour de la recherche de billets d’avion : plusieurs sites de voyagistes sont visités, qui ont pour la plupart un empan très important bien qu’ils n’apparaissent que dans peu de sessions (Travelprice, Degriftour, Lastminute, Anyway). Pour y accéder, l’internaute n’utilise pas un moteur de recherche : soit il connaît assez bien l’adresse pour l’entrer dans la barre d’adresse du navigateur, soit il l’a stockée dans ses Favoris. Dans les deux cas, l’internaute est en terrain connu pour la tâche qui l’occupe, et il parcourt les différentes offres dans une logique de comparaison.

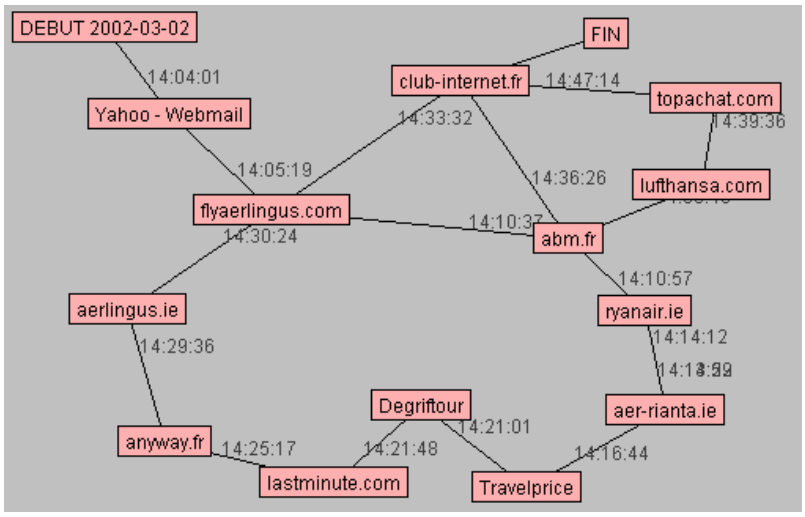


Figure 6.9. SN00-02 – session 17512 (pan. 334)

Dans ce type de parcours, l’appréhension des contenus Web est plus proche d’un format « guichet », « pages jaunes » ou « galerie marchande » : les activités induites sont plutôt rares pour chaque internaute, mais celui-ci adopte des comportements relativement stables dans ce contexte, et visite souvent les mêmes sites ou groupes de sites.

Tableau 6.25. SN00-02 – session 17512, vision séquentielle

Site	date	Nb Pages	Durée	Empan	Présence dans les sess. du pan.
Yahoo - WebMail	14:04:01	1	1'18	774	677 (47 %)
flyaerlingus.com	14:05:19	9	5'18	8	4 (0 %)
abm.fr	14:10:37	3	20"	0	2 (0 %)
ryanair.ie	14:10:57	19	3'02	0	1 (0 %)
aer-rianta.ie	14:13:59	5	13"	0	1 (0 %)
ryanair.ie	14:14:12	1	10"	0	1 (0 %)
aer-rianta.ie	14:14:22	1	2'22	0	1 (0 %)
Travelprice	14:16:44	8	4'17	799	12 (1 %)
Degriftour	14:21:01	2	47"	799	18 (1 %)
lastminute.com	14:21:48	1	3'31	839	15 (1 %)
anyway.fr	14:25:17	1	4'19	663	16 (1 %)
aerlingus.ie	14:29:36	2	48"	4	4 (0 %)
flyaerlingus.com	14:30:24	4	3'08	8	4 (0 %)
Club Internet	14:33:32	1	2'54	1006	319 (22 %)
abm.fr	14:36:26	3	22"	0	2 (0 %)
lufthansa.com	14:36:48	5	2'48	0	1 (0 %)
topachat.com	14:39:36	15	7'38	960	91 (6 %)
Club Internet	14:47:14	1	32"	1006	319 (22 %)

Corrélativement, les thèmes et services les plus visités dans ce contexte sont orientés vers le Web marchand et les activités hors Web (voir Tableau 6.26) : les catégories Nomade « Mes Courses » et « Vie pratique » sont sur-représentées dans ces sessions à contenus occasionnels, de même que les sites pornographiques.

Tableau 6.26. SN00-02 – Thèmes et services des sessions « contenus occasionnels » (Nomade)

Catégorie dominante dans la session	Sessions « contenus occasionnels »	Ensemble des sessions
Mes Courses	12,5 %	7,2 %
Société, Vie pratique	11,6 %	9,6 %
PORNO	10,2 %	5,9 %
Sport et détente	9,6 %	7,3 %
Espace B to B	8,9 %	11,3 %
Actu, médias	8,3 %	5,0 %
CS – WebMail	6,1 %	12,7 %
Nouvelles technologies	5,0 %	2,8 %
CS – Moteur	4,0 %	3,7 %
Culture et loisirs	3,7 %	3,0 %
PG - Page Accueil	3,4 %	12,5 %
Voyage, géographie	3,4 %	2,1 %
Autres	13,5 %	16,9 %

Clef de lecture : dans 12,5 % des sessions « contenus occasionnels », la catégorie Nomade « Mes Courses » occupe le plus de temps dans la session.

La démarche comparative de l'utilisateur induit un allongement global des sessions et une augmentation du nombre de sites visités ; ce phénomène peut être atténué lorsqu'un site connu répond aux attentes de l'utilisateur, comme pour celui

de la SNCF pour les horaires et billets de train, `pagesjaunes.fr` pour les fonctions d'annuaire, ou `service-public.fr` pour les démarches administratives. À l'inverse, les sites de vente en ligne de produits touristiques, culturels ou technologiques amènent souvent l'internaute à visiter plusieurs sites différents dans un même champ d'activité.

Prédateur *vs.* flâneur : usages différenciés des moteurs de recherche

La présence dans les sessions de sites occasionnels et de faible empan dénote un tout autre usage du Web, apparenté à une ressource d'informations ponctuelle mobilisée pour des besoins précis et limités dans le temps. Dans ce contexte, les moteurs de recherche sont hautement mis à contribution : on a pu voir que six sessions « découverte » sur dix contiennent une requête moteur, et que ces sessions sont pour 85 % d'entre elles des parcours complexes, avec la moitié de parcours à pivots et un quart de parcours à détours (voir Tableau 6.21, p. 258).

C'est au sein de ces sessions longues et complexes que l'utilisateur découvre de nouveaux sites, ouvre de nouveaux territoires, explore des contenus inédits ; pour autant, l'usage majoritaire des moteurs dans ce contexte nous rappelle qu'il ne s'agit pas là d'une navigation au gré du vent et des liens hypertexte, mais d'une recherche sur un sujet précis, guidée par un objectif explicitable par l'utilisateur. Au sein de ces parcours ouverts, on distingue deux types de comportements, que l'on qualifiera de « prédateur » et de « flâneur » : dans le premier cas, l'internaute cherche des informations précises et ciblées sur un thème donné, et feuillette rapidement les différentes ressources qu'il trouve pour aller au plus vite au résultat qui l'intéresse. Dans l'autre cas, les détours sont plus nombreux, les requêtes plus ouvertes, et l'internaute semble de manière générale disposer de plus de temps pour visiter les « à côté ». Au cours des entretiens menés avec les panélistes de BibUsages, ces éléments de différenciation étaient soulignés par les interviewés :

> *« Je n'utilise jamais Kartoo comme moteur de recherche au niveau professionnel, j'utilise Wanadoo, Google, Voila, et puis c'est tout. Et pour mes recherches personnelles, ça va être Kartoo par exemple parce que je ne sais pas exactement comment formuler ma recherche. Au niveau professionnel, le fil directeur, c'est l'efficacité ; c'est-à-dire qu'il faut assurer une rentabilité ou une mesure de son travail qui soit réelle. Donc une recherche personnelle, c'est une somme et une variété de documentation qui vont permettre de bâtir quelque chose autour d'un concept ou d'une idée. C'est de la curiosité pure ; mais ce n'est pas sanctionné de la même manière. » (Utilisateur H)*

On retrouve ces deux types de comportements dans les données au gré de l'examen manuel des sessions. Deux exemples illustreront cette dichotomie : dans le premier, l'internaute effectue une recherche sur un sujet précis, les encres fiduciaires. Il est amené à reformuler sa requête de nombreuses fois, et à consulter de nombreuses pages de résultats : le Tableau 6.27, qui présente l'ensemble des mots-clefs recherchés sur les différents moteurs dans cette session, montre les stratégies de précision (ajout de mots), de déplacement (« encres petrel » modifié en « encres fiduciaires » puis en « encres gommables »), de sophistication (ajout de guillemets) ou de changement d'outil (passage de Google à Altavista). Au total, ce sont trente-

quatre pages de résultats consultées dans la session qui amènent l'internaute à visiter quinze sites différents relatifs à sa recherche.

Tableau 6.27. Requêtes moteur pour la session 41772 (SN00-02)

Date	Moteur	Nb pages de résultats	Requête
22:31:05	Google	2	petrel
22:32:03	Google	3	encres pétrel
22:34:58	Google	2	encres fiduciaires pétrel
22:35:11	Google	10	encres fiduciaires
22:47:40	Google	1	encres fiduciaires pétrel
22:47:44	Google	2	encres fiduciaires
22:48:15	Google	2	encres petrel
22:50:19	Altavista.fr	1	"encres petrel"
22:50:28	Altavista.fr	6	petrel
22:53:11	Google	1	encres gommables et delebiles
22:53:17	Google	3	encres gommables
22:53:55	Google	1	encres delebiles

La forme du parcours est en ce cas centrée sur l'outil de recherche : l'internaute ne sort pas des sites proposés dans les résultats du moteur, et bien souvent ne consulte que la page indiquée par le moteur, sans naviguer dans le reste du site (voir Figure 6.10). Dans l'exemple, la séquence de recherche occupe la quasi-totalité de la session : vingt-quatre minutes sur la demi-heure que dure la session, les six dernières minutes étant consacrées à une toute autre activité, la consultation de la météo pour des stations de sports d'hiver.

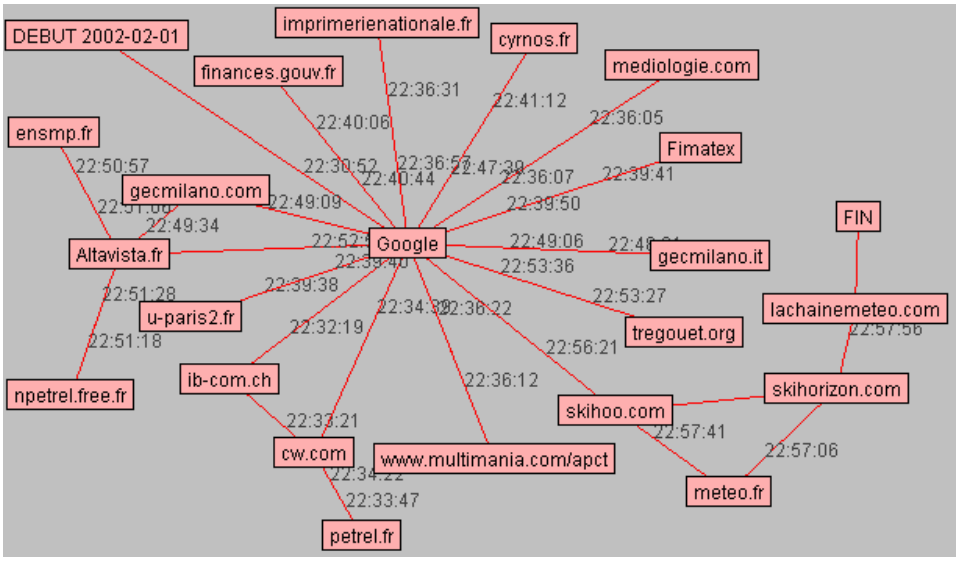


Figure 6.10. SN00-02 – session 41772 (pan 1291)

Ce premier exemple témoigne de comportements « prédateurs » : l'utilisateur a une idée relativement précise de ce qu'il cherche, dans un domaine restreint, et il a la

capacité de préciser ou de modifier le contenu de sa requête pour obtenir le résultat désiré.

Ce type de parcours s'oppose à des recherches plus ouvertes, qui embrassent une plus grande diversité, et où la succession des mots-clefs ouvre et complète le champ de recherche au lieu de le restreindre. L'exemple de session donné Figure 6.11 en est une illustration : d'une durée totale d'une heure, cette session est axée autour de trois recherches successives – « la Tranche-sur-Mer », « île de Ré » et « île de Noirmoutier ».

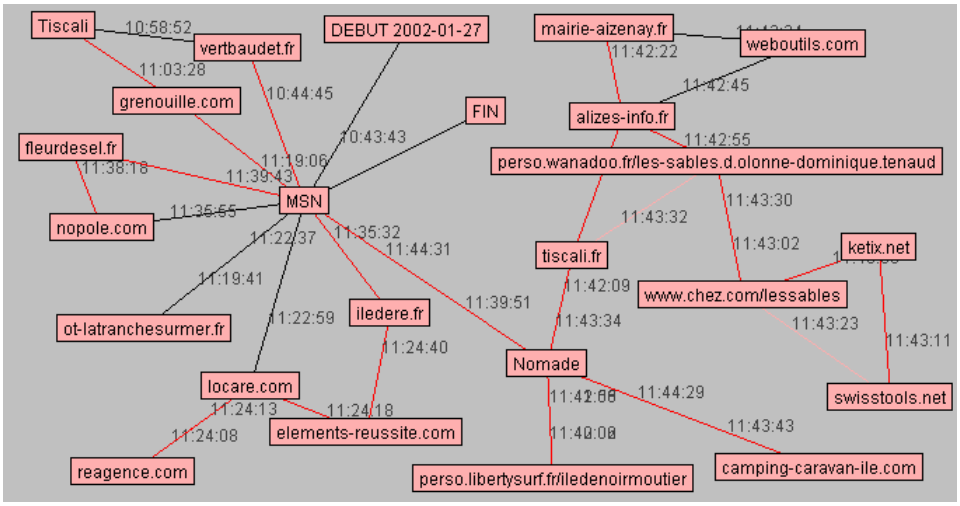


Figure 6.11. SN00-02 – session 7310 (pan 156)

Dans cette session, les requêtes ne sont pas reformulées, et peu de pages de résultats sont consultées : tout au plus l'utilisateur essaye sa dernière requête également sur Nomade après avoir utilisé MSN jusqu'alors. L'allongement et la complexification de la session vient d'un temps plus important accordé aux pages de résultats de recherche : visiblement intéressé par les éléments touristiques des trois lieux, l'internaute navigue sur les sites proposés par les moteurs, examine les listes d'hôtels et leur site le cas échéant, consulte les plans, les descriptifs des villes, etc.

Ce comportement de type « flâneur » est surtout observable dans les requêtes moteur visant des contenus pornographiques. Les mots-clefs sont dans ce cas souvent vagues et généraux : « sexe », « porno », « gratuit », « pics » ou encore « xxx » sont très souvent employés et combinés dans ce contexte, amenant à naviguer longuement de site en site à partir des pages de résultat des moteurs. Hors de ce champ, la navigation au gré des liens est globalement assez rare : le fonctionnement même des outils de recherche incite l'utilisateur à préciser autant que possible ses requêtes, de sorte que c'est vers les sites eux-mêmes que se déplace la demande des utilisateurs en matière de liens pertinents. La difficulté n'est plus tant de trouver les bons moteurs, mais de trouver les bon sites sur un thème donné, qui fourniront à la fois des contenus de qualité et des ouvertures pertinentes. Le comportement des internautes de BibUsages sur le site de Gallica est révélateur à cet égard : le moteur de recherche interne est largement majoritaire dans l'accès aux documents, mais les

« Dossiers », visites guidées thématiques, sont occasionnellement consultés, lorsque l'internaute dispose de plus de temps et de curiosité.

Signification des contenus à l'aune des pratiques

En définitive, l'analyse de la forme des parcours et de leurs contenus en regard des territoires personnels sur le Web nous amène à distinguer trois types d'activité, auxquelles correspondent trois modes d'appréhension et d'interprétation des contenus. Les actions routinières valorisent les contenus de type « flux d'information » ou « flux de communication », au sein de sessions plutôt courtes où se juxtaposent les différents sites habituels ; ce territoire familier est régulièrement balisé par l'internaute, et sous-tend l'ensemble de sa navigation. Les contenus occasionnels correspondent au Web marchand ou de service, et à un mode d'usage de type « guichet de renseignements » ; ce contexte, plutôt rare, amène l'utilisateur à visiter un nombre restreint de sites connus, qui sont balayés au cours de la session selon un mode de feuilletage orienté vers la comparaison. Enfin, les parcours ouverts, qui amènent à découvrir de nouveaux sites, font massivement appel aux moteurs de recherche, mais sont orientés vers la résolution d'un problème ou d'une question particulière. Dans ce dernier cas, le Web n'est pas tant considéré comme un espace documentaire que comme une collection de sources possibles au sein desquelles il convient de trouver la plus complète et la plus fiable pour une interrogation ponctuelle.

Ces trois schémas d'action prototypiques permettent de décrire la majorité des activités de navigation sur le Web. Ils impliquent des formes de parcours et des rythmiques particulières, et conditionnent l'appréhension des sites et des contenus : un même site pourra apparaître dans ces trois contextes, mais s'insérer dans des structures interprétatives différentes. C'est ce que nous avons pu voir avec les moteurs de recherche, où l'outil peut être utilisé comme pseudo-signet ou comme outils de découverte de ressources. De manière générale, certains sites sont plus sujets aux utilisations multiples tandis que d'autres orientent vers des schémas actantiels fermés : le WebMail est un outil de communication fermé, tandis qu'un forum pourra être perçu par les uns comme une ressource pour un besoin ponctuel et accédé via une requête moteur, et par d'autres comme un flux impliquant une visite routinière. Autre exemple, les sites de vente en ligne impliquent globalement des comportements de type « galerie marchande », mais on a pu voir que dans certains cas, ils servent de réservoir à images pour la confection de CD récupérés *via* des réseaux de *peer-to-peer*.

Il est probable que l'évolution actuelle de l'offre en matière d'outils de publication sur le Web soutienne cette tendance : les systèmes de gestion de contenu[1] permettent à tout un chacun d'installer facilement des outils de publication collaborative, des forums, des blogs, etc. qui transforment un site statique en flux continu. Cet aspect protéiforme des contenus Web, que l'on peut considérer tantôt comme ressource documentaire, tantôt comme canal de communication et d'information, tantôt comme fournisseur de services, montre en définitive que

[1] Ou « CMS » : Content Management System.

textes, documents et outillages accessibles sur la Toile n'ont de signification que par le biais du parcours qu'y trace l'utilisateur, et du cours d'action dans lequel s'inscrit sa pratique.

> ***Synthèse.*** *En croisant la typologie des sessions fondée sur l'approche topologique et temporelle, et celle basée sur les territoires personnels sur le Web, trois comportement prototypiques de navigation sont mis à jour, qui reflètent des contextes d'usage différenciés. En premier lieu, les sessions routinières, liées aux parcours éclairs et ciblés, s'apparentent à des balayages récurrents des terrains connus, où les internautes se connectent à des flux d'information (informations générales, boursières, sites spécialisés) ou de communication (WebMail, forums,* chat*) ; dans les deux cas, le mode de réception des contenus ne semble pas si éloigné de celui des médias traditionnels. Ensuite, dans les parcours orientés « Web pratique », plus complexes et allongés, l'appréhension des contenus est plus proche d'un format « guichet », « pages jaunes » ou « galerie marchande » : les activités induites sont plutôt rares pour chaque internaute, mais celui-ci adopte des comportements relativement stables dans ce contexte, et visite souvent les mêmes sites ou groupes de sites. Enfin, les parcours ouverts, qui amènent à découvrir de nouveaux sites, font massivement appel aux moteurs de recherche, dans le cadre de requêtes précises. Dans ces trois modes prototypiques d'appréhension du Web, les comportements de navigation sont profondément influencés par la double dynamique de la familiarité territoriale et du contexte d'usage.*

6.3 Le document numérique, l'usuel et l'œuvre[1]

Dans la littérature sur les hypertextes, il est souvent fait implicitement référence à l'appréhension des contenus en contexte de recherche : en Recherche d'Information, le Web est considéré comme un réservoir d'informations et de connaissances non structuré ; dans [Ghitalla *et al.* 2003], il est envisagé comme un espace documentaire. S'il est réducteur de considérer le Web uniquement sous cet angle, cette approche est pertinente dans le cadre de son utilisation en contexte de recherche ciblée, mais elle mérite d'être revisitée à l'aune des différentes modalités d'appréhension des contenus que nous avons mises à jour. Le panel BibUsages, constitué de chercheurs et de consommateurs de fonds textuels numérisés, nous permet d'étudier plus avant ces parcours ouverts de type « recherche » et de voir comment l'accès numérique à des ressources traditionnellement imprimées en renouvelle l'appréhension.

---

[1] Cette partie reprend des éléments du rapport du projet BibUsages ; voir [Assadi *et al.* 2003a].

6.3.1 Internautes lecteurs, internautes chercheurs

Nous avons déjà eu l'occasion de souligner les spécificités du panel BibUsages en ce qui concerne les contenus visités et les types de sessions, plutôt orientées vers la recherche. Nous approfondissons ici cette description, pour voir la place qu'occupent les contenus culturels et les sessions de recherche chez ces internautes, et comment s'articulent les différents contenus visités dans le contexte de la session.

Intérêt prononcé pour les contenus culturels

Dans les réponses au questionnaire en ligne soumis en mars 2003 aux visiteurs de Gallica, nous pouvions déjà observer que les « gallicanautes » constituent une population de lecteurs particulièrement intéressés par les contenus culturels et l'actualité en ligne (voir Tableau 6.28). Parmi les centres d'intérêt déclarés, la quasi-totalité des répondants mentionne des contenus de type culturel ou académique : « Art et littérature » pour 62 % des répondants, « Recherche documentaire » pour les trois quarts ; les sites d'actualités intéressent également la moitié des interrogés, tandis que les sphères ludiques (jeux, sport) et économiques (information sur les entreprises, services bancaires) ou les services de communication restent minoritaires.

Tableau 6.28. Questionnaire BibUsages en ligne : centres d'intérêt sur le Web

Quels sont vos principaux centres d'intérêt sur le Web ?	
Contenus culturels, dont :	98,0 %
Recherche documentaire ou bibliographique	76,8 %
Art et Littérature	62,4 %
Sciences Humaines et sociales	50,7 %
Informatique et multimédia	31,2 %
Sciences et technologies	30,5 %
Autres informations culturelles	31,9 %
Informations économiques ou institutionnelles, dont :	64,1 %
Actualités	47,3 %
Institutions et service public	25,8 %
Banque et finances	13,3 %
Emploi, Stage (recherche ou offre)	10,3 %
Économie et entreprise	9,1 %
Autres informations économiques ou institutionnelles	10,3 %
Loisirs, dont :	44,2 %
Voyages, tourisme	27,2 %
Sorties, divertissements	16,9 %
Jeux	7,3 %
Sports	7,1 %
Autres Loisirs	7,1 %
Communication	21,3 %
Autres centres d'intérêt	10,4 %

L'identification des différents types de portails et de services avec *CatService* permet d'observer dans les données de trafic ces éléments de manière plus fine, et nous montre que les « gallicanautes » constituent un véritable panel d'internautes-lecteurs et d'internautes-chercheurs (voir Tableau 6.29).

Tableau 6.29. Audience par types de sites et de portails (juillet – décembre 2002)

Type de portail (*CatService*)	Présence dans les sessions BibUsages	Nombre de panélistes de BibUsages	Présence dans les sessions SN2002
Portail généraliste	60,3 %	71	55,2 %
Moteur	30,1 %	67	20,2 %
Site « perso »	21,1 %	69	15,7 %
WebMail	17,6 %	15	22,4 %
Bibliothèque électronique	8,0 %	59	0,2 %
Media / Presse	7,6 %	49	2,7 %
e-commerce / biens culturels	5,5 %	63	3,1 %
Généalogie	4,6 %	32	0,3 %
e-commerce / Banque Bourse	2,9 %	30	12,3 %
Media / Radio	2,0 %	33	1,1 %
Media / TV	1,5 %	38	2,5 %
e-commerce / tourisme	0,9 %	46	1,5 %
e-commerce / courses	0,1 %	4	0,2 %

En premier lieu, portails généralistes et moteurs de recherche occupent une place prépondérante dans l'activité Web, et touchent la quasi-totalité du panel : les moteurs de recherche entrent en tête des sites les plus visités et les plus présents dans les pratiques. On les retrouve dans 30 % des sessions du panel, contre 20 % pour le panel SN2002. Cet élément est confirmé dans les entretiens, où beaucoup d'interviewés évoquent des pratiques de recherche d'information au premier rang de leur activité sur le Web. D'autre part, le panel se caractérise par une très forte fréquentation des portails « culturels » : les bibliothèques électroniques, les sites de commerce de biens culturels (Alapage, Fnac, Amazon...) et les sites de médias (presse, radio, télévision) occupent une place privilégiée dans les pratiques. Le détail de la fréquentation des portails de vente de biens culturels révèle en particulier l'importance de la bibliophilie, avec une part non négligeable du trafic sur des sites comme Chapitre.com, Livre-rare-book, Galaxidion ou Librissimo (voir Tableau 6.30).

Tableau 6.30. Fréquentation des sites de vente de biens culturels

Portail	Nombre de panélistes	Nombre de sessions	Nombre de sessions par panéliste	Temps moyen dans la session (min.)
Amazon	55	386	7,0	4 min. 06
Fnac	42	299	7,1	2 min. 25
Alapage	29	107	3,7	2 min. 13
Chapitre.com	26	160	6,2	9 min. 17
Livre-rare-book	17	92	5,4	5 min. 02
Galaxidion	6	73	12,2	6 min. 24
Librissimo	6	7	1,2	4 min. 31
Numilog	5	14	2,8	4 min. 42
Eyrolles	3	4	1,3	1 min. 03
CNRS Éditions	2	8	4,0	8 min. 37
Autres (cinq portails)	2	2	1	0 min. 20

Enfin, les sites consacrés à la généalogie ont une place importante dans les usages, et leur présence témoigne de centres d'intérêt personnels très forts ; il faut sans doute ajouter à cela l'audience particulière des sites personnels, présents dans 21% des sessions, qui sont d'importantes sources d'informations pour les recherches généalogiques sur le Web.

Cette première vue montre un lien global entre la fréquentation des bibliothèques électroniques et les outils de recherche d'information d'une part, et les sites de « contenus à lire » et de commerce de biens culturels d'autre part. Les consommateurs de bibliothèques numériques sont de manière générale de grands consommateurs de produits culturels, que ce soit dans une optique de consultation ou de consommation.

Audience des bibliothèques électroniques en contexte

Au sein de la plateforme *CatService*, vingt-trois sites ont été identifiés comme relevant de la catégorie « bibliothèques électroniques », parmi lesquels dix-huit ont été visités par le panel entre juillet et décembre 2002. Nous avons retenu une acception large de cette catégorie en y incluant, outre les versions numériques des bibliothèques patrimoniales, des sites d'éditeurs, des bibliothèques numériques régionales ou municipales, ainsi que des collections numériques issues d'initiatives associatives. Nous constituons ainsi une catégorie assez vaste de « collections de textes d'ouvrages accessibles en ligne ». L'audience détaillée de ces différents sites montre que Gallica arrive en tête, en nombre de sessions comme de panélistes, mais que la visite des autres sites est loin d'être anecdotique (voir Tableau 6.31). Si la richesse du fonds explique la fréquentation plus forte du fond numérisé de la BnF, on peut penser à la lumière des entretiens que Gallica figure pour les internautes comme une source de textes parmi d'autres.

Tableau 6.31 - Audience des bibliothèques électroniques

Portail	Nombre de sessions	Nombre de panélistes	Temps moyen dans une session
BNF-Gallica	1039	59	25 min. 31
BNF-Autres	647	52	6 min. 37
ABU	39	15	2 min. 56
Bibliothèque de Lisieux	26	13	4 min, 03
Revues.org	25	13	2 min, 05
Athena	16	11	1 min. 48
Bibliopolis	16	6	7 min. 36
ClicNet	15	9	40 sec.
CNUM	13	4	6 min. 24
Mozambook	13	3	2 min. 56
Electronic Text Center	11	5	2 min. 13
Online Books Page	10	4	2 min. 05
INALF	8	2	5 min. 40
Autres (moins de 5 sessions)	24	10	2 min. 50

L'examen des types de sites et de services visités au sein des sessions comportant un accès à une bibliothèque électronique permet de voir comment l'usage des fonds numériques s'articule avec les autres contenus et services disponibles sur le Web, et

permet d'en cerner plus finement les contextes d'usage. Pour cela, nous comparons la présence des différents types de sites identifiés dans *CatService* au sein des sessions avec bibliothèque électronique et dans l'ensemble des sessions (voir Tableau 6.32).

Tableau 6.32 - Présence des différents types de sites et services dans les sessions avec accès à une bibliothèque électronique et dans l'ensemble des sessions

	Présence dans les sessions bib. élec.	Présence dans l'ensemble des sessions	Variation de la présence dans les sessions bib. élec.
Bibliothèque électronique	100,0 %	8,0 %	
e-commerce / biens culturels	10,4 %	5,5 %	+ 88,2 %
Sites personnels	29,6 %	21,1 %	+ 40,8 %
Moteur	40,8 %	30,1 %	+ 35,7 %
Généalogie	5,2 %	4,6 %	+ 13,4 %
e-commerce / tourisme	1,0 %	0,9 %	+ 7,7 %
Media / Radio	2,0 %	2,0 %	- 0,8 %
Portail généraliste	55,0 %	60,3 %	- 8,9 %
Media / Presse	5,5 %	7,6 %	- 27,6 %
e-commerce / Banque Bourse	1,8 %	2,9 %	- 37,2 %
Media / TV	0,8 %	1,5 %	- 43,0 %

Les moteurs de recherche sont sur-représentés dans les sessions Bibliothèques électroniques, où ils sont 1,8 fois plus présents que dans l'ensemble des sessions. Cet usage fort des moteurs montre l'importance de l'utilisation des bibliothèques électroniques dans un contexte de recherche d'information, où le fonds numérisé est une source parmi d'autres pour trouver de l'information.

Il faut sans doute ajouter à cela les sites personnels, sur-représentés dans les sessions avec Gallica, et qui s'imposent comme sources de données sur des sujets pointus, comme la généalogie. On retrouve cet engouement à travers des requêtes sur les moteurs de recherche très particulières et précises souvent adressées par un seul utilisateur, et qui consistent en des noms propres (patronymes, noms de lieu). La généalogie et le régionalisme s'imposent ainsi comme des centres d'intérêt prépondérants pour une part importante du panel, et Gallica apparaît comme une source d'information parmi d'autres pour ce type de recherches.

L'usage des sites de e-commerce de biens culturels, également corrélé à l'accès aux fonds numérisés, semble correspondre à un effet de catalogue ou de « test avant achat », ce que confirment les entretiens. Nous sommes ici probablement en présence d'une manifestation des liens vertueux pouvant exister entre Web marchand et non-marchand : les bibliothèques numériques permettent de consulter des *fac simile* d'ouvrages anciens, qu'il s'agit ici de feuilleter avant de les acquérir sur un site spécialisé de bibliophilie[1].

[1] On rejoint ici [Gensollen 1999], qui posait dès 1999, à propos de la création de valeur sur Internet, que « la partie non marchande du Web joue sans doute un rôle crucial dans l'économie du système » (p. 19).

En revanche, les sites de type « Média – presse » sont moins présents dans les sessions Bibliothèques électroniques que dans l'ensemble : si les utilisateurs des bibliothèques électroniques sont, comme nous l'avons déjà vu, globalement de gros consommateurs de journaux en ligne, les accès à ces deux types de sites ne correspondent pas aux mêmes pratiques, et se font dans des sessions et des contextes différenciés. Ce résultat illustre bien l'apport d'une analyse fine en termes de sessions, rendue possible grâce à la technologie de recueil de trafic sur le poste de l'utilisateur mise en œuvre dans le projet BibUsages ; en effet, l'analyse de l'audience globale a montré que les usages des bibliothèques électroniques et des sites de type « média / presse » étaient globalement liés, en ce qu'ils sont particulièrement importants pour la population étudiée. L'analyse fine à l'échelle des sessions apporte une information complémentaire importante : même si, globalement, la population étudiée a un intérêt et un usage fort pour ces deux types de sites, les contextes d'usage au sein de sessions de navigation sont bien distincts.

Navigation sur Gallica

Sur les 17 000 sessions enregistrées entre juillet et décembre, 6,1 % comportent un accès à Gallica, qui concernent un nombre important d'utilisateurs : 59 des 72 panélistes de BibUsages sont allés au moins une fois sur Gallica, soit huit personnes sur dix.

L'intensité d'usage de Gallica répond à une répartition similaire à celle du Web : un tiers des visiteurs de Gallica fait plus de 77% des sessions comportant un accès au site, tandis qu'un autre tiers fait seulement 4% de ces sessions. Plus encore, les utilisateurs les plus intensifs de Gallica sont aussi des utilisateurs intensifs du Web : sur la période, ils comptent 400 sessions en moyenne, pour une moyenne de 270 sessions au total pour les autres visiteurs de Gallica, et 150 sessions pour ceux qui n'y sont jamais allés. Ainsi, par son mode de recrutement à partir d'un questionnaire affiché sur le site Gallica, le panel est constitué d'utilisateurs intensifs de ce site, et ce sont eux qui tirent le trafic Web général du panel vers le haut sur l'ensemble de la période, sur Gallica comme sur les autres sites.

Les 1 039 sessions comportant un accès à Gallica sont globalement plus longues que les autres : 1 h 01 min. en moyenne, contre 28 minutes pour les autres sessions. Par ailleurs, dans une session comportant un accès à Gallica, le temps total passé sur ce site proprement dit est en moyenne de 25 minutes, soit presque la durée moyenne d'une session sans Gallica. La visite de la bibliothèque numérique de la BnF s'apparente ainsi à une activité de longue durée.

Plus encore, la consultation de Gallica s'impose comme une activité excluant la visite alternée d'autres sites. Nous avons analysé, dans les sessions, l'alternance de la visite des différents sites, et constaté que dans plus de la moitié des cas, la navigation sur Gallica occupe une seule séquence, elle n'est pas alternée avec la visite d'un autre site ; et pour 22 % des sessions seulement, on compte deux séquences distinctes sur Gallica. Le « multi-tâches » est donc rarement pratiqué, et Gallica induit une activité longue et monolithique.

Au sein de ces longues séquences de navigation, la consultation de documents (texte, image, audio) est majoritaire, et s'appuie sur le duo Recherche – Feuilletage (voir Tableau 6.33). Le téléchargement d'ouvrages, au format TIFF ou PDF, en retrait

par rapport au feuilletage, est très important : dans 38% des sessions Gallica, l'utilisateur est amené à sauvegarder une copie locale d'un document sur son poste, pour un total de plus de 2 000 ouvrages ou extraits d'ouvrages téléchargés.

Tableau 6.33. Services visités sur Gallica

Service	Nombre de sessions	Part des sess. Gallica	Nombre de panélistes
Consultation	628	76,7 %	48
Page Accueil	617	75,3 %	53
Recherche	596	72,8 %	50
Feuilletage d'un ouvrage	548	66,9 %	47
Téléchargement d'ouvrage (tout ou partie)	314	38,3 %	34
Découverte	92	11,2 %	30
Dossiers	89	10,9 %	31
Aide	20	2,4 %	13

La consultation des dossiers reste marginale en nombre de sessions, mais concerne la moitié du panel, ce qui dénote un effet de « visite par curiosité » pour cette rubrique, élément confirmé dans les entretiens :

> *« Je suis allé voir ce qu'il y avait dans les dossiers pour savoir si ça m'intéressait ; il y en a je sais plus combien ; pour l'instant il n'y en a pas qui sont dans mes centres d'intérêt pour le moment, mais je me suis promis un jour où j'aurais du temps d'approfondir un petit peu ça. » (Utilisateur F)*

> *« Je trouve que c'est un peu « gadgétique » et je trouve ça un peu limité. Puis bon… Comme mon domaine est ciblé, si vous voulez… Ça porte souvent sur des petits aspects qui sont pas dans mes préoccupations générales. » (Utilisateur H)*

> *« J'y vais systématiquement rien que pour avoir le plaisir de voir quelques belles enluminures comme ça. Si vous voulez, c'est plus de la recherche, c'est le plaisir de voir de belles choses. » (Utilisateur P)*

Pour la plupart des interviewés, la visite des Dossiers dans Gallica s'inscrit explicitement dans une autre logique de parcours : la flânerie des expositions numériques est opposée à l'efficacité de la recherche ciblée *via* le moteur de recherche interne au site. Les deux contextes d'usage sont ainsi clairement différenciés, le mode « prédateur » étant le plus fréquent.

> ***Synthèse.*** *Chez les panélistes de BibUsages, on observe un lien global entre la fréquentation des bibliothèques électroniques et les outils de recherche d'information d'une part, et les sites de « contenus à lire » et de commerce de biens culturels d'autre part. Les consommateurs de bibliothèques numériques sont de manière générale de grands consommateurs de produits culturels, que ce soit dans une optique de consultation (fonds numériques, journaux) ou de consommation (achat en ligne). Au-delà de cette approche globale, le trafic et les entretiens montrent qu'au sein des sessions, l'usage des bibliothèques électroniques est orienté vers deux types*

de contextes : la recherche documentaire, avec un recours important aux moteurs de recherche et à des sites personnels, et le test avant achat de livres anciens, lié à la visite de sites de bibliophilie et de biens culturels. Dans les deux cas, la visite du site de Gallica est effectuée sur un mode « prédateur » où l'internaute mène une recherche ciblée et feuillette les documents dans une longue période ininterrompue ; partant, flânerie et parcours au gré des liens sont de l'ordre de l'occasionnel.

6.3.2 Le document numérique dans les pratiques

L'accès aux bibliothèques numériques dans un contexte de recherche conduit à reconsidérer le statut des documents numériques proposés. S'il s'agit souvent d'œuvres littéraires dont la lecture est *a priori* linéaire et complète, les contextes d'usage, le profil des utilisateurs et les déclarations faites au cours des entretiens amènent à considérer les fonds électroniques comme des bibliothèques de recherche plus que de lecture.

Lecture parcellaire et ciblée

Lors des entretiens, la lecture en ligne a été le plus souvent évoquée dans le cas des recherches ponctuelles et précises. Elle intervient généralement pour la lecture de petits documents ou pour localiser l'information avant enregistrement. Un cas particulier est à mettre en avant, l'utilisation de *e-books* et de lecture sur PDA[1] :

> *« Lire sur l'écran, d'abord on s'habitue, et puis il y a un mode de saisie que je retrouve un peu sur le Palm en petit format. On a une lecture qui est d'un certain point de vue accélérée parce que le coup d'œil est synthétique. (…) On a une saisie immédiate, évidemment un petit nombre de lignes, il faut que le reader soit bon, mais s'il est bon, c'est au contraire très confortable, on a lu la page d'un coup. Je me souviens d'avoir relu la Chartreuse de Parme de cette façon, avec un plaisir considérable, parce que c'était assez voluptueux comme ça, petites pages par petites pages de saisir le texte de manière très visuelle, presque comme au cinéma. Et ça donne une sensation particulière. » (Utilisateur A)*

Hormis ce cas exceptionnel, la lecture sur écran est peu pratiquée. Jugée « fatigante » par la plupart des interviewés, elle intervient dans le cadre d'une recherche très précise d'informations, de citations ou de termes définis pour des corrections orthographiques, et concerne des portions très ciblées de documents.

Malgré cela, l'impression reste minoritaire, elle n'intervient que pour de petits documents ou des extraits, le coût important étant la cause première invoquée. L'impression permet entre autres d'extraire des parcelles de document et de les traiter de manière plus aisée.

> *« Mais je vais imprimer le chapitre qui m'intéresse, c'est plus facile à visionner sur un papier que sur écran. » (Utilisateur L)*

---

[1] *Personal Digital Assistant* (Palm, Pocket PC).

> *« Sinon j'imprime la page qui m'intéresse avec les références. » (Utilisateur I)*

> *« Euh, au début j'imprimais, j'imprime moins, hein. Au début j'imprimais beaucoup, euh… j'imprime ce qui est utile, hein, par exemple je viens de vous donner un exemple pour Casanova ; quand j'ai repéré, je lis à l'écran, c'est fatiguant mais je lis à l'écran, et quand je repère les pages qui vont me servir, alors là je les imprime. » (Utilisateur O)*

D'autres imprimeront systématiquement, mais cette pratique reste anecdotique sur l'ensemble des panélistes interrogés : en définitive, l'impression comme la lecture sont des pratiques ciblées qui concernent des portions bien délimitées de texte répondant à un intérêt particulier et momentané de l'utilisateur. La consultation de fonds numérisés se rapproche alors définitivement de la recherche ciblée : il n'est pas question de lire les œuvres proposées en ligne (à l'écran ou sur papier), mais d'y puiser des informations précises pour répondre à des besoins particuliers. Ce mode de lecture rapproche les gallicanautes des chercheurs classiques dans leur mode de lecture, où l'ouvrage n'est pas lu de manière linéaire, tandis que l'index et la table des matières sont privilégiés.

Sauvegarder et manipuler le document numérique

Ce mode d'utilisation des documents numériques a deux conséquences dans l'appréhension des contenus en ligne : d'une part la constitution de fonds personnels accumulant les sources potentiellement mobilisables sur un thème donné, et d'autre part le remaniement des documents pour en permettre une manipulation facile.

En ce qui concerne le téléchargement, nous avons déjà constaté que 38 % des sessions sur Gallica comportent une action de téléchargement d'un ou plusieurs ouvrages, ce qui représente un total de plus de 2 000 documents téléchargés par 34 des 53 personnes du panel ayant fréquenté Gallica au cours des six mois d'observation. Les entretiens ont permis de mettre en évidence l'usage régulier voire systématique du téléchargement lorsque le contenu des documents correspond aux besoins de recherche de l'utilisateur. Il s'agit à la fois d'assurer la pérennité des ressources dans l'incertitude relative au Web, et de disposer à tout moment des documents sans avoir à les télécharger à nouveau – problème particulièrement sensible pour les connexions RTC :

> *« Certains sites sont éphémères et je m'en rappelle une fois un site que j'ai téléchargé pour mémoire » (Utilisateur K)*

> *« J'ai vraiment l'habitude d'enregistrer sur le disque dur, de les garder en mémoire » (Utilisateur L)*

> *« Le plus difficile c'est ensuite d'archiver et d'avoir le temps de regarder les archives. » (Utilisateur M)*

Au sein de recherches souvent très ciblées, l'information précise contenue dans un document n'est pas seule sauvegardée, et l'entour de ces informations est également objet de conservation. Le téléchargement va de l'enregistrement de

documents avec conservation des indications d'auteur et de provenance, à l'aspiration de sites entiers.

> *« J'enregistrerai peut-être pas forcément uniquement l'info qui me semble intéressante mais j'enregistrerai la page entière ou si je trouve que le site est intéressant, j'avoue que je copie le site sur le disque dur, ou une partie du site. » (Utilisateur L)*

> *« Si c'est quelque chose que je vais consulter par la suite, donc là, je l'enregistre et je le garde pour après. » (Utilisateur G)*

Un nombre important d'interviewés déclare constituer ainsi des sortes de fonds numériques personnels centrés sur leurs centres d'intérêt propres, dont certains les documentent et en assurent la pérennité en les gravant sur CD-Rom, ce qui atteste une véritable politique de conservation à long terme :

> *« Donc ensuite sous […] Access, je remets : nom d'auteur, titre, et le numéro du Cdrom sur lequel, parce que c'est encore archivé avec des numéros, et quand par exemple je travaille actuellement sur l'évolution de la mécanique, je prends par exemple les trucs de Pierre Duhem qui sont numérisés sauf le volume titre, c'est pas grave, et après en fonction de ça je compulse, alors je compulse directement sur l'ordinateur. Je n'imprime pas. » (Utilisateur M)*

> *"Je télécharge et puis je le stocke. Même, parfois, je ne l'édite pas. Mais je sais que ça existe et puis bon, le jour où je vais rédiger une partie de ce que je fais, bien à ce moment-là, je l'imprimerai ou je le regarderai". (Utilisateur N)*

Cette politique de conservation se situe dans le cadre d'un travail plus général d'accumulation systématique de documents sur un sujet donné. Dans cette perspective, le mode image prépondérant dans le fonds proposé par Gallica est pour beaucoup un frein à la manipulation des documents, et certains procèdent à leur retranscription, manuelle ou par reconnaissance optique de caractères.

> *« Alors quand c'est des textes, c'est facile, avec la recherche on peut voir, on va chercher un mot, hein […] Si vous voulez parler des médecins, vous cherchez médecins, vous allez trouver quelque chose hein. Comme il apparaît un peu de tout, mais seulement on repère tout de suite, hein. En images, allez voir ! » (Utilisateur O)*

> *« Alors, je le télécharge, puis je fais une reconnaissance de caractères, puis je lis tout le texte et je corrige ma reconnaissance de caractères qui va comporter énormément de fautes, bien sûr, puisque les ouvrages donnés en ligne par la Bibliothèque Nationale, ils ont raison de les faire suffisamment légers, il y a un compromis entre la qualité du texte et puis le poids qu'ils vont avoir. Donc ce poids ne permet pas d'avoir quelque chose de parfait, mais ça ne fait rien ! À ce moment là, cet ouvrage-là, en même temps que je le lis, dans mon texte, je corrige. Et après, l'ouvrage, ça y est, j'en fais un autre document PDF mais ça n'a rien à voir avec celui de la Bibliothèque Nationale, il doit être parfait. Je respecte la mise en page du bouquin bien sûr. » (Utilisateur P)*

Ces éléments nous amènent à penser que le statut des documents électroniques est plus proche de celui de l'usuel que de l'œuvre, leur utilité étant définie dans le cadre d'une pratique ciblée et d'interrogations précises. La conservation des documents numérisés et la constitution de fonds électroniques personnels peuvent être ainsi vues comme l'accumulation de documents de référence pour un usage ultérieur, même s'ils ne sont pas effectivement consultés par la suite. L'attitude est bien ici celle de chercheurs balayant leur champ d'investigation, et il est intéressant de constater qu'elle concerne non seulement des enseignants et des chercheurs universitaires, mais également une population de « chercheurs amateurs », qui trouvent avec Gallica et les bibliothèques électroniques en général le moyen d'accéder aux fonds habituellement réservés aux universitaires et aux chercheurs professionnels[1].

> ***Synthèse.*** *Les entretiens renforcent et affinent les éléments mis en avant à l'aide des données de trafic : le recours aux bibliothèques électroniques dans un contexte de recherche se retrouve dans le traitement et la manipulation des documents numériques. La pratique de lectures ciblées, la constitution de fonds personnels et d'archives sur un thème donné, la documentation des corpus personnels sont le reflet d'une manipulation des documents qui les rangent du côté de l'usuel ou de la ressource documentaire plus que de l'œuvre.*

6.3.3 Usages-types

Sur la base des traces d'usages issues du trafic des utilisateurs ainsi que des entretiens menés avec seize d'entre eux, trois « portraits-types » d'utilisateurs ont été identifiés :

- Le chercheur d'information.
- Le bibliophile.
- Le « lecteur à l'écran ».

Il ne s'agit nullement ici de donner une segmentation des utilisateurs, mais plutôt de dresser des profils correspondant à des usages récurrents observés dans notre échantillon.

Le chercheur d'information

Ce qui est mis en avant dans ce type d'usage, c'est bien le rôle de médiation de Gallica à l'intérieur d'Internet. Gallica est une source de documents (de référence) sur le Web, qui permet souvent d'accéder à des informations attestées et de poursuivre sa recherche. On utilise le moteur de Gallica (presque) comme tout autre moteur de recherche sur Internet et les objets visés – les contenus recherchés – sont de nature essentiellement électronique.

---

[1] Pour une comparaison approfondie entre « gallicanautes » et public traditionnel de la BnF, voir [Assadi *et al.* 2003b].

Dans notre panel, nous avons rencontré des utilisateurs menant des recherches dans un cadre professionnel, mais surtout, une majorité de personnes menant des recherches à titre personnel autour d'un hobby, d'une passion, un sujet de recherche très présent étant la mémoire familiale et régionale.

> *« J'étais financier dans des industries. […] Je suis passé sur les recherches sur la famille de Lorraine et la Lorraine d'abord, par le biais de ma famille. Il se trouve que ma famille a été liée à la Maison de Lorraine depuis le XVIIe siècle jusqu'au XIXe, alors c'est comme ça que je suis passé… » (Utilisateur E)*

Ces recherches personnelles occupent une partie importante de leur temps, et donnent parfois lieu soit à des publications, à la constitution d'archives personnelles (CD-ROM) ou à la création de sites personnels.

> *« Mon intention, c'est de rédiger quelque chose, alors je ne sais pas si ça intéressera mon éditeur, parce que ce sera beaucoup moins illustré que mon livre précédent. » (Utilisateur N)*

> *« J'ai publié un petit bouquin il y a une dizaine d'années ; je vais en republier un avant la fin de l'année, un livre pour enfants, il y a le polar qui arrive et qui, je pense, sera pour le mois de mai. » (Utilisateur F)*

Dans ce type d'usage, c'est le texte lui-même qui prime : « Alors le problème, c'est que j'aime bien les beaux livres anciens, c'est très beau dans une bibliothèque, mais les rééditions sont très chouettes aussi. Encore une fois je suis plus intéressé par le contenu que par la relique. » déclare un interviewé. Les questions de manipulation et de documentation des contenus numériques passent alors au premier plan : le mode texte est préféré au mode image, et l'identification des sources et des auteurs valorise les bibliothèques patrimoniales comme Gallica, même si celle-ci peut être mobilisée conjointement avec d'autres ressources sur le Web.

Le bibliophile

Contrairement au premier cas, Gallica joue ici un rôle de médiation vers le monde « réel » (y compris la sphère marchande). Pour le bibliophile , Gallica est utilisé en pré-achat, le but ultime reste l'objet-livre.

> *« Je vais sur le site de la BNF parce que j'y trouve quelque chose, j'ai le souvenir que je pourrai trouver quelque chose, et que je vais me consacrer un peu de temps pour voir quelques images, ou quelques documents que je ne pourrai pas consulter, ou être sûr que je vais acheter un ouvrage ; je parlais de [nom d'auteur], je me suis acheté en consultant la BNF, ça m'a beaucoup intéressé, car j'ai acheté un ouvrage qui coûtait quand même 1 200 F. » (Utilisateur H)*

Dans ce cas, l'impression ne supplante pas l'ouvrage relié, et l'attrait pour les ouvrages anciens est à mettre en parallèle avec la valorisation du livre traditionnel en général :

> *« Il y a un côté sensuel du livre, le toucher, le prendre ; il y a la faculté de le prendre à n'importe quel moment, qu'on fasse beaucoup de Kms, prendre un bouquin et en lire trois pages, c'est aussi reposant que pendant une demi-heure,*

> *en avion, un peu partout, dans le train parfois. Donc c'est agréable. Et puis il y a surtout le plaisir de partager, de prêter un livre, ou se le faire prêter, c'est une sorte de convivialité très importante. » (Utilisateur H)*

Le format électronique ne supplante pas l'édition papier dans son statut : « Le bouquin, c'est quand même mieux ; il y a une pérennité aussi, ça se conserve, ça c'est le côté objet que l'on possède », déclare un interviewé. Ici, au-delà des arguments d'ordre pratique (maniabilité, prix, conservation), le livre reste un objet culturellement valorisé en tant que tel.

Le lecteur à l'écran

Enfin, nous avons choisi d'évoquer ici un profil plutôt atypique, correspondant à un usage rare d'après nos observations. Il s'agit d'utilisateurs ayant une pratique intensive de la lecture à l'écran. Dans ce cas, Gallica – ou tout autre site offrant des collections de textes et d'ouvrages – est considéré comme un lieu de "consommation" et non pas de médiation : « J'écume les sites où on peut télécharger des textes, c'est ma passion. » déclare l'utilisatrice interviewée qui correspond à ce profil.

> *« La motivation de la fréquentation est toujours le téléchargement de textes ; je recherche bien sûr des textes classiques ; je suis agrégée de lettres au départ. Donc c'est un peu normal. Donc c'est par rafales, brusquement l'idée me vient, je suis assez fantasque, l'idée me vient tout d'un coup de relire Scaron ; je l'ai dans ma bibliothèque, mais maintenant je lis de la sorte sur écran, et j'ai besoin du texte tout de suite, et donc je vais voir s'il est disponible à la BNF, et si l'ouvrage que je recherche n'est pas là, je vais chercher ailleurs. » (Utilisateur A)*

On note ici un déplacement des pratiques de lecture intensive du papier vers l'écran, accompagné d'une redécouverte du plaisir lié à la lecture et à l'appréhension de l'objet-livre. L'intérêt se porte sur les œuvres littéraires, qui impliquent une lecture globale et linéaire, et dont le support de lecture renouvelle l'appréhension.

Le livre renouvelé par le Web

En somme, dans le cadre des pratiques observées, le statut du document numérique semble plus proche de l'usuel que de l'œuvre. Sa consultation et son usage s'inscrivent le plus souvent dans le cadre de recherches personnelles sur des sujets précis : les textes consultés sont majoritairement envisagés comme support de recherche bien plus que comme objet de lecture. Il est frappant de constater combien ces recherches personnelles sont menées avec rigueur et méthode : démarche de recherche systématique (avec en particulier un usage avancé des moteurs de recherche sur le Web), classement rigoureux des documents allant jusqu'à la constitution de fonds documentaires personnels, tendance à privilégier les éditions de référence (ce qui représente un des motifs du recours à Gallica). En ce sens, les fonds numériques tendent à rendre disponibles, au sein des ressources accessibles sur Internet, des documents et des possibilités de recherche nouveaux, et élargissent ainsi le public des chercheurs à l'amateur et à l'érudit. On souscrit ici à l'analyse faite par Chartier dans [Chartier 2003], qui projette une coexistence entre écriture manuscrite, publication imprimée et textualité électronique, où cette dernière

organise « de manière nouvelle la relation entre la démonstration et les sources, les modalités de l'argumentation et les critères de la preuve ».

Ce public semble assez différent de celui des bibliothèques classiques, et les chercheurs « professionnels » y sont comparativement peu représentés. Les plus de quarante ans, actifs ou retraités, sont majoritaires dans la population observée et les bibliothèques électroniques sont avant tout pour eux une source d'informations dans le cadre de recherches personnelles. L'intensité d'usage est ici bien supérieure à celle de la population générale des internautes français et va de pair avec un très fort taux d'équipement en haut débit (câble, ADSL). Nous découvrons ici une population d'internautes seniors fortement équipés en haut débit, et dont les centres d'intérêt, outre l'offre de services et de communication classique, gravitent autour des contenus « culturels ». Cette population atypique dans le paysage des internautes français constitue en elle-même un groupe d'autant plus intéressant qu'il est mal identifié dans les études d'usages à large panorama car minoritaire.

Si d'une manière générale les utilisateurs des bibliothèques électroniques sont également de forts consommateurs de « contenus à lire » (journaux en ligne en particulier), au sein des sessions, l'usage des bibliothèques numériques est fortement corrélé à celui des moteurs de recherche d'une part, et des sites de vente de biens culturels d'autre part. Deux profils se dégagent : celui du « chercheur amateur », dont les centres d'intérêt sont pointus et déjà bien connus de l'utilisateur, et celui du bibliophile pour qui Gallica fait office de catalogue avant achat. Dans les deux cas, la lecture en ligne est rare, tout autant que l'impression des documents téléchargés, et la lecture s'apparente à la recherche de fragments ciblés au sein de vastes collections laissant de côté la totalité des œuvres. Dans ce cadre, le statut des documents en ligne semble remis en cause : tandis que l'édition papier reste du côté de l'œuvre, l'édition électronique s'apparente à l'usuel.

L'étude des pratiques en contexte souligne également les passerelles entre Web marchand et non marchand pour les utilisateurs. Alors que les acteurs d'Internet (fournisseurs de contenus et d'accès) perçoivent une dichotomie forte entre sites marchands et non marchands, les internautes passent indifféremment d'un type de site à un autre, et l'on doit plutôt parler d'enrichissement mutuel entre sites marchands et non marchands dès lors qu'on les envisage sous l'angle des pratiques.

Dès lors, attirant de nouveaux publics, induisant de nouveaux modes d'appréhension des textes, s'inscrivant dans des parcours de lecture inédits, les bibliothèques électroniques, loin d'être une simple version numérisée des fonds traditionnels, s'apparentent à un nouvel espace de lecture et de consultation aux côtés des bibliothèques classiques.

> ***Synthèse.*** *Trois profils-type d'utilisateurs des bibliothèques électroniques se dégagent : le chercheur d'information, pour qui les fonds numérisés constituent une ressource parmi d'autres ; le bibliophile, plutôt tourné vers l'objet-livre et l'achat ; et le lecteur à l'écran, profil rare renouvelant le plaisir de la lecture sur des terminaux informatiques. Dans les trois cas, le livre-papier conserve le statut d'œuvre, tandis que l'édition électronique (*fac-simile *ou texte manipulable), plongée dans le contexte du Web et profondément influencée par celui-ci, s'inscrit dans de nouvelles formes d'usage où prime avant tout le texte.*

Conclusion

Dans le Chapitre 5, nous avons construit une typologie des parcours sur la base de leur forme et de leur temporalité, et observé les corrélations que ces parcours-type entretiennent avec les contenus ; replacés dans le contexte des pratiques individuelles, ces modes de navigation prototypiques prennent un sens particulier. Ils ne sont effectivement liés ni à l'ancienneté de la pratique d'Internet, ni au type de connexion, ni aux types de pratiques d'Internet en général, mais semblent dépendre de contextes particuliers liés à l'activité locale de l'utilisateur. La notion de territoire personnel sur le Web nous permet de cerner ces contextes et de leur donner sens : la vue longitudinale sur trois ans qu'offrent les données SN00-02 permet de voir comment sont structurés les espaces individuels de navigation tant du côté des thèmes ou services accédés que des sites visités. Dans les deux cas, il apparaît qu'un très petit nombre de sites constitue le support principal de la navigation, quel que soit le contexte de l'utilisateur. On le retrouve dans les sessions routinières, renvoyant à un territoire restreint fréquemment balayé, ainsi que dans un deuxième ensemble de sessions impliquant des sites vus de manière occasionnelle mais régulièrement dans un contexte donné. Plus encore dans un dernier groupe de sessions tournées vers la découverte de nouveaux sites, dont la plupart ne seront jamais revus par la suite, c'est encore le noyau dur des sites habituels de l'internaute qui sont le support de la découverte et du mouvement vers l'inconnu. En définitive, et contrairement à l'idée souvent évoquée de navigation dans l'hypertexte au gré des liens, l'internaute reste le plus souvent dans un espace qui lui est familier et qui sert de pivot à l'ensemble de ses parcours.

Confrontée à la segmentation des sessions en parcours-type, cette description des sessions par le biais des territoires se révèle très productive : le croisement de ces deux typologies met à jour des comportement de navigation en situation intégrant la double dynamique des contenus et des visites par l'utilisateur. Trois modes d'appréhension prototypiques des contenus du Web sont à distinguer : en premier lieu, le mode routinier, lié aux parcours éclairs et ciblés, fonctionne sur le mode d'un balayage régulier et fréquent de ressources connues qui ont la particularité, sur le plan du contenu, d'être des flux – flux d'information (informations générales, boursières, sites spécialisés) ou flux de communication (WebMail, forums, *chat*). Ensuite, un mode « galerie marchande » ou « guichet de renseignements », qui intervient de manière moins fréquente, implique des parcours orientés vers le « Web pratique », plus complexes et plus allongés ; ces contextes particuliers surviennent peu fréquemment, mais les comportements sont très stables dans ces contextes en termes de sites visités. Enfin, les parcours ouverts, qui amènent à découvrir de nouveaux sites, font massivement appel aux moteurs de recherche, dans le cadre de requêtes précises, et sont liés aux parcours à pivots ou parcours éclatés.

Ces trois modes prototypiques d'appréhension du Web contrastent à la diversité thématique et surtout fonctionnelle des contenus Web : services proposés, outils de recherche, fréquence de mise à jour, type de publication sont autant d'éléments qui suggèrent une grande diversité dans les pratiques de la Toile et leur possible mobilisation dans des chaînes opératoires très diverses. Pour autant, la mise à jour de modes génériques d'appréhension des contenus atteste la prévalence des pratiques et des modes d'appréhension dans la valorisation des contenus : malgré des modes de

navigation parfois prédéfinis de manière très contrainte par leurs concepteurs, les sites se caractérisent avant tout par ce qu'en font les internautes.

L'usage massif des bibliothèques électroniques dans des contextes de recherche documentaire ou d'achat va dans ce sens : on rejoint ici F. Rastier qui proposait de « formuler un principe d'*architextualité* : tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent »[1]. En étendant ce principe à Internet, on peut ainsi affirmer que tout contenu plongé dans le Web s'expose à en subir l'influence non seulement présentationnelle, formelle et fonctionnelle, mais aussi perceptuelle, interprétative et usuelle. La juxtaposition au sein d'une même interface de l'ensemble de l'hypertexte opère ainsi sur les sites une mise à plat, que les parcours viennent assembler au fil des pratiques individuelles et des usages collectifs ; en somme, ce n'est pas tant le contenu des sites qui donne une signification au parcours que la somme des parcours qui confère un sens individuel ou collectif aux contenus du Web au sein de pratiques stabilisées.

[1] [Rastier 2001a], p. 92.

Conclusion, perspectives

Pour clore ce travail, nous souhaitons revenir sur quelques résultats importants que nous avons avancés, tant dans l'analyse des pratiques que sur le plan méthodologique, et tracer sur ces deux axes quelques perspectives de recherche.

1. Modes de navigation

Genres de parcours

Au terme de notre analyse, nous serions presque tentés, en suivant F. Rastier, de parler de *genres* de parcours sur le Web, c'est-à-dire de modalités normées de la pratique du Web partagées par l'ensemble des individus, et spécifiques dans leur structure et leur contenu. Ces modalités d'usage du Web « héritent » de pratiques existantes : au sein des trois modes prototypiques de parcours que nous avons mis à jour, l'appréhension de contenus apparentés à des flux renvoie à la consommation de médias de masse ; les parcours orientés vers les services et les sites de e-commerce rappellent les galeries marchandes ou les guichets de renseignement ; les parcours de recherche évoquent la compulsion de sources diverses sur un problème donné. Ces similitudes sont favorisées par le fait que, souvent, les contenus Web sont la transposition en ligne de services préexistants : messagerie, courses, bibliothèques.

Pour autant, le parallèle s'arrête là : de manière générale, la sociologie des usages nous montre que la situation d'action, le format des outils manipulés, la structure des interfaces importent autant que les fonctionnalités des outils techniques. Dans le détail de l'activité sur le Web, la spécificité ergonomique et fonctionnelle des interfaces Web nous amène à découvrir une métrique particulière de ces modes d'appréhension, que l'on retrouve chez l'ensemble des internautes : les temporalités, les formes spécifiques de parcours, les outils et services particuliers mobilisés dans ces différents contextes permettent de les différencier. À chaque type de pratique, correspondent un mode de déplacement particulier au sein de l'hypertexte, et la mobilisation de territoires spécifiques par chaque individu.

Le contenu modelé par l'usage

Le Web suscite également une mise à plat de l'ensemble des contenus *via* une même interface logicielle et ergonomique : l'adjacence de contenus hétérogènes, intrinsèquement ergonomique et renforcée par l'hypertexte, opère une « mise en corpus » qui influence leur mise en forme, leur valorisation, leur statut. Un double mouvement se produit : certes, ce matériau particulier détermine le format des pratiques sur le mode d'un « espace des possibles », mais c'est à travers ces pratiques et dans les situations particulières d'usage que les contenus prennent sens.

En premier lieu, les formats traditionnels se trouvent mobilisés différemment une fois plongés dans le contexte du Web : on a pu voir, dans le cas des bibliothèques électroniques, que les textes sont utilisés comme ressource pour la recherche ou pour l'achat, et que les œuvres sont, du point de vue de l'usage, mobilisées comme des usuels. D'autre part, une même ressource sera mobilisée et valorisée de manière différente par plusieurs utilisateurs, ou même par un même utilisateur dans deux contextes différents : on citera l'exemple des moteurs de recherche, tantôt employés comme pseudo-bookmarks lorsqu'un site particulier est visé, tantôt comme outil exploratoire de zones inconnues sur le Web. Certes, certains contenus se prêtent à des usages ciblés, d'autres sont plus ouverts, mais la tendance générale est à la plasticité : une forme de normalisation des interfaces Web fait se côtoyer dans une même page un moteur de recherche interne au site, des bandeaux de navigation, et le contenu unique de la page lui-même. Devant cette multiplicité des possibilités d'interaction avec les interfaces et de leur mobilisation dans des situations variées, c'est au sein des cours d'action organisant la navigation que se construit le sens des contenus du Web.

Dans ce cadre, le corpus de parcours s'apparente à un corpus d'expériences où l'observation de régularités fait émerger les usages. Notre travail met en évidence, au sein du Web, des *structures vécues* qui diffèrent profondément de la structure hypertextuelle sous-jacente : l'appréhension l'emporte sur la proposition et la mise à disposition. La structure et la dynamique des territoires personnels sur la Toile montrent que cette appropriation des contenus se structure autour d'une poignée de sites récurrents qui occupent, quel que soit le contexte, la majorité de la durée des parcours. On est ici très loin de l'idée de surf ou de butinage ; il est frappant de constater que les sessions sont la plupart du temps dédiées à un cours d'action unique, et lorsque ce n'est pas le cas, les différents cours d'action sont rarement entrelacés. Le parcours apparaît bien comme l'accomplissement d'un projet dans le cadre de structures d'actions normées exclusives les unes des autres.

2. Données de trafic

Méthodologie adaptée

Des données de trafic brutes à l'analyse fine des usages en situation, nous avons développé une méthodologie et un outillage dont nous souhaitons souligner trois aspects importants.

En premier lieu, notre approche est globale, à plus d'un titre. Les deux panels issus des projets TypWeb et SensNet sont représentatifs des usages à domicile des internautes français, et leur taille et leur durée d'observation permettent d'atteindre une masse critique nécessaire au repérage de régularités de comportements.

La globalité se retrouve également dans l'exhaustivité des données recueillies : si les panélistes avaient à tout moment le loisir de suspendre le recueil de données par les sondes, la variété des contenus que nous observons amène à penser qu'ils ne l'ont que très peu fait. Loin de réduire cette diversité, nous avons cherché au contraire à en

rendre compte autant que possible, en observant systématiquement l'ensemble des parcours analysables dans la construction des typologies.

Enfin, l'approche globale se traduit également par une mise en contexte des parcours au sein de situations de navigation et de territoires personnels construits sur le long terme. Nous avons pour cela élaboré et mis en œuvre un appareil analytique qui tient autant que faire se peut l'ensemble des constituants d'un parcours, de la page à l'individu, au sein d'un même objet et d'une même démarche.

Deuxième élément méthodologique notable, nous nous sommes efforcés de bâtir, à ces différents paliers d'analyse, des descriptions adaptées aux contenus Web et à leur mode d'appréhension.

Sur le plan des contenus, c'est en considérant le Web comme outil et non uniquement comme support que nous avons pu mener à bien une approche praxéologique. La Toile, trop souvent perçue comme simple vecteur d'informations, est aussi un ensemble de dispositifs d'interaction proposant des services, des outils de recherche, des systèmes de gestion de contenu, etc. La description fonctionnelle des pages et des sites avec les annuaires du Web et le module CatService a permis, au-delà des thématiques et des centres d'intérêt propres à chaque individu, de cerner des modes de navigation s'appuyant sur ces éléments fonctionnels des contenus Web.

Sur le plan de l'activité de navigation, nous avons tenté de rendre compte de manière simple des éléments dynamiques de la pratique : les indicateurs que nous avons construits permettent d'appréhender la topologie des parcours, leur rythmique et leur temporalité afin d'intégrer dans la description des éléments de la « gestuelle » navigationnelle. Cette description est ainsi cohérente avec une vision du Web comme espace d'action que nous mettons en avant sur le plan des contenus.

Enfin, dans la mobilisation de ces différents descripteurs, nous avons dû tenir compte de leur hétérogénéité et de leur structure. Pour les rendre manipulables par des outils de statistique descriptive et exploratoire, nous avons opéré des découpages et des regroupements *ad hoc* dans les variables, en prenant systématiquement en compte la réalité des pratiques sous-jacentes. Ceci est particulièrement vrai pour la mobilisation des indicateurs topologiques, où les variables continues masquent, à certaines valeurs, des réalités bien distinctes. Cela se retrouve également à une échelle plus globale dans la manipulation des données de trafic : quelles que soient les variables considérées (intensité de trafic, sites visités, durées, etc.), les courbes de distribution ont quasi-systématiquement des allures zipfiennes, ce qui conduit à opérer des regroupements et des discrétisations si l'on veut éviter les interprétations erronées.

Parcours : variables actives

À terme, le travail exploratoire et descriptif que nous avons mené permet de dresser, au sein de la profusion des descripteurs, une liste des variables pertinentes pour définir un parcours. Nous en comptons principalement trois :

- *morphologie* : la forme d'un parcours en est une composante essentielle, et elle représente un bon indice du cours d'action auquel celui-ci est soumis. Pour le représenter en termes statistiques, on pourra à gros grain se contenter d'envisager la session à l'échelle des sites uniquement, en mobilisant quatre

descripteurs principaux : la durée de la session, le nombre de sites visités, le taux de linéarité et le degré de concentration sur sites revisités.

- *types de contenus* : la description des parcours s'appuiera particulièrement sur une description fonctionnelle des contenus, en termes de services et d'outils de navigation. Les aspects thématiques sont ici secondaires, dans la mesure où ils sont soumis à d'importantes variations inter-individuelles.
- *territoires sur le Web* : ce dernier point, qui implique une vue longitudinale des usages individuels, est particulièrement éclairant pour comprendre la valorisation des contenus proposés au sein des pratiques. Sur la base d'une distinction entre fréquence et régularité, on scinde le corpus de sites d'un utilisateur en espaces routiniers, occasionnels et exploratoires. Ces trois grandes zones éthologiques tracent au sein de l'hypertexte des espaces de compréhension qui confèrent aux parcours leur valeur d'usage.

Ces trois types de variables permettent de rendre compte des genres de parcours en tenant compte, autour des éléments de la dynamique locale production/réception, du support de l'action et de son inscription dans les pratiques individuelles.

3. Pour aller plus loin

La description des parcours que nous proposons pourra servir, nous l'espérons, de point d'appui à d'autres travaux. Nous voyons pour notre part deux axes de recherche majeurs pour prolonger ce travail, vers la formalisation d'une part et l'affinage de l'autre.

Formaliser

Le premier axe de recherche touche l'exploitation des données de trafic elles-mêmes. Le travail de description et de segmentation que nous avons réalisé doit pouvoir servir de base à des approches plus formelles, notamment pour une représentation des parcours sous forme de séquences et de graphes. Les représentations synthétiques des contenus et de leur place dans les territoires personnels, qui créent de la redondance dans les données, réintroduisent également la possibilité de mettre en œuvre des analyses de type chaînes de Markov, séries temporelles, etc.

D'autre part, nous nous sommes attachés à décrire les variables que nous avons mobilisées, les seuils utilisés pour les discrétiser, le poids des variables et des modalités dans la construction des classes. Tout cela doit pouvoir rendre réutilisable ce travail non plus dans la perspective d'une classification, mais pour le classement automatique de sessions. Sur cette base, les cinq parcours-types et les trois modes d'appréhension prototypiques que nous avons mis au jour peuvent notamment, une fois reconnus, autoriser des traitements différenciés par la suite, par exemple pour la mise en œuvre d'agents d'aide à la navigation, ou l'adaptation dynamique des contenus des sites.

Enfin, ces développements sont appelés à s'intégrer au sein d'une plateforme de catégorisation des usages d'Internet. Les outils de visualisation font déjà l'objet, dans le cadre du projet SensNet, d'une intégration à une telle plateforme ; les segmentations produites peuvent également être introduites sous forme de nouvelles variables, et compléter la description des données pour la fouille et la synthèse.

Affiner et compléter la description

Le second axe de recherche s'oriente vers une description plus fine des usages et des pratiques en situation. Au sein des parcours, on souhaiterait disposer d'une typologie des contenus plus fine et adaptée aux contenus Web : nous avons déjà exposé les travaux sur les genres et les types de pages Web, leur avancée apporterait à l'analyse des parcours une source descriptive précieuse. Les recherches en cours au sein du projet SensNet vont dans ce sens, en tentant d'identifier les traits caractéristiques de types de pages et de sites motivés sur le plan fonctionnel.

Dans la même perspective, on souhaiterait descendre plus bas dans la description des interactions avec les interfaces, et examiner dans quelle mesure les éléments ergonomiques locaux de manipulation des dispositifs sont soumis aux déterminations contextuelles générales que nous avons identifiées. Pour cela, des dispositifs tels que la capture d'écran ou le développement de sondes interceptant les événements sur les IHM (Interfaces Homme Machine) fourniraient des données précieuses pour apprécier les mouvements et les rythmiques locales à l'échelle de la page ou du site.

Enfin, il apparaît souhaitable de compléter l'analyse restreinte au Web par la prise en compte dans les sessions des agencements avec les autres outils Internet d'une part et l'entour de l'utilisateur de l'autre. Ces deux éléments concourent à une appréhension de l'articulation des différents médias et de la multimodalité en situation. La mobilisation de l'ensemble des données recueillies par les sondes et le recours à des dispositifs externes tels que la vidéo constituent à nos yeux de bonnes pistes pour avancer dans cette problématique.

Bibliographie

ACHARYYA, S. et GHOSH, J. (2003), Context-Sensitive Modeling of Web-Surfing Behaviour Using Concept Trees, Actes de *Proceedings of the Fifth WEBKDD workshop: Webmining as a Premise to Effective and Intelligent Web Applications (WEBKDD'2003)*, Washington, USA, http://www.acm.org/sigkdd/proceedings/webkdd03/wkdd03-paper1.pdf.

ADAMIC, L. A. (1999), The Small World Web, Actes de *ECDL'99*, Springer.

ADAMIC, L. A. (2001), Networks dynamics: the World Wide Web, (Stanford University), http://www.hpl.hp.com/shl/people/ladamic/thesis/ladamic_thesis.pdf.

ADAMIC, L. A. et HUBERMAN, B. A. (2001), The Web's Hidden Order, *Communications of the ACM* 44(9), http://www.hpl.hp.com/research/papers/weborder.pdf.

AMITAY, E. (1997), Hypertext : the importance of being different, MSc Dissertation, University of Edinburgh, (Centre for Cognitive Science).

AMITAY, E. (1999), Anchors in Context: A corpus analysis of web pages authoring conventions, *in* L. Pemberton et S. Shurville, *Words on the Web - Computer Mediated Communication*, U.K., Intellect Books.

AMITAY, E., CARMEL, D., DARLOW, A., LEMPEL, R. et SOFFER, A. (2003), The Connectivity Sonar: Detecting Site Functionality by Structure Patterns, Actes de *Fourteenth conference on Hypertext and Hypermedia (Hypertext'03)*, ACM, http://www.ht03.org/papers/pdfs/5.pdf.

AMITAY, E. et PARIS, C. (2000), Automatically Summarising Web Sites - Is There A Way Around It?

ASSADI, H. et BEAUDOUIN, V. (2002), Comment utilise-t-on les moteurs de recherche sur Internet ?, *Réseaux* 20(116), pp. 171-198.

ASSADI, H. et BEAUVISAGE, T. (2002), A comparative study of six French-language Web directories, Actes de *ISKO 2002*, Granada, Spain.

ASSADI, H., BEAUVISAGE, T., DE CHARENTENAY, F., CLOAREC, T., LUPOVICI, C. et TOUTUT, H. (2003a), Usages des bibliothèques électroniques en ligne. Projet BibUsages, France Télécom R&D.

ASSADI, H., BEAUVISAGE, T., LUPOVICI, C. et CLOAREC, T. (2003b), Users and uses of online digital libraries in France, Actes de *Research and Advanced Technology for Digital Libraries. 7th European Conference on Digital Libraries (ECDL 2003)*, Trondheim, Norway, Springer.

BAEZA-YATES, R. et POBLETE, B. J. (2003), Evolution of the Web Structure, Actes de *WWW Conference 2003*, http://www.www2003.org/cdrom/papers/poster/p103/p103-baeza-yates/p103-baeza-yates.html.

BALPE, J.-P., LELU, A., PAPY, F. et SALEH, H. (1996), *Techniques avancées pour l'hypertexte*, collection Paris.

BEAUDOUIN, V., ASSADI, H., BEAUVISAGE, T., LELONG, B., LICOPPE, C., ZIEMLICKI, C., ARBUES, L. et LENDREVIE, J. (2002), Parcours sur Internet : analyse des traces d'usage, France Télécom R&D.

BEAUDOUIN, V., BEAUVISAGE, T., CARDON, D. et VELKOVSKA, J. (2003a), L'entrelacement des médias dans la constitution des publics de Loft Story, France Télécom R&D.

BEAUDOUIN, V., FLEURY, S., HABERT, B., ILLOUZ, G., LICOPPE, C. et PASQUIER, M. (2001), TyPWeb : décrire la Toile pour mieux comprendre les parcours, Actes de

CIUST'01, Colloque International sur les Usages et les Services des Télécommunications, Paris, France.

BEAUDOUIN, V., FLEURY, S., PASQUIER, M., HABERT, B. et LICOPPE, C. (2003b), Décrire la Toile pour mieux comprendre les parcours, *Réseaux*(116), pp. 19-51.

BEAUDOUIN, V. et VELKOVSKA, J. (1999), Constitution d'un espace de communication sur Internet, *Réseaux* 17(97), pp. 121-177.

BERNERS-LEE, T., HENDLER, T. et LASSILA, O. (2001), The Semantic Web, *Scientific American*(276).

BIDEL, S., LEMOINE, L., PIAT, F., ARTIRES, T. et P., G. (2003), Statistical machine learning for tracking hypermedia user behavior, Actes de *MLIRUM'03: Second Workshop on Machine Learning, Information Retrieval and User Modeling*, http://www.cs.rutgers.edu/mlirum/mlirum-2003/final/Bidel.pdf.

BORGES, J. et LEVENE, M. (1998), Mining Association Rules in Hypertext Databases, Actes de *4th International Conference on Knowledge Discovery and Data Mining*.

BORGES, J. et LEVENE, M. (1999), Data Mining of User Navigation Patterns, Actes de *WEBKDD'99*.

BORGES, J. et LEVENE, M. (2000), A Fine Grained Heuristic to Capture Web Navigation Patterns, *SIGKDD Explorations* 2(1), pp. 40-50.

BRETAN, I., KARLGREN, J., HALLBERG, A. et WOLKERT, N. (1998), Web-specific genre visualization, Actes de *3rd World Conference on the WWW and Internet*, Orlando, http://www.sics.se/~jussi/Papers/1998_WebNet_DropJaw/dropjaw_webnet98.pdf.

BROADBENT, S. et CARA, F. (2003), Les nouvelles architectures de l'information, *in*, *Text-e. Le texte à l'heure de l'Internet.*, Bibliothèque Centre Pompidou.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A. et WIENER, J. (2000), Graph structure in the Web, Actes de *Nineth International World Wide Web Conference*, Amsterdam, The Netherlands.

BRUSILOVSKY, P. (1996), Methods and Techniques of Adaptive Hypermedia, *User Modeling and User Adapted Interaction* 6(2-3), pp. 87-129, http://www.ifi.ntnu.no/~oshea/pdf/brusilovsky96methods.pdf.

BRUSILOVSKY, P. (2001), Adaptive Hypermedia, *User Modeling and User Adapted Interaction* 11(1-2), pp. 87-110, http://www2.sis.pitt.edu/~peterb/papers/brusilovsky-umuai-2001.pdf.

BUSH, V. (1946), As we may think, *The Atlantic Monthly* 176(1), pp. 101-108.

BYRNE, M. D., JOHN, B. E. et JOYCE, E. (1999a), A day in the life of ten WWW users.

BYRNE, M. D., JOHN, B. E., WEHRLE, N. S. et CROW, D. C. (1999b), The tangled Web we wove: a taskonomy of WWW use, Actes de *CHI'99 Human Factors in Computing Systems*, ACM press.

CANTER, D., RIVERS, R. et STORRS, G. (1985), Characterizing user navigation through complex data structures, *Behavioural and Information Technology*(4), pp. 93-102.

CATLEDGE, L. D. et PITKOW, J. E. (1995), Characterizing browsing strategies in the World-Wide Web, *Computer Networks and ISDN Systems* 27(6), pp. 1065-1073.

CHARTIER, R. (2003), Lecteurs et lectures à l'âge de la textualité électronique, *in*, *Text-e. Le texte à l'heure de l'Internet.*, Bibliothèque Centre Pompidou.

CHEVALIER, K., BOTHOREL, C. et CORRUBLE, V. (2003), Discovering Rich Navigation Patterns on a Web Site, Actes de *Discovery Science 2003*, Springer.

CHOO, C. W., DETLOR, B. et TURNBULL, D. (1999), Information Seeking on the Web - An Integrated Model of Browsing and Searching, Actes de *ASIS '99 Annual Meeting*, Washington DC, USA, http://donturn.fis.utoronto.ca/papers/asis99/asis99.html.

CHOO, C. W., DETLOR, B. et TURNBULL, D. (2000), Working The Web: An Empirical Model of Web Use, Actes de *33rd Hawaii International Conference on System Science*

(HICSS), Maui, Hawaii, http://donturn.fis.utoronto.ca/papers/hicss2000/hicss2000.html.

COCKBURN, A. et MCKENZIE, B. (2000), What do Web users do ? An empirical analysis of Web use, Actes de *International Journal of Human-Computer Studies*, http://www.cosc.canterbury.ac.nz/~andy/papers/ijhcsAnalysis.pdf.

COOLEY, R., MOBASHER, B. et SRIVASTAVA, J. (1997), Grouping web page references into transactions for mining world wide web browsing patterns, Minneapolis, USA, Dept. of Computer Science, University of Minnesota.

COOLEY, R., MOBASHER, B. et SRIVASTAVA, J. (1999a), Data Preparation for Mining World Wide Web Browsing Patterns, *Knowledge and Information Systems* 1(1).

COOLEY, R., TAN, P.-N. et SRIVASTAVA, J. (1999b), Websift: The Web Site Information Filter System, Actes de *1999 KDD Workshop on Web Mining*, San Diego, CA.

COTTE, D. (2002), L'approche néophyte de la page Web. Ou "Mais où je clique là ?" *Les Cahiers du Numérique* 3(3), pp. 17-32.

CROVELLA, M. E. et BESTAVROS, A. (1996), Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes, Actes de *Conference on Measurement and Modeling of Computer Systems (Sigmetrics)*, ACM.

CROWSTON, K. et WILLIAMS, M. (1999), The effects of linking of Web documents, Actes de *32nd Hawaii International Conference on System Sciences*, Maui, Hawaii, http://crowston.syr.edu/papers/ddgen04.pdf.

CUNHA, C., BESTAVROS, A. et CROVELLA, M. E. (1995), Characteristics of WWW Client-based Traces, Computer Science Department, Boston University.

DANIELSON, D. (2003), Transitional Volatility in Web Navigation, *ITC&Society* 1(3), pp. 131-158, http://www.stanford.edu/group/siqss/itandsociety/v01i03/v01i03a08.pdf.

DILLON, A. et GUSHROWSKI, B. (2000), Digital genres and the web: is the home page the first digital genre?, *Journal of the American Society for Information Science* 51(2), pp. 202-205, http://www.ischool.utexas.edu/~adillon/publications/genres_web.pdf.

DIMAGGIO, P., HARGITTAI, E. et NEUMAN, R. (2001), Social Implications of the Internet, *Annual Review of Sociology* 27, pp. 307-336, http://webuse.umd.edu/handouts/publications/ARS2001.pdf.

DIMITROVA, M., FINN, A., KUSHMERICK, N. et SMYTH, B. (2002), Web genre visualization, *Conference on Human Factors in Computing Systems (CHI'2002)*, Minneapolis, USA, http://www.cs.ucd.ie/staff/nick/home/research/download/dimitrova-chi2002.pdf.

FALOUTSOS, M., FALOUTSOS, P. et FALOUTSOS, C. (1999), On power-law relationships of the Internet topology, Actes de *Conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '99)*, Cambridge, USA.

GENSOLLEN, M. (1999), La création de valeur sur Internet, *Réseaux* 17(97), pp. 15-76.

GHITALLA, F., BOULLIER, D., GKOUSKOU-GIANNAKOU, P., LE DOUARIN, L. et NEAU, A. (2003), *L'outre-lecture. Manipuler, (s')approprier, interpréter le Web*, collection Etudes et recherches, Paris, Bibliothèque publique d'information/Centre Pompidou.

HEER, J. et CHI, E. (2002), Separating the Swarm: Categorization Methods for User Sessions on the Web, Actes de *Proceedings of ACM CHI 2002 Conference on Human Factors in Computing Systems*, Minneapolis, USA, http://www-users.cs.umn.edu/~echi/papers/2002-CHI/UIR-R-2001-05-Heer-CHI2002-ScentMMC.pdf.

HÖLSCHER, C. et STRUBE, G. (2000), Web search behavior of Internet experts and newbies, Actes de *9th World Wide Web Conference (WWW'9)*, http://www9.org/w9cdrom/81/81.html.

HUBERMAN, B. A., PIROLLI, P. L. T., PITKOW, J. E. et LUKOSE, R. M. (1998), Strong Regularities in World Wide Web Surfing, *Science* 280(5360), pp. 95-97,

http://www.parc.xerox.com/istl/projects/uir/pubs/pdf/UIR-R-1998-06-Pitkow-Science-Surfing.pdf.

IVORY, M. et HEARST, M. (2002), Statistical Profiles of Highly-Rated Web Sites, Actes de *CHI 2002*.

JANSEN, B. J., BATEMAN, J. et SPINK, A. (1998a), Searching heterogeneous collections on the Web : behaviour of Excite users, *Information Research* 4(2).

JANSEN, B. J., SPINK, A. et TEFKO BATEMAN, S. (1998b), Real Life Information Retrieval: A Study Of User Queries On The Web, *IRFORUM: SIGIR Forum (ACM Special Interest Group on Information Retrieval)* 32.

JEANNERET, Y. et SOUCHIER, E. (1999), Pour une poétique de l'écrit d'écran, *Xoana*(6/7), pp. 97-107.

JENKINS, C., CORRITORE, C. et WIEDENBECK, S. (2003), Patterns of Information Seeking on the Web: A Qualitative Study of Domain Expertise and Web Expertise, *IT&Society* 1(3), pp. 64-89, http://www.stanford.edu/group/siqss/itandsociety/v01i03/v01i03a05.pdf.

JONES, S., CUNNINGHAM, S. J. et MCNAB, R. (1998), An Analysis of Usage of a Digital Library, Actes de *European Conference on Digital Libraries*.

JOUËT, J. (2000), Retour critique sur la sociologie des usages, *Réseaux* 18(100), pp. 487-521.

JOUËT, J. (2003), Technologies de communication et genre. Des relations en construction, *Réseaux* 21(120), pp. 53-86.

KARLGREN, J., BRETAN, I., DEWE, J., HALLBERG, A. et WOLKERT, N. (1998), Genres defined for a purpose, fast clustering, and an iterative information retrieval interface, Actes de *Eighth DELOS Workshop on User Interfaces in Digital Libraries*.

KARLGREN, J. et CUTTING, D. (1994), Recognizing text genres with simple metrics using discriminant analysis, Actes de *COLING 94*, Kyoto, Japon.

KILGARRIFF, A. et GREFENSTETTE (2003), Introduction to the Special Issue on the Web as Corpus, *Computational Linguistics* 29(3), pp. 333-347.

KOSALA, R. et BLOCKEEL, H. (2000), Web Mining Research : A Survey, Actes de *SIGKDD Explorations*.

LELONG, B. (2003), Quel "fossé numérique" ? Clivages sociaux et appropriation des nouvelles technologies, *in* E. Maigret, *Communication et médias*, La Documentation Française**:** 112-116.

LEROI-GOURHAN, A. (1943), *Evolution et techniques. L' Homme et la matière.*, collection Paris, A. Michel Lagny-sur-Marne, Seine-et-Marne, impr. de E. Grevin et fils.

LEROI-GOURHAN, A. (1964), *Le Geste et la parole*, collection Paris, A. Michel Lagny, impr. E. Grevin et fils.

LEVENE, M. et LOIZOU, G. (1999), A probabilistic approach to navigation in hypertext, *Information Sciences* 114, pp. 165-186, http://www.dcs.bbk.ac.uk/~mark/download/probht.ps.gz.

LICOPPE, C., PHARABOD, A.-S. et ASSADI, H. (2002), Contribution à une sociologie des échanges marchands sur Internet, *Réseaux* 20(116), pp. 97-140.

MASAND, B. et SPILIOPOULOU, M. (2000), Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), Actes de *Proceedings of SIGKDD Explorations*.

MAURER, S., HUBERMAN, B. et ADAR, E. (2000), Web rings, Xerox Palo Alto Research Center (PARC), http://www.hpl.hp.com/shl/papers/rings/rings.pdf.

MCKENZIE, B. et COCKBURN, A. (2001), An empirical analysis of web page revisitation, Actes de *34th Hawaiian International Conference on System Sciences (HICSS'34)*, Maui, Hawaii, http://www.cosc.canterbury.ac.nz/~andy/papers/hiccsWeb.pdf.

MOBASHER, B., DAI, H., LUO, T., NAKAGAWA, M., SUN, Y. et WITSHIRE, J. (2000), Discovery of aggregate usage profiles for web personalization, Actes de *WebKDD'2000*, http://maya.cs.depaul.edu/~mobasher/papers/webkdd2000.pdf.

MODJESKA, D. (1997), Navigation in ElectronicWorlds: A Research Review, Toronto, Computer Systems Research Group, University of Toronto, http://www.dgp.utoronto.ca/people/modjeska/Pubs/lit_rvw.pdf.

MULLIER, D. (2000), Examining How Users Interact with Hypermedia Using A Neural Network, Actes de *Proceedings of ICAI-00*, Las Vegas, USA, http://www.lmu.ac.uk/ies/comp/staff/dmullier/icai.pdf.

MULLIER, D., HOBBS, D. et MOORE, D. (2002), Identifying and Using Hypermedia Browsing Patterns, *Journal of Educational Multimedia and Hypermedia* 11(1), http://www.lmu.ac.uk/ies/comp/research/isle/eduMedia/papers/J01Experiments.pdf.

MURRAY, D. et DURRELL, K. (1999), Inferring Demographic Attributes of Anonymous Internet Users, Actes de *WEBKDD'99*, http://www.i2pi.com/papers/analysis/inferring-demographic-attributes-of.pdf.

OWEZARSKI, P. (2001), Que nous dit la métrologie sur le futur d'Internet ?, Actes de *Journées Réseaux (JRES 2001)*, Lyon, France, http://www.laas.fr/~owe/PUBLIS/01428.pdf.

PADMANABHAN, B., ZHENG, Z. et KIMBROUGH, S. (2001), Personalization from Incomplete Data: What You Don't Know Can Hurt, Actes de *Seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2001)*, ACM, http://opim.wharton.upenn.edu/~balaji/bpkdd01.pdf.

RASTIER, F. (1987), *Sémantique interprétative*, collection.

RASTIER, F. (1989), *Sens et textualité*, collection, PUF.

RASTIER, F. (2001a), *Arts et sciences du texte*, collection, PUF.

RASTIER, F. (2001b), L'action et le sens - Pour une sémiotique des cultures, *Journal des anthropologues* 85-86, pp. 183-219.

RASTIER, F. (2003), Deniers et Veau d'or : des fétiches à l'idole, *Dits et inédits*, Revue Texto, http://www.revue-texto.net/Inedits/Inedits.html.

REHM, G. (2002), Towards automatic Web genre identification, Actes de *35th Hawaii International Conference on System Science*, Hawaii, http://dlib2.computer.org/conferen/hicss/1435/pdf/14350101.pdf.

REINERT, M. (1993), Les "mondes lexicaux" et leur logique, *Langage et société*(66), pp. 5-39.

RELIEU, M. et OLSZEWSKA, B. (2004), La matérialisation de l'internet dans l'espace domestique. Une approche située de la vie domestique, *Réseaux* 22(123), pp. 119-148.

RIGNAULT, M.-P. (2003), Thalassotherapy as a real choice: ethical, bio-ontological and cognitive aspects of seaweed, Actes de *Sixth InterGersoise Conference (IGC'03)*, Filartigue, France.

RODDICK, J. et SPILIOPOULOU, M. (2002), A Survey of Temporal Knowledge Discovery Paradigms and Methods, *IEEE Transactions on Knowledge and Data Engineering* 14(4), pp. 750-767, http://csdl.computer.org/dl/trans/tk/2002/04/k0750.pdf.

ROUSSINOV, D., CROWSTON, K., NILAN, M., KWASNIK, B., CAI, J. et LIU, X. (2001), Genre based Navigation on the Web, Actes de *34th Annual Hawaii International Conference on System Sciences (HICSS-34)*, http://csdl.computer.org/comp/proceedings/hicss/2001/0981/04/09814013.pdf.

SOUCHIER, E. (2000), Internet, Multimédia... mais de quelle écriture, de quelle lecture parlons-nous ?, Actes de *Ecritures et médiations - médias-cité 2000*, Bordeaux, France, http://www.medias-cite.org/docs/ecritures_numeriques.PDF.

SPILIOPOULOU, M., FAULSTICH, L. C. et WINKLER, K. (1999), A Data Miner Analyzing the Navigational Behaviour of Web Users, Actes de *Workshop on Machine Learning in User Modelling of the ACAI'99*, Creta, Greece, http://www.wiwi.hu-berlin.de/~myra/W_ACAI99.ps.gz.

SRIVASTAVA, J., DESIKAN, P. et KUMAR, V. (2003), Web Mining – Accomplishments & Future Directions, Actes de *Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD03)*, Seoul, Korea, http://www.csee.umbc.edu/%7Ekolari1/Mining/papers/srivastava.pdf.

TAUSCHER, L. et GREENBERG, S. (1997a), How people revisit web pages : empirical findings and implications for the design of history systems, *International Journal of Human Computer Studies* 47(1), pp. 97-138, http://ijhcs.open.ac.uk/tauscher/tauscher.pdf.

TAUSCHER, L. et GREENBERG, S. (1997b), Revisitation Patterns in World Wide Web Navigation, *in* ACM, *Conference on Human Factors in Computing Systems (CHI'97)*, Atlanta, Georgia, USA, ACM Press.

VAN DER WALT, M. (1998), The structure of classification schemes used in internet search engines, Actes de *Fifth Internationale ISKO Conference*, Lille, France.

XU, G., COCKBURN, A. et MCKENZIE, B. (2001), Lost on the Web: An Introduction to Web Navigation Research, Actes de *The Fourth New Zealand Computer Science Research Students Conference*.

III

Annexes

Annexe 1

Projets

Notre travail de thèse s'inscrit dans trois projets auxquels nous avons participé, menés au laboratoire « Usages, Créativité, Ergonomie » de la Direction des Interactions Humaines de France Télécom R&D : TypWeb, SensNet et BibUsages, dont nous donnons ici une description globale. Par le biais de TypWeb (2000-2001) et SensNet (2002-2003), nous avons accès aux données de trafic de plusieurs milliers d'internautes résidentiels issus du panel de la société de mesure d'audience NetValue[1], ainsi qu'à la plateforme et aux outils de traitement des données de trafic développée dans ce cadre par France Télécom R&D ; le projet BibUsages a quant à lui permis de constituer, en partenariat avec la Bibliothèque Nationale de France, un panel ciblé d'usagers des bibliothèques électroniques en 2002.

1.1 Projet TypWeb[2]

Un partenariat entre France Télécom R&D, NetValue, HEC et Wanadoo SA a été constitué en 2000 avec pour objectif d'exploiter de manière approfondie les données de trafic du panel France de NetValue sur l'année 2000 : il s'agit de toutes les données de trafic sur Internet d'un échantillon d'internautes à domicile, représentatif de la population connectée à Internet et résidant en France. Cette exploration approfondie a pour finalité de donner une meilleure connaissance des usages d'Internet et de comprendre comment évoluent les usages pour une cohorte donnée, ce qui permet d'anticiper ce que pourraient devenir les usages d'Internet une fois le marché arrivé à maturité.

1.1.1 Historique et objectifs

Les premiers contacts ont été établis entre la division multimédia de la branche Grand Public de France Télécom et NetValue fin 1999. NetValue était alors le seul

[1] En 2002, NetValue a été rachetée et intégrée au sein de Nielsen-NetRatings.

[2] Cette partie reprend la présentation du projet faite dans [Beaudouin *et al.* 2002].

opérateur dans le domaine de la mesure d'audience sur Internet. Le partenariat a finalement été signé en juin 2000 et porte sur la période avril 2000-décembre 2001 (un prolongement au contrat initial a été établi en juin 2001).

NetValue possède des données d'une finesse remarquable sur les pratiques de ses panélistes sur Internet, qui permettent en particulier d'étudier l'usage des différents protocoles Internet et leur évolution dans le temps. Grâce à une sonde NetMeter, installée sur l'ordinateur de l'internaute, toute l'activité de l'usager sur Internet est enregistrée : pages visitées, messages reçus et envoyés, etc. France Télécom R&D développe des méthodes avancées d'analyse des usages d'Internet, mobilisant ingénierie linguistique et fouille de données. Des méthodes de traitement ont été mises au point spécialement sur les données de NetValue. HEC dispose d'une expertise dans le domaine de la publicité et du marketing appliqué à Internet et Wanadoo, offreur d'accès et de services, a des besoins en matière de connaissance des usagers et des services utilisés.

Les données du panel NetValue ont les avantages suivants :

1. il s'agit d'un échantillon représentatif des internautes connectés à domicile. Tous les mois, NetValue fait réaliser par la Sofres une enquête téléphonique de cadrage auprès d'un échantillon de 4000 personnes tirées au hasard, ce qui permet de définir les caractéristiques de la population connectée à Internet. Mois par mois, le panel NetValue est ainsi réajusté (nouveaux recrutements) de manière à être représentatif de cette population connectée. Comme le marché est en pleine croissance, la cohorte que nous avons définie début 2000 perd de sa représentativité au fil de l'année : nous avons pris comme option de définir une population fermée, dont nous suivons de manière longitudinale les évolutions ;
2. grâce aux enquêtes de recrutement des panélistes, des informations fines sur les utilisateurs sont recueillies (sexe, âge, PCS, équipement, etc.) ;
3. les internautes sont suivis sur une longue période (engagement minimal d'un an) ce qui permet de suivre de manière longitudinale l'évolution des usages ;
4. les usages sur l'ensemble des protocoles Internet (et pas seulement sur le Web) sont recueillis : ainsi pouvons-nous évaluer l'usage des applications du type messagerie instantanée, *peer-to-peer*, etc.

Les axes de recherche

Le projet TypWeb a cherché à montrer comment les usages d'Internet évoluent au fil du temps pour une population donnée. Une cohorte de 1140 internautes extraite du panel grand public de NetValue a été suivie tout au long de l'année 2000, ce qui permet de se défaire du biais que pose, en raison du processus d'apprentissage, l'accroissement constant de la population des internautes. Les informations détaillées disponibles sur les internautes permettent de différencier les profils d'usage selon des variables socio-démographiques. Enfin, la méthodologie NetMeter de NetValue permet de récupérer les données de trafic Internet, quelque soit le protocole utilisé (*chat*, messagerie instantanée, FTP, audio, vidéo…).

Les différents axes de recherche suivis dans le cadre du projet TypWeb sont les suivants :

1. Vision globale d'usage d'Internet en différenciant les protocoles
2. Utilisation des services sur les principaux portails

3. Les usages des moteurs de recherche
4. La fréquentation des sites marchands (Tourisme et Biens culturels)
5. Les pages personnelles
6. Le courrier électronique et les réseaux de sociabilité
7. Segmentation des internautes et services de communication
8. La publicité sur Internet

Les travaux prévus dans le cadre du partenariat ont été menés jusqu'à leur terme, y compris la construction d'une segmentation des internautes sur la base des pratiques réelles d'Internet.

Ces travaux présentent quelques limites qu'il nous faut souligner :

- ils nous donnent une représentation des usages d'Internet à domicile exclusivement. Les usages au travail ou à l'université n'apparaissent pas.
- Si ces travaux permettent d'avoir une bonne *description* des usages, ils ne nous donnent pas accès à une *compréhension* des usages. Ils servent de base à des explorations qualitatives.

Rôle des partenaires

NetValue a mis à disposition du partenariat les données issues de son panel d'internautes résidentiels français : une cohorte de 1140 internautes a été extraite du panel et les données de trafic concernent toute l'année 2000. NetValue a participé au traitement par le biais d'un statisticien.

France Télécom R&D coordonne le partenariat et exploite les données recueillies pour dégager une analyse fine des usages d'Internet et des typologies d'internautes.

Wanadoo SA (Direction de la Stratégie et Wanadoo Régie) oriente les recherches, notamment sur l'axe publicité ; HEC a participé au démarrage à la coordination du partenariat et a pris en charge l'analyse du thème publicité.

Les résultats de travaux menés dans le cadre d'un autre partenariat entre France Télécom R&D, Paris III, Paris X et le LIMSI-CNRS (Benoît Habert, Serge Fleury et Marie Pasquier) sur la structure et le contenu des sites personnels ont été en partie réexploités dans le cadre de ce partenariat.

1.1.2 Principaux résultats

Sont présentés ici des résultats globaux sur les usages et des résultats détaillés sur quelques pratiques spécifiques : l'utilisation des moteurs de recherche, l'accès aux services sur les principaux portails, la fréquentation des sites marchands, les page personnelles selon l'hébergeur et le mail.

Les points qui marquent l'originalité de ces travaux sont :

- étude de l'évolution des usages pour une population définie d'internautes ;
- prise en compte de l'ensemble des protocoles, ce qui permet de comparer les usages du mail classique avec ceux du Web Mail, de suivre les usages du *chat*, de l'IRC…
- analyse des usages en analysant les contenus visités : identification des services sur les portails, identification des requêtes moteur et extraction des mots clefs, identification des rubriques sur les sites marchands et analyse des contenus textuels sur les pages personnelles.

Le point fort de l'approche mise en place dans TypWeb réside dans la capacité à articuler l'analyse de la production (quels sont les contenus des sites visités : types de services, mots-clefs des requêtes éventuellement contenu des pages) et celle de la réception (comment sont visités ces contenus). Pour cela nous croisons des méthodes d'ingénierie linguistique et de statistiques (*data* et *text-mining*). En parallèle des entretiens approfondis nous permettent de comprendre la logique de ces usages.

Évolution des usages, segmentation des internautes

Globalement le nombre de sessions Internet augmente au fil de l'année, mais le groupe des très faibles utilisateurs (un quart du panel) qui fait moins de 2 % de l'ensemble des sessions voit ses usages décliner au fil des mois. La dispersion des usages augmente donc avec le temps.

Dans 76 % des sessions, l'internaute accède au Web et dans 46 % des sessions à la messagerie électronique. Si les usages du Web et de la messagerie électronique sont répandus chez la grande majorité des internautes, ce n'est pas le cas pour les Messageries Instantanées et le *chat* qui ne sont utilisés que par un quart des internautes. Ces outils de communication synchrones sont principalement utilisés par les jeunes. Le recrutement social de ces utilisateurs est plus faible que celui de la messagerie électronique, ce qui incite à penser que le caractère synchrone et sans mémoire de l'échange lève certaines des barrières que pose l'utilisation de l'écrit. Enfin, le fait d'être célibataire influe positivement sur l'intensité d'usage de ces types d'outils.

Deux grands groupes d'internautes se distinguent : ceux qui accordent une place prépondérante au Web dans leurs usages d'Internet et ceux qui favorisent au contraire l'usage des services de communication. Dans chacun de ces groupes se constituent des axes de différenciation, en fonction de l'intensité d'usage pour le premier groupe et en fonction du ou des outils de communication utilisés (mail classique, WebMail, *chat*, messagerie instantanée…) pour le second groupe. Les utilisateurs de *chat* et de messageries instantanées se recrutent surtout chez les jeunes. Ils se distinguent par leur capacité à articuler au cours d'une même session consultation du Web, utilisation de la messagerie et conversations synchrones.

Les moteurs de recherche

Dans les parcours sur le Web, on a identifié toutes les pages vues qui correspondent à une requête auprès d'un moteur de recherche, puis extrait les mots-clefs et les opérateurs dans les requêtes ce qui permet à la fois d'explorer les usages des différents moteurs mais aussi le contenu des recherches. On a pu ainsi constater que 29 moteurs de recherche différents ont été utilisés par les panélistes en 2000.

Les requêtes sur les moteurs ne représentent que 100 000 pages vues sur les 7,5 millions de pages vues au total en 2000, soit 1,3 % ; cependant, ils sont très présents dans la navigation : 20 % des sessions de navigation sur le Web comprennent une requête dans un moteur de recherche.

Les moteurs de recherche sont utilisés par une large majorité (85 %) du panel étudié. Parmi les 15 % des panélistes n'ont jamais utilisé de moteur de recherche, on

trouve surtout des femmes et des jeunes de moins de 15 ans : il y a une corrélation positive forte entre l'intensité d'usage d'Internet et l'intensité d'usage des moteurs.

Il semble qu'un usage intense des moteurs passe en 2000 par une diversification des moteurs utilisés. Les faibles utilisateurs n'utilisent qu'un seul moteur, tandis que les forts utilisateurs explorent et testent en permanence l'offre en termes de moteurs. Même au sein des sessions, on a remarqué que dans 32% des cas, plusieurs moteurs étaient utilisés.

L'analyse des requêtes adressées à chaque moteur permet de mettre en évidence la proximité des moteurs des grands portails et des spécialisations assez fortes pour certains moteurs (les requêtes « sexe, piratage et vidéo » sont plus fréquentes sur un moteur comme Altavista, tandis que d'autres attirent des requêtes « vie pratique »). Les moteurs qui possèdent les identités les plus marquées et les plus opposées sont *Altavista* et *Wanadoo*. Dans ce cadre, les internautes se divisent en deux grandes catégories d'utilisateurs des moteurs :

- ceux dont les requêtes tournent autour de la « culture internet » (multimédia, sexe, jeux, informatique), qui sont plus fréquemment des hommes, des moins de 24 ans et des très gros utilisateurs du Web,
- ceux qui recherchent des informations pour la vie « ordinaire » : ce sont principalement des femmes, des personnes d'âge moyen et appartenant à des catégories professionnelles intermédiaires.

Les portails généralistes

On a identifié sur les principaux portails généralistes (Voila, Wanadoo, Yahoo France, Yahoo US, Altavista, Free, etc.) les différents services proposés, et évalué quels en étaient les usages. Premier résultat, on constate qu'il y a plusieurs dizaines de services offerts, mais que les quatre à six premiers services recueillent à eux seuls 80 % de l'audience.

Le nombre de pages vues par les internautes sur les principaux portails est stable au cours de l'année 2000. Mais cette stabilité globale cache des évolutions contrastées :

- l'usage des moteurs diminue au fil de l'année. Pour expliquer ce phénomène, on suppose qu'avec l'ancienneté, les internautes se repèrent plus facilement sur le Web et se servent moins des moteurs, au profit d'autres outils de repérage sur le Web : signets, de sites avec liens, etc.
- l'usage des services de communication (mail, *chat*…) augmente. La diversification des portails en termes de services proposés (à l'origine le portail était essentiellement centré autour du moteur de recherche ou de l'annuaire pour Yahoo) a donc un effet visible sur les usages.

Les pages personnelles

Les pages personnelles visitées ont globalement des tonalités différentes selon leur serveur d'hébergement : le domaine donne un style à ses habitants. L'analyse des contenus d'un échantillon de pages personnelles *visitées* (sites hébergés chez des fournisseurs d'accès ou des portails) permet d'identifier des styles propres aux hébergeurs, comme on peut le voir pour Wanadoo et Free :

- Wanadoo : les pages visitées se caractérisent par une forte présence des verbes *dire*, *parler*, *penser* ; la mise en scène de l'échange (*moi*, *nous* / *toi*, *vous*) ; thèmes du gravage de CD ; les thématiques du travail (*bureau*, *directeur*, *patron*, *licenciement*…), de l'amour (*rencontrer*, *regard*, *plaire*), de la vie (*vieillir*, *mourir*…) et d'autres préoccupations d'ordre existentiel. Le site est alors un lieu d'expression intime du moi qui s'adresse à l'autre. Les pages visitées hébergées par Club-internet présentent des caractéristiques proches de celles de Wanadoo.
- Free : quelques domaines sémantiques peuvent clairement être identifiés : les messages renvoyés par les serveurs d'interdiction d'accès ou de redirection (*you don't have permission*, *forbidden*, *click here*), le champ sémantique du sexe (y compris les mises en garde pour les visiteurs), celui des logiciels (*cracks*, *download*…) et celui de la gratuité. Chez Free, on observe un entrelacement intéressant entre la liberté (sexuelle et logicielle) et la gratuité, porté par le double sens du mot *free*.

Les sites marchands

Pour étudier la fréquentation de sites marchands, on a identifié les sites marchands des secteurs du tourisme, des courses et des biens culturels (disque, cd…) et au sein de ces sites les différents services visités (information, achat, réservation, recherche, promotions…). Cela permet de repérer les rubriques des sites qui sont effectivement visitées mais aussi d'analyser comment l'internaute explore et compare les offres des sites au cours de ses parcours sur le Web. C'est la première fois que cette exploration des parcours sur les sites marchands est réalisée.

La moitié des internautes est allée au moins une fois en 2000 sur un site marchand lié au tourisme, et il en va de même pour les sites de biens culturels (Fnac, Alapage, Amazon...). Les internautes qui fréquentent les sites marchands sont plutôt des hommes, d'anciens internautes, avec une intensité d'usage d'Internet élevée. Pourtant dans seulement 2 % des sessions est identifié un accès à un site marchand de tourisme ou de biens culturels.

Près de la moitié des internautes qui consultent une agence de voyage virtuelle, consultent d'autres sites du même type au cours de la même session. En revanche, les internautes sont plus fidèles sur les sites de biens culturels : 80 % ne visitent qu'un site de ce type au cours de la session.

Le profil d'usage des sites marchands suit de près la structure de l'offre des sites. Par exemple, Promovacances valorise ses offres de dernière minute, et c'est bien cette partie du site qui est la plus visitée, à l'inverse de Travelprice dont les fonctions de recherche sont les plus valorisées et les plus visitées. Le profil d'usage reflète le positionnement des sites.

Conclusions

Que l'on étudie les usages des moteurs, des portails, le contenu des pages personnelles des différents hébergeurs ou la fréquentation des sites marchands, on est étonné d'observer un ajustement aussi serré entre l'offre et la demande :

- les portails diversifient en 2000 leur offre de service en mettant l'accent sur les services de communication, et ce sont ces derniers qui voient leurs usages croître ;
- les hébergeurs de pages personnelles ont chacun des positionnements spécifiques (en termes de communication, de cible marketing…) et rencontrent des utilisateurs qui renforcent leur positionnement. Les pages perso chez Free valorisent liberté et gratuité comme leur hébergeur, les pages chez Wanadoo le profil français moyen.... ;
- les moteurs, même les plus généralistes sont utilisés de manière différente par les internautes, ce qui semble montrer qu'ils ont des identités marquées : Altavista est plutôt utilisé pour les requêtes sexe, multimédia et piratage, Voila davantage pour la vie pratique ;
- sur les sites marchands ce sont les parties les plus mises en valeur par l'ergonomie et le discours de communication qui rencontrent le plus de visiteurs.

Il y a donc bel et bien un ajustement réciproque entre l'offre et la demande, et c'est une spécificité d'Internet. Sans doute le fait que les utilisateurs soient intégrés dans les processus de conception des services et des outils favorise-t-il cet ajustement entre la production et la réception.

1.2 Projet SensNet[1]

Le projet SensNet (2002-2004) se situe dans le prolongement du projet TypWeb ; il bénéficie d'un financement du Réseau National de Recherche en Télécommunications (RNRT) du Ministère de la Recherche, et compte quatre partenaires : France Télécom R&D, NetValue (devenue Nielsen/NetRatings), le LIMSI – CNRS et l'Université de Paris III.

L'objectif final de ce projet est de mettre en place un système de catégorisation sémantique des usages et des parcours du Web. En s'appuyant sur les données d'usages des internautes du panel NetValue, il a pour objectif de proposer un système de catégorisation qui prend en compte les particularités du Web :

1. Celui-ci n'est pas seulement un espace de consultation d'information ; il autorise un nombre élevé de types d'activités (s'informer, rechercher, communiquer, acheter…) ;
2. Le Web est un hypermedia, cela implique que les aspects formels (réseau de liens, éléments multimedia, zones interactives...) soient intégrés dans la catégorisation ;
3. La page vue est un moment dans le parcours de l'internaute mais aussi un des éléments constitutifs d'un site. Il faut prendre en compte la conception des sites dans l'analyse des usages du Web. Cette démarche d'analyse appliquée à des usages spécifiques (utilisation des portails, des sites marchands, parcours de recherche d'information...) permettra de mieux

[1] Dans cette partie, nous reprenons la description du projet soumise au RNRT.

catégoriser les sites, les parcours et de définir des profils d'internautes en fonction de leurs usages.

1.2.1 Objectifs

L'objectif global du projet est de mettre en œuvre un système d'analyse sémantique du Web constitué à partir des usages effectifs du Web, qui tienne compte des types d'activité et des aspects hypermédia ; qui situe les pages vues dans leur site d'origine et dans les parcours, comme étant à la croisée entre un site et un parcours et qui s'appuie sur le récit des pratiques des utilisateurs et concepteurs pour donner du sens.

Le premier objectif est de constituer un prototype de plate-forme de catégorisation automatique qui permette

1. de catégoriser les types d'activité (communiquer, consulter, acheter…), ce qui implique d'établir un inventaire de ces types d'activité ;
2. de capturer les traits formels (par exemple la présence de liens externes ou d'images sur la page) et textuels (par exemple les pronoms personnels ou les noms rares…) prédéfinis, correspondants à des pages vues et à des parcours ;
3. d'affecter des catégories thématiques aux pages consultées.

Le deuxième objectif consiste à :

1. identifier les traits formels et textuels pertinents pour caractériser les objets du Web qui seront capturés dans la plateforme qui vient d'être décrite ;
2. mettre au point des méthodes de traitement adaptées à chaque type de traits. Il pourra y avoir des stratégies de catégorisation complémentaires. On pourra par exemple considérer que les contenus de la balise HTML META, que remplit le concepteur de site et qui sont largement utilisés pour l'indexation, constituent un jeu de traits pertinents pour catégoriser thématiquement les sites. Et ces traits pourront être soumis à différents types de traitement (catégorisation inductive, supervisée…).

Dans ce contexte, la mise en place d'un système informatique requiert la confrontation permanente avec les données. C'est pourquoi la mise au point de l'outil se fera par l'exploration systématique des données d'usage et de parcours.

Le troisième objectif de SensNet est d'explorer de manière approfondie plusieurs usages d'Internet. Comme il est hors de propos de catégoriser tout le Web, il a été choisi de sélectionner des types de sites (portails, sites marchands et serveurs communautaires, sites consacrés à la musique) et des types de pratiques (recherche d'information, achat en ligne, consultation d'archives en ligne) sur lesquels nous projetterons les parcours. L'exploration de ces usages et le croisement avec des entretiens qualitatifs permettront de définir précisément les traits les plus pertinents pour catégoriser les sites et les parcours. Un autre aspect important et original du projet est de relier ces parcours catégorisés au profil socio-démographique des internautes. En effet, l'utilisation d'un panel représentatif des internautes permet d'obtenir des données précises de comportement d'individus dont le profil est

connu. Les profils permettent d'enrichir la catégorisation des sites et des parcours, de même que la catégorisation thématique va enrichir le profil des internautes.

Enfin, le dernier objectif correspond à la démarche de validation des outils mis en place et des méthodes d'analyse qui s'étendra tout au long du projet et fera l'objet d'un sous-projet particulier. Il est essentiel dans ce projet d'identifier précisément les avancées et les limites de la catégorisation sémantique automatique telle que nous la proposons, afin de l'améliorer *via* une confrontation permanente avec le terrain (professionnels de l'Internet, internautes). Un bilan sera réalisé en fin de projet.

1.2.2 Mise en œuvre et état de l'art

Ce projet met en oeuvre une approche pluridisciplinaire et s'appuie sur des méthodes et outils issus de différents domaines :

- Linguistique informatique, et notamment la linguistique de corpus.
- Statistiques et analyse de données.
- Techniques de recueil de trafic Internet.
- Méthodes de la sociologie des usages.

Il y a deux types de verrous à lever :

- Verrou technologique : insuffisance de l'information contenue dans les URL
  L'analyse des adresses (URL) seules ne permet pas d'obtenir une information suffisamment fine. En effet, à titre d'exemple, les contenus générés dynamiquement ne donnent aucune information sur les thématiques dans les URL, mais des informations techniques (n° de fichier par exemple). Il est donc indispensable d'analyser le contenu des pages.
- Verrou économique : coût de la catégorisation manuelle
  Une classification manuelle des sites les plus importants est déjà réalisée par les équipes de NetValue, en fonction d'une typologie propre aux sites. La complexité des sites de type « portail », qui proposent l'ensemble des services accessibles sur le Web (information, messagerie, sports, finance, etc.) rend très difficile la classification de leurs contenus. Par ailleurs, il est impossible en l'état de catégoriser l'ensemble des pages vues par le panel tous les mois (plusieurs millions de pages par pays et par mois). Cette difficulté est accentuée par le changement rapide du contenu des pages et de la structure des sites.

1.2.3 Organisation du projet

Le projet est décomposé en cinq sous-projets.

- Sous-projet 1 : Prototype de plate-forme de catégorisation automatique (pilotage : NetValue). Il s'agit de développer un système qui permette 1) de capturer les traits formels et textuels définis en amont, correspondants à des pages vues et à des parcours ; 2) de catégoriser les

types d'activité (ce qui implique d'établir un inventaire de ces types d'activité).

- Sous-projet 2 : Définition des traits et méthodes de traitement associées (pilotage : LIMSI). Il vise 1) à identifier les traits formels et textuels pertinents pour caractériser les objets du Web et 2) à mettre au point des méthodes de traitement adaptées à chaque type de traits. Il pourra y avoir des stratégies de catégorisation complémentaires (catégorisation inductive, supervisée…).
- Sous-projet 3 : Sites, parcours et utilisateurs (pilotage : France Télécom R&D). Le sous-projet consiste à explorer de manière approfondie plusieurs usages d'Internet, en sélectionnant des types de sites et des types de pratiques.
- Sous-projet 4 : Validation des outils et des méthodes d'analyse (pilotage : NetValue). Ce sous-projet consiste à confronter la catégorisation induite à 1) celle des professionnels 2) celle perçue par les internautes et à mettre en évidence le caractère discriminant ou non des traits.
- Sous-projet 5 : Pilotage et coordination (pilotage : France Télécom R&D). Ce sous-projet est dédié 1) à la mise en place des moyens (serveurs, outils de travail coopératifs) nécessaires pour partager les données, les outils et les avancées des différents sous-projets et 2) au suivi du bon déroulement du projet. Il sera pris en charge par le comité de pilotage regroupant des représentants de chaque partenaires.

1.2.4 Retombées du projet

Des résultats scientifiques sont attendus dans le domaine des usages, allant dans le sens d'une meilleure connaissance des profils des utilisateurs d'Internet et de la manière dont ils perçoivent les services et contenus qui leur sont proposés. Les méthodes et outils d'analyse sémantique qui sont proposés présentent une démarche scientifique originale qui s'intègre dans le cadre de la linguistique de corpus. La communauté scientifique « Web sémantique » sera également très réceptive aux résultats de SensNet. En effet, il est envisagé de faire des propositions dans le cadre de l'action *semantic web* du W3C à partir des résultats obtenus dans SensNet.

En termes de retombées industrielles et économiques, ce projet devrait aboutir au développement ou au prototypage d'outils pour :

- le classement des sites Web (ou rubriques) qui traitent principalement d'un thème donné ;
- le classement des thèmes les plus consultés pour un site donné (ou un ensemble de sites) ;
- la mise en relation des profils socio-démographiques des internautes avec leurs thèmes de prédilection (outil marketing) ;
- l'aide à la navigation dans les sites complexes ;
- l'aide à la construction d'annuaires thématiques du Web.

1.3 Projet BibUsages[1]

Le projet BibUsages est un partenariat entre France Télécom R&D et la Bibliothèque Nationale de France mené en 2002 avec le soutien du RNRT ; il a pour objectif l'analyse des usages des bibliothèques électroniques en France.

1.3.1 Objectifs et méthodologie

Le projet BibUsages s'intéresse aux usages des bibliothèques électroniques en ligne. De tels usages sont innovants mais ils s'insèrent dans des pratiques stabilisées, en particulier au sein de la population des enseignants et des chercheurs, mais également auprès du grand public. L'accès immédiat à un corpus volumineux de d'œuvres permet à des chercheurs d'envisager des études inédites, car techniquement impossibles auparavant. Par ailleurs, les enseignants, du collège au premier cycle universitaire, trouvent dans les bibliothèques électroniques une ressource pédagogique inestimable.

L'objectif principal du projet est de décrire les usages des bibliothèques en ligne et en particulier ceux de Gallica, la bibliothèque électronique en ligne de la Bibliothèque Nationale de France (`http://gallica.bnf.fr`), en les croisant avec les caractéristiques de la population des utilisateurs. Cette étude permet également de mettre en évidence la manière dont des usages émergeants infléchissent et modifient des pratiques bien établies ; dans le cas présent, la recherche académique et l'enseignement, ainsi que les pratiques de lecture et de manipulation de textes en ligne en général (lecture à l'écran, téléchargement de textes et d'ouvrages, etc.).

Dans ce contexte, il s'agit d'expliquer, par des méthodes issues des sciences sociales et cognitives, des usages déjà largement diffusés (il existe déjà plusieurs bibliothèques électroniques librement consultables sur le Web) mais dont une compréhension plus rigoureuse permettrait d'élargir la palette des fonctionnalités innovantes proposées tout en s'adaptant aux besoins et aux caractéristiques des utilisateurs.

Dans cette étude, nous avons mis en œuvre une méthodologie combinant des approches qualitatives et quantitatives et mettant en œuvre en particulier une technologie innovante de capture et d'analyse de trafic IP. Nous avons ainsi mis en œuvre une approche « centrée utilisateur » qui reste rarement mise en oeuvre dans les études d'usages d'envergure sur le Web.

État de l'art

Dans les études d'usage d'Internet, on distingue deux grandes catégories : les études dites « centrées serveur », qui s'appuient sur l'analyse des journaux de connexion (*access logs*) disponibles sur les serveurs d'une part et les études dites « centrées utilisateur » qui s'appuient sur l'enregistrement du trafic au niveau de l'ordinateur personnel de l'utilisateur d'autre part.

---

[1] Cette partie reprend les éléments de synthèse présentés dans [Assadi *et al.* 2003a].

La plupart des études menées à ce jour relèvent de la première catégorie. Les études « centrées utilisateur » sont à la fois plus rares et plus riches du point de vue de la compréhension des usages. En effet, le fait d'avoir accès à l'utilisateur et à ses caractéristiques permet de croiser ces données avec les données de trafic. En outre, il est plus aisé et plus productif de compléter des études centrées utilisateur par des études qualitatives (entretiens et observations auprès d'un échantillon d'utilisateurs).

En ce qui concerne le thème spécifique des usages des bibliothèques électroniques en ligne, ce domaine reste largement à explorer. La population des utilisateurs de bibliothèques est relativement bien connue, grâce aux enquêtes et études menées par les grandes bibliothèques (dont la BnF) auprès de leurs publics sur place. En revanche, il n'existe pas à notre connaissance d'étude globale des usages d'une population diversifiée d'utilisateurs distants d'une bibliothèque, population composée de chercheurs universitaires et d'étudiants de troisième cycle, mais également d'enseignants de collège et de lycée, d'élèves ou de particuliers menant des recherches à titre personnel.

Organisation du projet et méthodologie

Le projet BibUsages a été mené en partenariat entre France Télécom R&D et la Bibliothèque nationale de France et a bénéficié du soutien du Réseau National de Recherches en Télécommunications (RNRT). Le projet a duré 12 mois et s'est déroulé en 3 étapes :

1. Enquête en ligne sur le site de Gallica (mars 2002).
   Un questionnaire a été soumis aux visiteurs du site Gallica en mars 2002 durant trois semaines. Il permet à la fois d'avoir une connaissance plus précise du public de Gallica, et de recruter les volontaires pour faire partie du panel d'utilisateurs dont le trafic Web a été enregistré.
   Outre les caractéristiques socio-démographiques des répondants, le questionnaire s'articule autour de deux thématiques principales : d'une part, l'usage de Gallica (fréquence des visites, rubriques consultées, etc.), et d'autre part les usages d'Internet en général (intensité d'usage, services utilisés, types de sites visités, etc.). À la fin du questionnaire, les répondants se sont vu proposer de participer au panel d'utilisateurs mis en place.
   Au terme de cette première étape, 2340 personnes ont répondu au questionnaire, et 589 ont accepté de faire partie du panel d'utilisateurs, soit près d'un quart.
2. Constitution d'un panel d'utilisateur, installation du dispositif de capture de trafic chez les utilisateurs du panel et recueil des données.
   Au terme de la procédure d'inscription et d'installation, le panel est composé de 72 volontaires dont les caractéristiques socio-démographiques correspondent à celles de l'ensemble des répondants à l'enquête. Les données d'usage de ce panel ont été rapatriées sur un serveur centralisé de traitement de juillet à décembre 2002.
3. Conduite d'entretiens avec un échantillon d'utilisateurs volontaires faisant partie du panel (octobre 2002).
   Ces entretiens ont concerné 16 des 72 participants du panel, et ont été axés autour de trois problématiques particulières : leurs usages d'Internet

en général, leurs usages des bibliothèques numériques et de Gallica, et liens avec les pratiques de lecture et culturelles « off-line. ».

L'analyse croisée des trois sources de données – questionnaire en ligne, données de trafic, entretiens – permet ainsi de dresser un panorama des usages riche et dépassant les pratiques on-line proprement dites.

1.3.2 Retombées du projet

Quelques résultats du projet

En premier lieu, le projet BibUsages a permis de mieux appréhender les utilisateurs des bibliothèques électroniques. Celles-ci attirent un public qui n'est pas nécessairement habitué aux bibliothèques, mais qui y vient par le biais de recherches spécifiques : dans les entretiens autant que dans le trafic observé chez les participants de l'étude, les fonds numérisés apportent la possibilité de disposer de manière simple et rapide de documents de référence, difficilement trouvables, et qui s'inscrivent dans le cadre de contextes de recherche précis. Ce public semble assez différent de celui des bibliothèques classiques, et les chercheurs « professionnels » y sont comparativement peu représentés. Les plus de quarante ans, actifs ou retraités, sont majoritaires dans la population observée, et les bibliothèques électroniques sont avant tout pour eux une source d'informations dans le cadre de recherches personnelles. L'intensité d'usage est ici bien supérieure à celle de la population générale des internautes français, et va de pair avec un très fort taux d'équipement en haut débit (câble, ADSL)[1] ; nous avons ici affaire à une population d'utilisateurs avancés, qui pourrait être considérée comme une population leader dans les usages du haut-débit.

Le projet BibUsages a également permis d'appréhender les contextes d'usage des fonds numérisés. Il apparaît que si d'une manière générale les utilisateurs des bibliothèques électroniques sont également de forts consommateurs de « contenus à lire » (journaux en ligne en particulier). Au sein des sessions de navigation, l'usage des bibliothèques numériques est fortement corrélé à celui des moteurs de recherche d'une part, et à celui des sites de vente de biens culturels d'autre part. Deux profils se dégagent : celui du « chercheur amateur », dont les centres d'intérêt sont pointus et déjà bien connus de l'utilisateur, et celui du bibliophile pour qui Gallica fait office de catalogue avant achat. Dans les deux cas, la lecture en ligne est rare, tout autant que l'impression des documents téléchargés et la lecture s'apparente à la recherche de fragments ciblés au sein de vastes collections laissant de côté la totalité des œuvres. Dans ce cadre, le statut des documents en ligne semble remis en cause : tandis que l'édition papier reste du côté de l'œuvre, l'édition électronique s'apparente à l'usuel.

Dès lors, attirant de nouveaux publics, induisant de nouveaux modes d'appréhension des textes, s'inscrivant dans des parcours de lecture inédits, les

---

[1] 39% de nos utilisateurs déclarent être équipés d'une connexion haut-débit (enquête en ligne sur le site de Gallica, mars 2002, 2340 réponses), contre 8,9% de la population des internautes en France (source : NetValue, rapport de décembre 2001).

bibliothèques électroniques, loin d'être une simple version numérisée des fonds, s'apparentent à un nouvel espace de lecture et de consultation aux côtés des bibliothèques traditionnelles.

Retombées du projet et perspectives

La connaissance des publics des bibliothèques numériques et des contextes d'usage dans lesquels ils y accèdent fournit des retombées intéressantes pour les deux partenaires du projet.

Pour France Télécom, trois points particuliers sont à retenir, en premier lieu en termes de connaissance client : nous découvrons ici une population d'internautes seniors fortement équipés en haut débit, et dont les centres d'intérêt, outre l'offre de services et de communication classique, gravitent autour des contenus « culturels ». Cette population atypique dans le paysage des internautes français constitue en elle-même une cible intéressante pour France Télécom, pour laquelle il est maintenant possible d'adapter l'offre de services en l'orientant plus vers les outils de recherche et les contenus « à lire ».

Ensuite, l'étude montre les passerelles entre Web marchand et non marchand pour les utilisateurs. Alors que les acteurs d'Internet (fournisseurs de contenus et d'accès) perçoivent une dichotomie forte entre sites marchands et non marchands, les internautes passent indifféremment d'un type de site à un autre et l'on doit plutôt parler d'enrichissement mutuel entre sites marchands et non marchands dès lors qu'on les envisage sous l'angle des pratiques.

Enfin, en termes d'expertise technique, l'expérimentation a permis d'asseoir et de compléter les outils et les méthodes utilisées pour l'analyse de données de trafic centrées utilisateur. Cette expertise permet à FTR&D de proposer des études ciblées sur des pratiques prédéfinies et de fournir des analyses fines des usages.

Pour la Bibliothèque Nationale de France, le projet permet avant tout de mieux connaître son public numérique : pratique forte du téléchargement, recours quasi-systématique à l'outil de recherche, points d'améliorations ergonomiques sont autant d'enseignements du projet, tant par l'analyse du trafic que par les entretiens, qui permettront à la BnF d'adapter son offre.

Par ailleurs, BibUsages permet à la BnF de mieux connaître les contextes dans lesquels son fond numérique est visité. Le point de vue utilisateur adopté dans l'étude renseigne sur la fréquentation par les utilisateurs des autres sites proposant des collections de textes en ligne, « concurrents » directs de Gallica ; il montre également quels liens avec les sites marchands sont envisageables à partir de Gallica (bibliophilie, par exemple) et lesquels ne sont pas pertinents (sites de journaux en ligne en particulier).

Annexe 2

Requêtes Web : mille-feuille technique

Cette section a pour objectif de présenter de manière simple certains des éléments du dispositif technique qui sous-tend la navigation sur le Web, afin de mieux comprendre les parcours et les traces qu'ils peuvent laisser. L'affichage d'une page Web est le résultat d'une série d'opérations informatiques impliquant différents niveaux techniques de communication entre le navigateur et le site Web. Chacune de ces couches remplit une fonction particulière, qui peut être décrite simplement[1].

2.1 Acheminement et adressage

Le couple TCP/IP, utilisé pour la transmission de données sur Internet, assure le transport et l'adressage des données : le protocole IP permet de connaître la provenance et la destination des informations, et le protocole TCP assure l'intégrité des données transmises. À titre de comparaison, IP est l'équivalent d'une adresse postale, et TCP gère les lettres et colis échangés entre deux adresses.

2.1.1 Le rôle de TCP/IP

TCP (qui signifie Transmission Control Protocol, soit en français : Protocole de Contrôle de Transmission) est un des principaux protocoles de la couche transport du modèle TCP/IP. Il permet, au niveau des applications, de gérer les données en provenance (ou à destination) de la couche inférieure du modèle (c'est-à-dire le protocole IP). Lorsque les données sont fournies au protocole IP, celui-ci les encapsule dans des datagrammes IP, en fixant le champ protocole à 6 (pour savoir que le protocole en amont est TCP). TCP est un protocole orienté connexion, c'est-à-

[1] Dans cette annexe, nous reprenons en filigrane des extraits des articles présentés sur le site http://www.commentcamarche.net (© Jean-François Pillou), soumis à la licence GNU FDL (http://www.gnu.org/copyleft/fdl.html).

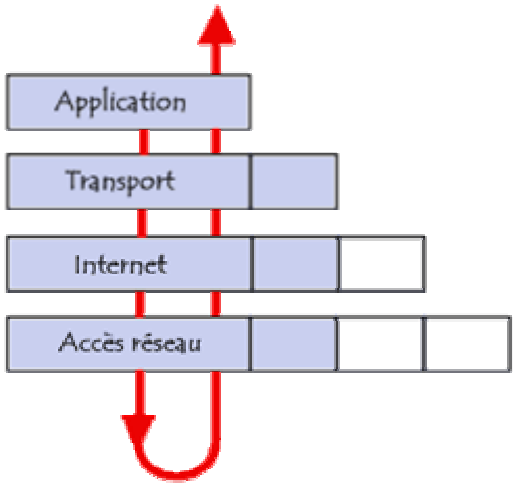
dire qu'il permet à deux machines qui communiquent de contrôler l'état de la transmission.

Lors d'une communication à travers le protocole TCP, les deux machines doivent établir une connexion. La machine émettrice (celle qui demande la connexion) est appelée client, tandis que la machine réceptrice est appelée serveur. On dit qu'on est alors dans un environnement Client-Serveur.

Afin de pouvoir appliquer le modèle TCP/IP à n'importe quelles machines, c'est-à-dire indépendamment du système d'exploitation, le système de protocoles TCP/IP a été décomposé en plusieurs modules effectuant chacun une tâche précise. De plus, ces modules effectuent ces tâches les uns après les autres dans un ordre précis, on a donc un système stratifié, c'est la raison pour laquelle on parle de modèle en couches.

Le terme de couche est utilisé pour évoquer le fait que les données qui transitent sur le réseau traversent plusieurs niveaux de protocoles. Ainsi, les données (paquets d'informations) qui circulent sur le réseau sont traitées successivement par chaque couche, qui vient rajouter un élément d'information (appelé en-tête) puis sont transmises à la couche suivante. Le modèle TCP/IP est très proche du modèle OSI (modèle comportant 7 couches) dont il reprend l'approche modulaire, mais en contient uniquement quatre :

Couche	Fonction
Couche Application	englobe toutes les applications accédant au réseau (Telnet, SMTP, FTP, etc.)
Couche Transport (TCP)	assure l'acheminement des données, ainsi que les mécanismes permettant de connaître l'état de la transmission
Couche Internet (IP)	fournit le paquet de données (datagramme)
Couche Accès réseau	spécifie la forme sous laquelle les données doivent être acheminées quel que soit le type de réseau utilisé



Lors d'une transmission, les données traversent chacune des couches au niveau de la machine émettrice, de l'application (par exemple un navigateur) jusqu'à la couche réseau. À chaque couche, une information est ajoutée au paquet de données, il s'agit d'un en-tête, ensemble d'informations qui garantit la transmission. Au niveau de la

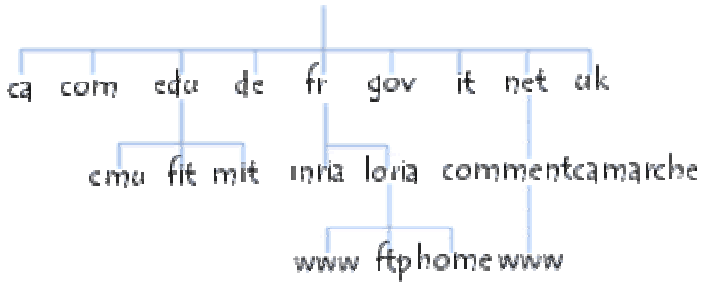
machine réceptrice, lors du passage dans chaque couche, l'en-tête est lu, puis supprimé.

2.1.2 Adresse IP et nom de domaine

Sur Internet, les ordinateurs communiquent entre eux grâce au protocole TCP/IP, que l'on écrit sous forme de 4 numéros allant de 0 à 255 (4 fois 8 bits), on les note donc sous la forme `xxx.xxx.xxx.xxx` où chaque `xxx` représente un entier de 0 à 255. Ces numéros servent aux ordinateurs du réseau pour se reconnaître : chaque machine sur le réseau possède une adresse IP propre. Cependant, les utilisateurs ne veulent pas travailler avec des adresses numériques du genre *194.153.205.26* mais avec des noms de stations ou des adresses plus explicites, par exemple `http://www.yahoo.fr/` ou `contact@wanadoo.fr`. TCP/IP permet d'associer des noms en caractères alphanumériques aux adresses numériques grâce à un système appelé DNS (*Domain Name System*).

On appelle *résolution de noms de domaines* (ou *résolution d'adresses*) la corrélation entre les adresses IP et le nom de domaine associé. Aux origines de TCP/IP, étant donné que les réseaux étaient très peu étendus, c'est-à-dire que le nombre d'ordinateurs connectés à un même réseau était faible, les administrateurs réseau créaient des fichiers appelés *tables de conversion manuelle* (fichiers généralement nommés *hosts* ou *hosts.txt*), associant sur une ligne l'adresse IP de la machine et le nom littéral associé, appelé *nom d'hôte*. Ce système avait l'inconvénient majeur de nécessiter la mise à jour des tables de tous les ordinateurs en cas d'ajout ou modification d'un nom de machine. Ainsi, avec l'explosion de la taille des réseaux, et de leur interconnexion, il a fallut mettre en place un système plus centralisé de gestion des noms. Ce système est nommé *Domain Name System* (*Système de nom de domaine*).

Ce système consiste en une hiérarchie de noms permettant de garantir l'unicité d'un nom dans une structure arborescente. On appelle nom de domaine, le nom à deux composantes, dont la première est un nom correspondant au nom de l'organisation ou de l'entreprise, le second à la classification de domaine de premier niveau, ou TLD (Top Level Domain : .fr, .com, etc.). Chaque machine d'un domaine est appelée hôte. Le nom d'hôte qui lui est attribué doit être unique dans le domaine considéré (le serveur Web d'un domaine porte généralement le nom *www*). L'ensemble constitué du nom d'hôte, d'un point, puis du nom de domaine est appelé *adresse FQDN* (*Fully Qualified Domain Name*, soit *Nom de Domaine Totalement Qualifié*). Cette adresse permet de repérer de façon unique une machine. Ainsi `www.yahoo.fr` représente une adresse FQDN. Les machines appelées *serveurs de nom de domaine* permettent d'établir la correspondance entre le nom de domaine et l'adresse IP sur les machines d'un réseau.



Chaque domaine possède un serveur de noms de domaines, relié à un serveur de nom de domaine de plus haut niveau. Ainsi, le système de nom est une architecture distribuée, c'est-à-dire qu'il n'existe pas d'organisme ayant à charge l'ensemble des noms de domaines. Par contre, il existe un organisme (l'InterNIC pour les noms de domaine en *.com*, *.net*, *.org* et *.edu* par exemple).

2.1.3 Domaines de premier niveau

Les domaines de premier niveau (TLD, Top Level Domain) sont de deux types : les domaines génériques correspondent (en principe) à une description thématique des contenus, tandis que les domaines par pays rattachent un site à l'État.

Il existe actuellement quatorze domaines génériques, ou gTLD (*generic Top Level Domain*), extensions Internet à caractère générique, formées de trois lettres et plus[1] :

gTLD	Signification	Enregistrement ouvert	ouverture
.aero	aéronautique	l'industrie du transport aérien	2002
.biz	business	tous	2001
.com	commercial	tous	1995
.coop	coopérative	Coopératives	2002
.edu	éducation	écoles supérieures et universités	1995
.gov	gouvernement	organismes gouvernementaux des États-Unis	1995
.info	information	tous	2001
.int	international	organismes internationaux établis par traités internationaux	1998
.name	nom de famille	personnes physiques / individuels	2002
.net	réseau	tous	1995
.mil	militaire	organismes militaires des États-Unis	1995
.museum	musée	musées répondant à la définition de l'International Council of Museums (ICOM)	2001
.org	organisation / association	tous	1995
.pro	professionnel libéral	avocats, médecins, et autres professionnels libéraux	2002

Chaque gTLD est géré par un organisme particulier, qui distribue les accréditations pour disposer d'un nom de domaine sur le domaine concerné ; une

[1] Liste fournie par l'AFNIC, mise à jour le 21 juillet 2003 (voir `http://www.afnic.fr`).

liste d'organismes, les *registrars*, est autorisée à servir d'intermédiaire pour l'achat d'un nom de domaine pour tel ou tel gTLD. Dans les faits, les domaines *.com*, *.org* et *.net* sont disponibles sans justification, et sont les moins chers, ce qui explique leur popularité auprès des concepteurs de sites.

À côté des TLD génériques, on trouve les domaines correspondants à des pays, les ccTLD (*country code Top Level Domains*). Les ccTLD sont ces codes en deux lettres utilisés pour désigner les domaines Internet géographiques ; on compte environ 240 ccTLD à l'heure actuelle : *.fr* pour la France, *.ru* pour la Russie, etc.

Afin de faciliter l'échange international des biens et des informations, l'Organisme International de Normalisation (ISO) a établi en 1974 une norme internationale de codes pays afin d'identifier les pays ou zones géographiques. Cette norme s'appelle ISO 3166 et est par exemple utilisée pour les codes postaux nationaux. La norme ISO 3166, dans sa version deux lettres (ISO 3166-1), est également utilisée pour déterminer les ccTLD, les domaines géographiques de deux lettres utilisés sur Internet. Par exemple, la France a pour code ISO *FR* et pour ccTLD *.fr*.

Chaque État gère lui-même le domaine qui lui revient, et dispose d'une instance régulatrice propre ; en France, il s'agit de l'AFNIC. Ainsi, les règles d'accès à des noms de domaines varient selon les pays : l'achat d'un domaine en *.fr* était jusqu'à récemment restreint aux entreprises et organismes institutionnels (toute personne souhaitant enregistrer un nom de domaine en *.fr* et *.re* devait posséder un droit sur le nom de domaine demandé : par exemple en justifiant d'une marque déposée, d'une raison sociale, d'une enseigne, etc.). L'AFNIC a également mis en place des sous-domaines spécifiques tels que *.asso.fr* pour les associations, *.nom.fr* pour les individus, etc. : dans ce cas, l'achat d'un nom de domaine se fait dans le sous-domaine spécifié, par exemple : `crimlangueso.asso.fr`.

2.2 Protocoles

Une fois l'échange de données rendu possible *via* TCP/IP, les deux machines qui communiquent doivent savoir comment échanger ces données : c'est le rôle des protocoles.

2.2.1 Principe

Les protocoles sont des spécifications techniques établissant les règles de communication entre deux machines. Ils spécifient très précisément quel est le format et le type des données échangées, les messages d'erreurs éventuels, et les interactions possibles entre client et serveur.

Dans une architecture client-serveur, le serveur « attend » les requêtes du client, et y répond. La chaîne de communication entre les deux peut être plus ou moins complexe : pour un serveur Web sans authentification, le client envoie une requête, que le serveur exécute, il n'y a donc qu'un aller et un retour. Dans le cas de procédures plus complexes incluant notamment l'authentification du client (par exemple un service FTP non anonyme), celui-ci commence dans un premier temps à

faire une demande d'accès à la ressource, accède à l'interface d'authentification, renvoie ces informations et, si celles-ci sont correctes, est finalement connecté au serveur FTP où il peut exécuter toute une série de commandes dans une durée *a priori* illimitée (dans les faits, un délai d'inactivité automatise la déconnexion par le serveur). L'ensemble de ces schémas de communication sont codifiés et décrits dans la définition du protocole utilisé.

Rappelons ici que le terme « serveur » recoupe souvent une acception physique, qui désigne une machine dotée d'une configuration particulière, et logique, qui indique qu'un logiciel faisant office de serveur fonctionne sur la machine. La confusion des deux sens tient au fait que, dans le cadre d'applications industrielles traitant de gros volumes de données, les fonctions logicielles de serveur (serveur Web, serveur FTP, etc.) sont supportées par des architectures matérielles dédiées à cette utilisation. En réalité, n'importe quelle machine peut remplir les fonctions de serveur au sens logiciel du terme, c'est uniquement sa capacité à répondre à un afflux de requêtes trop important qui pourra la rendre inapte à remplir cette fonction.

2.2.2 Protocoles les plus utilisés sur Internet

Les protocoles les plus couramment utilisés sur Internet sont les suivants :

- HTTP : utilisé dans la communication avec les serveurs Web (également appelés « serveur HTTP ») ; le logiciel client utilisé est un navigateur.
- FTP : protocole dédié au transfert de fichiers, il permet soit d'envoyer et de récupérer un ou plusieurs fichiers de/vers une machine distante. Il inclut des fonctions d'authentification par nom d'utilisateur et mot de passe. La plupart des navigateurs modernes prennent en charge le protocole FTP, mais il existe également des clients dédiés qui permettent de gérer plus finement les paramètres de connexion.
- POP3 et SMTP : protocoles utilisés pour le courrier électronique, respectivement pour la réception et l'envoi des messages vers un serveur de messagerie. Ils nécessitent l'utilisation d'un logiciel client spécifique, dont les plus connus sont Outlook Express, Mozilla (ou Thunderbird) et Eudora.
- NNTP : dédié aux échanges sur les forums, il est dans la plupart des cas géré par les clients de messagerie.
- ICQ, IRC, MSN MESSENGER, etc. : protocoles utilisés pour les services de messagerie instantanée (*chat*), permettant des échanges synchrones dans des espaces publics ou en privé deux à deux. Ces protocoles ne sont pas compatibles entre eux, et nécessitent des clients spécifiques ; certains logiciels permettent cependant d'accéder à ces différents types de serveurs.

D'autres protocoles complètent cette liste, notamment ceux utilisés dans les jeux en réseau, ceux dédiés aux échanges sécurisés (SSH et SFTP), ou ceux utilisés pour les échanges de fichiers en *peer-to-peer*. La plupart du temps, ils correspondent à des fonctionnalités particulières (communication, échange de fichier, etc.), bien que beaucoup d'applications soient accessibles *via* des interfaces HTTP (WebMail, WebChat, forums, jeux, etc.) alors qu'elles nécessitaient auparavant l'utilisation d'un logiciel client spécifique.

2.3 Requêtes HTTP

Nous donnons ici le détail du fonctionnement du protocole HTTP, afin de montrer quelles informations les sondes de recueil de trafic peuvent recueillir, comment elles le font, et à quels biais elles sont soumises.

2.3.1 Communication entre client et serveur

Une requête HTTP est un ensemble de lignes envoyées au serveur par le navigateur. Elle comprend :

- *une ligne de requête* : c'est une ligne précisant le type de document demandé, la méthode qui doit être appliquée, et la version du protocole utilisée. La ligne comprend trois éléments devant être séparés par un espace:
  - la *méthode*
  - la ressource demandée sur le serveur
  - la version du protocole utilisé par le client (généralement HTTP/1.0)
- *les champs d'en-tête de la requête* : il s'agit d'un ensemble de lignes facultatives permettant de donner des informations supplémentaires sur la requête et/ou le client (navigateur, système d'exploitation, langue, etc.). Chacune de ces lignes est composée d'un nom qualifiant le type d'en-tête, suivi de deux points (:) et de la valeur de l'en-tête.
- *une ligne vide* : elle assure la séparation entre l'en-tête et le reste de la requête.
- *le corps de la requête* : c'est un ensemble de lignes optionnel devant être séparé des lignes précédentes par une ligne vide et permettant par exemple un envoi de données. Dans l'envoi d'une requête par le client, le corps de la requête n'est renseigné que lors d'une requête de type POST (envoi de données au serveur par un formulaire) ; dans la réponse des serveurs, il l'est systématiquement et contient la source HTML des pages, le contenu des fichiers d'images, etc.

Une requête HTTP a donc la syntaxe suivante (<crlf> signifie retour chariot ou saut de ligne) :

```
METHODE URL VERSION<crlf>
EN-TÊTE : Valeur<crlf>
[…]
EN-TÊTE : Valeur<crlf>
Ligne vide<crlf>
CORPS DE LA REQUETE
```

Voici ci-dessous un exemple de requête HTTP, demandant la page `/index.html` sur le serveur `www.globz.net` ; cette requête correspond, dans la barre d'adresse d'un navigateur, à `http://www.globz.net/index.html`.

```
GET /index.html HTTP/1.1
Host: www.globz.net
Accept : text/html
If-Modified-Since : Saturday, 15-January-2000 14:37:11 GMT
User-Agent : Mozilla/4.0 (compatible; MSIE 5.0; Windows 95)
```

Lors de l'envoi de la requête par le client, plusieurs « méthodes » sont utilisables, qui correspondent à l'envoi de données à des formats différents (GET vs. POST), où à la demande de résultats sensiblement différentes (GET pour récupérer les données, HEAD pour connaître des informations sur la ressource). Au total, le protocole http en version 1.1 définit huit méthodes :

Méthode	Description
GET	Obtient le contenu de la ressource spécifiée
HEAD	Obtient l'en-tête de la réponse uniquement
POST	Envoie de contenu au serveur (utilisé par certains types de formulaires)
PUT	Demande au serveur d'enregistrer les données envoyées (peu utilisé)
DELETE	Permet d'effacer un fichier sur le serveur (peu utilisé)
TRACE	Permet de contrôler la requête reçue par le serveur (peu utilisé)
CONNECT	Mot réservé pour les proxies permettant de créer des tunnels
OPTIONS	Liste les options possibles pour une ressource donnée (peu utilisé)

Les en-têtes possibles de la requête du client sont les suivants :

Nom de l'en-tête	Description
Accept	Type de contenu MIME accepté par le navigateur (ex : *text/html*).
Accept-Charset	Jeu de caractères attendu par le browser
Accept-Encoding	Codage de données accepté par le browser
Accept-Language	Langage attendu par le browser (anglais par défaut)
Authorization	Identification du browser auprès du serveur
Content-Encoding	Type de codage du corps de la requête
Content-Language	Type de langage du corps de la requête
Content-Length	Longueur du corps de la requête
Content-Type	Type de contenu MIME du corps de la requête (ex : *text/html*).
Date	Date de début de transfert des données
Forwarded	Utilisé par les machines intermédiaires entre le browser et le serveur
From	Permet de spécifier l'adresse e-mail du client
Link	Relation entre deux URL
Orig-URL	URL d'origine de la requête
Referer	URL du lien à partir duquel la requête a été effectuée
User-Agent	Chaîne donnant des informations sur l'équipement du client : nom et la version du navigateur, système d'exploitation.

Les en-têtes de la réponse du serveur sont les suivants :

Nom de l'en-tête	Description
Content-Encoding	Type de codage du corps de la réponse
Content-Language	Type de langage du corps de la réponse
Content-Length	Longueur du corps de la réponse
Content-Type	Type de contenu MIME du corps de la réponse (ex : *text/html*)
Date	Date de début de transfert des données
Expires	Date limite de consommation des données
Forwarded	Utilisé par les machines intermédiaires entre le client et le serveur
Location	Redirection vers une nouvelle URL associée au document
Server	Caractéristiques du serveur ayant envoyé la réponse

Les codes de réponses sont envoyés par le serveur pour indiquer la réussite ou non de la requête et, en cas d'échec, en donner la cause. Ce sont les codes que l'on voit lorsque le navigateur n'arrive pas à fournir la page demandée. Le code de réponse est constitué de trois chiffres : le premier indique la classe de statut et les suivants la nature exacte de l'erreur. La famille 1xx n'est plus utilisée, on en compte quatre aujourd'hui dans la version 1.1 de HTTP :

- codes 2xx : réussite, indiquent le bon déroulement de la requête :

Code	Message	Description
201	CREATED	Elle suit une command POST, elle indique la réussite, le corps du reste du document est sensé indiquer l'URL a laquelle le document nouvellement créé devrait se trouver.
202	ACCEPTED	La requête a été acceptée, mais la procédure qui suit n'a pas été accomplie
203	PARTIAL INFORMATION	Lorsque ce code est reçu en réponse à une commande GET, cela indique que la réponse n'est pas complète.
204	NO RESPONSE	Le serveur a reçu la requête mais il n'y a pas d'information a renvoyer
205	RESET CONTENT	Le serveur indique au navigateur de supprimer le contenu des champs d'un formulaire
206	PARTIAL CONTENT	le serveur a répondu partiellement à la requête GET

- codes 3xx : redirection, indiquent que la ressource n'est plus à l'emplacement indiqué :

Code	Message	Description
301	MOVED	Les données demandées ont été transférées a une nouvelle adresse
302	FOUND	Les données demandées sont à une nouvelle URL, mais ont cependant peut-être été déplacées depuis
303	SEE OTHER	Cela implique que le client doit essayer une nouvelle adresse, en essayant de préférence une autre méthode que GET
304	NOT MODIFIED	Si le client a effectué une commande GET conditionnelle (en demandant si le document a été modifié depuis la dernière fois) et que le document n'a pas été modifié il renvoie ce code.
305	USE PROXY	la ressource demandée doit être accédée en utilisant le proxy indiqué
306	(Unused)	ce code est réservé (il était utilisé dans un premier draft de la RFC2616)
307	TEMPORARY REDIRECT	la ressource demandée se trouve temporairement à une autre URI

- codes 4xx : erreur due au client, la requête est incorrecte :

Code	Message	Description
400	BAD REQUEST	La syntaxe de la requête est mal formulée ou est impossible à satisfaire
401	UNAUTHORIZED	Le paramètre du message donne les spécifications des formes d'autorisation acceptables. Le client doit reformuler sa requête avec les bonnes données d'autorisation
402	PAYMENT REQUIRED	Le client doit reformuler sa demande avec les bonnes données de paiement
403	FORBIDDEN	L'accès à la ressource est tout simplement interdit
404	NOT FOUND	Le serveur n'a rien trouvé à l'adresse spécifiée
405	METHOD NOT ALLOWED	le client essaie d'utiliser une méthode non autorisée sur l'URI demandée. Le serveur renvoie alors une directive Allow: pour

		indiquer quelles méthodes sont autorisées.
406	NOT ACCEPTABLE	la réponse (entité) ne correspond pas aux caractéristiques de la directive Accept: de l'en-tête de la requête
407	PROXY AUTHENTICATION REQUIRED	identique au code 401, mais il indique que le client doit d'abord s'authentifier auprès du proxy
408	REQUEST TIMEOUT	le client n'a pas envoyé de requête durant la période de temps où le serveur attendait
409	CONFLICT	il y a un conflit entre la requête et l'état actuel de la ressource. Le client peut a priori résoudre le problème.
410	GONE	la ressource n'est plus disponible sur le serveur et aucune adresse alternative n'a été fournie
411	LENGTH REQUIRED	la requête doit contenir un Content-Length:
412	PRECONDITION FAILED	une des préconditions fournies en en-tête de la requête a produit un résulat négatif du côté serveur
413	REQUEST ENTITY TOO LARGE	la ressource demandée est plus grosse que ce que le serveur veut renvoyer
414	REQUEST-URI TOO LONG	l'URI de la ressource demadée est trop longue. Cette erreur se produit par exemple lorsque le client a mal converti une requête POST en requête GET.
415	UNSUPPORTED MEDIA TYPE	le format de l'entité demandée n'est pas supporté par la ressource demandée pour la méthode demandée
416	REQUESTED RANGE NOT SATISFIABLE	le client demande un Range: (portion de l'entité) impossible à déterminer sur la ressource
417	EXPECTATION FAILED	la prévision de ressource exprimée dans le champ Expect: de la requête ne peut pas être satisfaite

- codes 5xx : erreur due au serveur :

Code	Message	Description
500	INTERNAL ERROR	Le serveur a rencontré une condition inattendue qui l'a empêché de répondre à la requête
501	NOT IMPLEMENTED	Le serveur ne supporte pas le service demandé
502	BAD GATEWAY	Le serveur a reçu une réponse invalide de la part du serveur auquel il essayait d'accéder en agissant comme une passerelle ou un proxy
503	SERVICE UNAVAILABLE	Le serveur ne peut pas vous répondre à l'instant présent, car le trafic est trop dense
504	GATEWAY TIMEOUT	La réponse du serveur a été trop longue vis à vis du temps pendant lequel la passerelle était préparée à l'attendre.
505	HTTP VERSION NOT SUPPORTED	le serveur ne supporte pas la version HTTP demandée. Le serveur devrait répondre pourquoi cette version n'est pas supportée, et quelles versions le sont.

2.3.2 Rôle du navigateur

Le navigateur prend en charge l'ensemble de ces processus de communication avec les serveurs Web : il formule les requêtes, interprète les résultats, les met en forme. Il s'occupe également, lorsqu'il interprète les pages html, d'y repérer les appels à des éléments extérieurs entrant dans la composition de la page (les images, en particulier). Pour chaque élément, il reformule une requête auprès des serveurs, et incorpore le résultat dans la page affichée. En outre, il renseigne les champs optionnels dans les en-têtes des requêtes HTTP : il informe notamment les serveurs Web sur le navigateur utilisé (gestion des capacités des navigateurs à traiter le contenu des documents renvoyés, ce qui peut amener les serveurs à renvoyer vers

une page d'erreur), la langue préférentielle (utilisé souvent pour des redirections automatiques ou l'adaptation du contenu des pages), et le *referer* (URL d'où provient l'internaute lors du suivi d'un lien). Un exemple permet d'observer dans le détail ce travail du navigateur : il s'agit d'une requête envoyée à l'adresse `http://www.google.fr/` (page d'accueil française de Google) en utilisant le navigateur Firefox 0.8.

On entre l'adresse demandée dans la barre d'adresse : de ce fait, le champ *referer* n'est pas renseigné ; par contre, le serveur sait quelle est la langue préférentielle de l'utilisation (champ *Accept-language*) et le type de fichiers acceptés (champ *Accept*). La requête envoyée est la suivante :

```
GET / HTTP/1.1
Host: www.google.fr
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US;
rv:1.6) Gecko/20040206 Firefox/0.8
Accept: application/x-shockwave-flash,text/xml,application/xml,
application/xhtml+xml,text/html;q=0.9,text/plain;q=0.8,video/x-
mng,image/png,image/jpeg,image/gif;q=0.2,*/*;q=0.1
Accept-Language: fr,en-us;q=0.7,en;q=0.3
Accept-Encoding: gzip,deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Keep-Alive: 300
Connection: keep-alive
```

Le serveur renvoie un code 200, indiquant que la requête est traitée correctement, et indique qu'il renvoie un document au format HTML (en-tête *Content-type*). Les en-têtes HTTP sont suivi d'une ligne vide, puis du source HTML de la page de réponse :

```
HTTP/1.x 200 OK
Cache-Control: private
Content-Type: text/html
Content-Encoding: gzip
Server: GWS/2.1
Content-Length: 1237
Date: Sat, 05 Jun 2004 16:48:30 GMT

<html><head><meta http-equiv="content-type" content="text/html;
charset=UTF-8"><title>Google</title><style><!--
body,td,a,p,.h{font-family:arial,sans-serif;}
.h{font-size: 20px;}
.q{color:#0000cc;}
//-->
</style>
<script>
<!--
function sf(){document.f.q.focus();}
// -->
</script>
</head><body bgcolor=#ffffff text=#000000 link=#0000cc vlink=#551a8b
alink=#ff0000 onLoad=sf()><center><table border=0 cellspacing=0
cellpadding=0><tr><td><img src="/intl/fr_fr/images/logo.gif"
width=276 height=110 alt="Google"></td></tr></table><br><form
action="/search" name=f><span id=hf></span><script><!--
function qs(el) {if (window.RegExp && window.encodeURIComponent) {var
qe=encodeURIComponent(document.f.q.value);if
(el.href.indexOf("q=")!=-1) {el.href=el.href.replace(new
RegExp("q=[^&$]*"),"q="+qe);} else {el.href+="&q="+qe;}}return 1;}
```

```
// -->
</script><table border=0 cellspacing=0 cellpadding=4><tr><td nowrap
class=q><font size=-1><b><font
color=#000000>Web</font></b>    <a id=1a class=q
href="/imghp?hl=fr&tab=wi" onClick="return
qs(this);">Images</a>    <a id=2a class=q
href="/grphp?hl=fr&tab=wg" onClick="return
qs(this);">Groupes</a>    <a id=3a class=q
href="/dirhp?hl=fr&tab=wd" onClick="return
qs(this);">Annuaire</a>    <a id=4a class=q
href="/nwshp?hl=fr&tab=wn" onClick="return
qs(this);">Actualités</a>    </font></td>
</tr></table><table cellspacing=0 cellpadding=0><tr valign=middle><td
width=25%> </td><td align=center><input maxLength=256 size=55
name=q value="">
<script>
document.f.q.focus();
</script>
<input type=hidden name=ie value="UTF-8"><input name=hl type=hidden
value=fr><br><input type=submit value="Recherche Google"
name=btnG><input type=submit value="J'ai de la chance"
name=btnI></td><td valign=top nowrap width=25%><font size=-
2>  <a href=/advanced_search?hl=fr>Recherche
avancée</a><br>  <a
href=/preferences?hl=fr>Préférences</a><br>  <a
href=/language_tools?hl=fr>Outils
linguistiques</a></font></td></tr><tr><td colspan=3
align=center><font size=-1>Rechercher dans : <input id=all type=radio
name=meta value="" checked><label for=all> Web</label><input id=lgr
type=radio name=meta value="lr=lang_fr" ><label for=lgr> Pages
francophones</label><input id=cty type=radio name=meta
value="cr=countryFR" ><label for=cty>Pages :
France</label></font></td></tr></table></form><p><font size=-
1><p></font><br><br><font size=-1><a href=/intl/fr/ads/>Publicité</a>
- <a href=http://toolbar.google.com/intl/fr/>Google Toolbar</a> - <a
href=/intl/fr/about.html>À propos de Google</a> - <a
href=http://www.google.com/ncr>Google.com in
English</a></font><p><font size=-2>&copy;2004 Google - Nombre de
pages Web recensées par Google :
4,285,199,774.</font></p></center></body></html>
```

Le navigateur affiche la page en même temps qu'il repère que celle-ci contient une image (mis en gras dans le code HTML de la page). Pour la récupérer, il envoie une nouvelle requête au serveur, à l'adresse mentionnée dans le lien vers le fichier image (`http://www.google.fr/intl/fr_fr/images/logo.gif`) :

```
GET /intl/fr_fr/images/logo.gif HTTP/1.1
Host: www.google.fr
User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US;
rv:1.6) Gecko/20040206 Firefox/0.8
Accept: image/png,image/jpeg,image/gif;q=0.2,*/*;q=0.1
Accept-Language: fr,en-us;q=0.7,en;q=0.3
Accept-Encoding: gzip,deflate
Accept-Charset: ISO-8859-1,utf-8;q=0.7,*;q=0.7
Keep-Alive: 300
Connection: keep-alive
Referer: http://www.google.fr/
```

La réponse du serveur indique en en-tête que la requête a abouti, et que l'image est au format (champ *Content-type*) :

```
HTTP/1.x 200 OK
Content-Type: image/gif
Last-Modified: Mon, 22 Mar 2004 23:04:36 GMT
Expires: Sun, 17 Jan 2038 19:14:07 GMT
Server: GWS/2.1
Content-Length: 8866
Date: Sat, 05 Jun 2004 16:48:30 GMT
```

Pour composer cette page, le navigateur a donc envoyé deux requêtes à destination du serveur Web `www.google.fr`, et assemblé ces éléments pour composer l'interface graphique.

Les sondes de recueil de trafic, en se plaçant entre la couche applicative et la couche TCP/IP, tracent l'ensemble de ces requêtes HTTP : elles extraient certaines informations à partir des en-têtes HTTP (adresse du serveur, ressource demandée sur le serveur, *referer*, etc.) pour les envoyer vers des serveurs de collectes. Pour chaque protocole, une sonde doit avoir un module adapté à sa syntaxe afin de savoir quelles informations doivent être extraites, et dans quel format elles apparaissent. C'est ce matériau qui forme les données brutes pour l'analyse des usages d'Internet.

Annexe 3

Inverser la perspective

Nous avons construit et évalué des méthodes de description des contenus à partir des adresses des pages visitées afin de décrire les parcours. Si la recherche d'éléments de description des contenus a pour visée l'analyse des parcours, ces descriptions peuvent être mobilisées comme outil de fouille des données pour l'analyse des usages de certains types de contenus particuliers. Nous décrivons ici cette utilisation des descriptions de contenu dans ce contexte, l'outillage que nous avons développé pour cela, ainsi qu'une étude à laquelle nous avons participé qui met en application ces techniques.

3.1 Description

Nous avons dans notre travail de thèse proposé des méthodes de description du contenu des URL visitées par les internautes, que nous avons mobilisées pour décrire les thèmes et services. Nous avons tenté d'avoir une caractérisation la plus large possible des données de trafic, et mobilisé les descriptions en masse pour traiter la diversité des comportements. La mobilisation des descriptions de contenu ne se limite pas à cette utilisation : ils peuvent, à l'inverse, être employés pour la fouille des données de trafic afin d'identifier les pages correspondant à des thématiques particulières en vue d'études ciblées sur des usages particuliers.

Le problème alors posé est relativement simple : comment savoir, à partir de la liste des URL visitées par un panel d'internautes, lesquelles sont relatives à un thème donné ? Un moyen, long et fastidieux, serait de parcourir manuellement le Web à la recherche d'adresses, de noter les liens, de classer ces résultats, et de les projeter ensuite sur les URL effectivement visitées ; un autre moyen serait de visiter toutes les pages vues par les panélistes pour vérifier si elles correspondent au thème recherché.

Nos outils de description de contenu peuvent apporter une solution plus simple et surtout moins coûteuse en temps à ce type de recherche, en utilisant les données recueillies à partir des annuaires du Web. Il s'agit pour cela de chercher dans les annuaires les sites répondant aux critères de sélection recherchés, et de projeter les sites correspondant sur les URL visitées par les panélistes.

Nous avons pour cela développé une application baptisée *TopicFinder* capable de fouiller les annuaires pour en obtenir des listes d'URL répondant à un critère donné. Cet outil, développé sous la forme de servlets Java, est intégré à la plateforme de traitement des données de trafic développée dans le cadre du projet SensNet.

TopicFinder propose une interface Web permettant à l'utilisateur d'interroger les données recueillies auprès des annuaires à partir d'un mot-clef correspondant au thème recherché dans les données de trafic. L'application cherche ensuite l'ensemble des intitulés de catégories d'annuaires qui contiennent ce mot-clef (requête SQL de type 'LIKE'). Les catégories correspondantes ainsi que leurs sous-catégories sont ensuite soumises à la validation de l'utilisateur, qui peut désélectionner celles qui ne répondent pas à ses besoins : la projection du mot-clef ne fait pas l'économie de la polysémie ni des inclusions au sein d'un mot de la chaîne recherchée. La Figure 3.1 présente un exemple de cette interface de validation : on cherche ici dans les données les pages visitées relatives à Victor Hugo, à l'aide du mot-clefs 'hugo'. Quatre annuaires contiennent des catégories correspondant à cette requête, mais il faut écart les catégories « Hugo Boss » et « Hugo Pratt », décochées dans l'exemple.

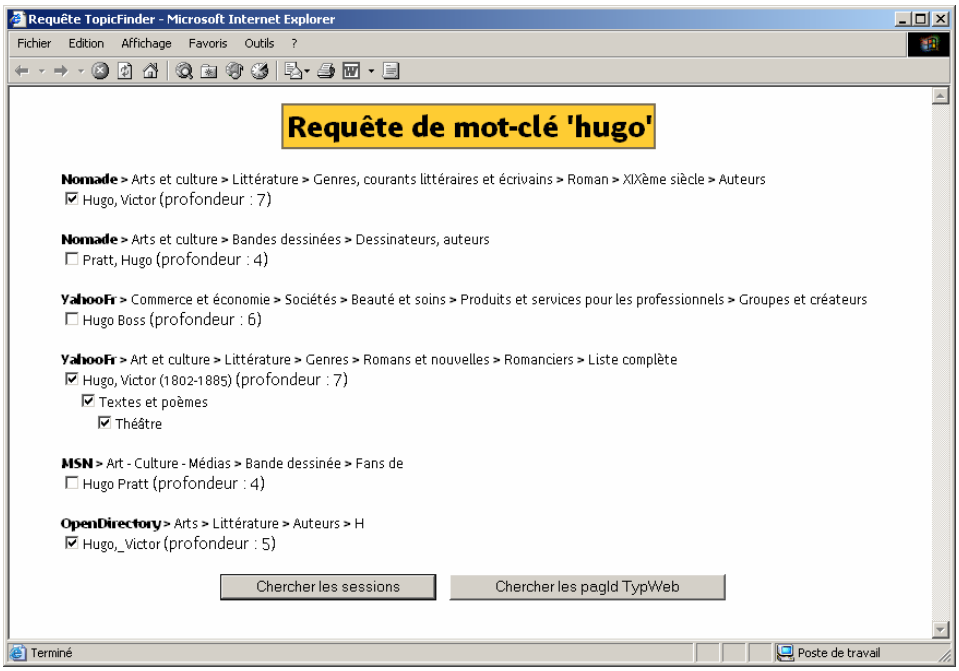


Figure 3.1. Interface de validation des catégories d'annuaires de TopicFinder

À partir des catégories retenues, *TopicFinder* recherche l'ensemble des URL qu'elles contiennent, les projette sur les URL visitées, et renvoie la liste de ces URL, ainsi que les identifiants des sessions et des panélistes correspondants.

Cette méthode donne des résultats variables selon les critères de recherche, et privilégie les sites à forte notoriété, et dont les concepteurs souhaitent les faire connaître, dans la mesure où l'inscription dans un annuaire est initialement une

démarche de la part du site. Ainsi, sur la musique par exemple, les sites de piratage, qui cultivent la confidentialité, échappent à l'analyse.

Partant de ce constat, nous considérons que cette méthode n'est pas suffisante en soi, mais peut servir d'amorçage pour identifier des panélistes potentiellement intéressants, qu'un examen manuel au niveau de la session viendra compléter (en utilisation l'application *RePlay* notamment). Elle se veut également complémentaire de deux autres méthodes de fouille sélective :

- requêtes moteur : à partir de l'extraction par *CatService* du contenu des requêtes adressées aux moteurs de recherche, il s'agit d'identifier les requêtes relatives au thème visé. Ceci implique de dresser une liste des lexicalisations possible du thème en question (liste de noms communs, mais aussi de noms propres et de marques) ;
- noms de sites et noms de répertoires dans les URL : à la manière du travail que nous avons mené sur les noms de répertoires pour y identifier des mots de la langue (voir 3.1.2, « Noms de répertoires »), on peut chercher dans les URL visitées des chaînes de caractères correspondant à une lexicalisation du thème recherché.

Ces deux méthodes impliquent en premier lieu de construire une liste des mots relatives au thème recherché, opération qui peut être longue. Elle nécessite surtout une validation de la projection de ce lexique sur les requêtes moteur ou les URL visitées, travail coûteux en temps. Pour autant, ces méthodes sont complémentaires de l'approche par les annuaires ; leur mise en application a permis d'en évaluer la difficulté et l'efficacité.

3.2 Mise en application : étude « Loft Story »[1]

À la demande du département Développement et Prospective de Wanadoo, le laboratoire UCE de France Télécom R&D a mené une étude sur l'entrelacement des médias dans la constitution des publics de l'émission Loft Story. Cette enquête mêle approche quantitative des parcours des internautes lofteurs (méthodologie SensNet) et entretiens qualitatifs avec des « fans » du Loft participant au *chats* et aux forums sur le thème du Loft. Elle montre les potentialités d'Internet, comme relais d'information et d'échanges interpersonnels, pour les émissions de télévision à grande audience. Les usages d'Internet sont partiellement « synchronisés » avec le flux télévisuel et les usagers créent des communautés extrêmement denses, actives, ironiques et éphémères pour discuter du programme.

L'analyse de la fréquentation des sites liés à l'émission repose sur des données de trafic Internet d'une population de 1540 internautes observés pendant toute l'année 2001. Ces données sont fournies par la société NetValue dans le cadre du partenariat SensNet. Nous disposons ainsi, pour chaque internaute du panel, de la liste complète et horodatée des adresses des pages qu'il a visitées en 2001.

---

[1] Nous reprenons ici les éléments présentés dans [Beaudouin *et al.* 2003a].

Dans un premier temps, nous avons identifié, au sein de l'ensemble des pages visitées par le panel en 2001, celles de sites en lien avec l'émission Loft Story. Pour cela, trois méthodes ont été mises en œuvre :

- utilisation des annuaires du Web : l'application *TopicFinder* permet de projeter sur le trafic du panel NetValue les sites des catégories « Loft Story » de 8 annuaires du Web francophone.
- examen manuel : on examine manuellement les URL visitées contenant les chaînes de caractère 'loft', 'story', 'loana', etc. Par exemple : `http://loanaforever.free.fr`.
- requêtes moteur : l'application *CatService* permet d'extraire les mots-clefs demandés sur les moteurs de recherche. On identifie ensuite manuellement les recherches relatives au Loft.

Les trois méthodes sont complémentaires ; au terme de ces trois opérations, nous identifions près de 7 900 URL relatives au Loft (voir Tableau 3.1), représentant plus de 70 000 pages vues (une page pouvant être vue plusieurs fois, par un ou plusieurs internautes).

Tableau 3.1. Méthodes d'identification des pages relatives au Loft

	URL uniques	URL vues
Annuaires	925 (11,7 %)	38 148 (53,9 %)
Manuel	7 025 (89,1 %)	67 302 (95,1 %)
Moteurs	677 (8,6 %)	1777 (2,5 %)
Total	7 885 (100 %)	70 769 (100 %)

Les annuaires identifient peu d'adresses (11,7 %) du total, mais celles-ci correspondent aux sites à forte notoriété concentrant une part importante de l'audience (53,9 % du total) ; à l'inverse, les URL correspondant à des requêtes moteur ne permettent d'identifier que 2,5 des pages vues, pour 8,6 % des pages distinctes, mais elles sont toutefois un moyen efficace d'amorçage pour identifier des sessions. La méthode la plus efficace demeure l'examen manuel des URL après projection de mots-clefs, mais c'est aussi la plus coûteuse en temps.

On identifie avec cette méthodologie près de 600 sites relatifs à « Loft Story » visités en 2001. Parmi ce nombre important, on constate que les 5 sites officiels (`www.loftstory.fr`, `www.loftstory.com`, sites de M6) drainent 70% des pages vues :

- les ¾ des internautes ayant vu un « site Loft » ont visité un site officiel (425 sur les 573 ayant visité un « site Loft ») ;
- les sites officiels sont présents dans 64 % des sessions contenant des pages « Loft ».

Les sites officiels apparaissent de ce point de vue comme un point de passage incontournable, et l'audience des sites non officiels se révèle plus éparse, mais elle demeure importante :

- ces sites ont attiré 39 % des panélistes Loftiens (224 panélistes sur les 573) ;
- ils ne représentent que 27 % des pages vues, mais ils sont présents dans 55 % des sessions ;

- beaucoup de ces sites sont ironiques vis à vis de l'émission (Bofstory, Loststory, etc.) ou érotiques / pornographiques ; les sites « sérieux » de fans font moins recette.

Cette forte dispersion justifie que l'on déploie les efforts nécessaires pour identifier dans les données de trafic les sites relatifs à l'émission ; pour efficace et rapide qu'elle soit, l'utilisation seule des annuaires nous aurait fait perdre de vue la moitié de l'audience des sites sur le Loft, sites personnels critiques, ironiques, d'information « de deuxième main » à l'audience éparse qui constituent néanmoins un contrepoint au flux « officiel » indispensable à la construction des publics de l'émission.

Cette identification a permis par la suite de cerner les publics-internautes de l'émission : distinction entre faibles loftiens et fans, profil socio-démographique particulier, structuration temporelle des visites par la rythmique quotidienne et hebdomadaire de l'émission. S'appuyant également sur des enregistrements d'échanges sur des *chats* et des entretiens avec des participants de forums dédiés à l'émissions, cette enquête a permis, en confrontant ces différents matériaux, de montrer à quel point les programmes télévisés pouvaient être un support et un promoteur décisif pour la pratique de l'Internet. L'étude des pratiques concrètes remet en question les stéréotypes attachés au média Internet. En effet, on considère trop souvent les pratiques des internautes comme des activités individuelles détachées de toute dimension collective et indépendante du temps public des grands événements du direct de la télévision. Nous nous représentons Internet comme un média de communication interpersonnel, inadapté aux grandes audiences de masse de la télévision et à la diffusion de l'information en temps réel. Comme le rappellent les militants de l'Internet qui défendent une relation interactive, égalitaire et symétrique entre producteurs et récepteurs d'informations, beaucoup de choses opposent ces deux médias : offre limitée/illimitée, logique de masse transclassiste/audience communautaire, terminal partagé/personnalisé, réception collective/pratique individuelle, etc. Il n'en reste pas moins que, à l'instar de Loft Story et des autres programmes de télé-réalité, l'imbrication progressive des médias interpersonnels dans le fonctionnement des médias de masse de l'espace public traditionnel est de plus en plus importante, ce qui atteste des logiques croissantes d'association ou de complémentarité entre programmes télévisuels et ressources en ligne.

Annexe 4

Matériau d'enquête BibUsages

Cette annexe présente le matériau d'enquête utilisé au cours du projet BibUsages : il s'agit d'une part du questionnaire soumis aux visiteurs de Gallica, et d'autre part de la grille construite pour mener les entretiens.

4.1 Questionnaire en ligne

Nous présentons ici le questionnaire en ligne soumis aux visiteurs de Gallica (`http://gallica.bnf.fr`) durant trois semaines en mars 2002. Ce questionnaire était proposé aux internautes accédant à la page d'accueil du site, sous forme de *pop-up* ; il n'était vu qu'une fois, que l'internaute y ait répondu ou non.

Présentation de l'étude : *La Bibliothèque Nationale de France réalise une étude afin de mieux connaître les visiteurs de son site gallica.bnf.fr et mieux cerner la façon dont ils utilisent ce site Gallica. Soyez assuré(e) que toutes les réponses resteront strictement confidentielles. Ce questionnaire durera environ 10 minutes.*

Q1. Approximativement, combien de fois avez-vous déjà visité le site « gallica.bnf.fr » au cours des 6 derniers mois ?

1. C'est ma première visite
2. moins de 5 fois
3. 5 à 10 fois environ
4. 10 à 20 fois environ
5. Plus de 20 fois
6. (vous ne savez pas)

Q2. Diriez-vous qu' au cours des 6 prochains mois vous consulterez le site « gallica.bnf.fr » …..

1. Régulièrement
2. De temps en temps
3. Rarement
4. (vous n'aurez plus l'occasion de le consulter)
5. (vous ne savez pas)

Q3. D'où vous connectez-vous pour consulter Le site Gallica ?
(Plusieurs réponses possibles)
1. De chez vous
2. De votre lieu de travail
3. De l'Université (pour les étudiants) ou d'une école supérieure d'ingénieurs, de commerce, …
4. D'un lycée ou collège
5. D'un lieu public tel qu'une bibliothèque, un cybercafé …
6. D'un autre lieu

Q4. (si Q3=6) **De quel autre lieu ?**
(question ouverte)

Q5. Comment avez-vous découvert notre site ?
1. Vous avez vu l'adresse du site Gallica dans une brochure ou documentation de la Bibliothèque Nationale de France
2. Vous l'avez trouvé par un lien à partir d'un autre site ou dans un message (e-mail, forum, …)
3. Vous l'avez trouvé par un moteur de recherche comme Yahoo!, Voilà, Google ……
4. Par un ami, une relation
5. Autre mode

Q6. (si Q5=5) **Lequel ?**
(question ouverte)

Q7. Approximativement, combien de temps avez-vous l'habitude de passer sur Gallica ?
1. Moins de 5 minutes
2. 5 à 10 minutes environ
3. 10 à 30 minutes environ
4. Plus de 30 minutes
5. (vous ne savez pas)

Q8. Habituellement, pourquoi venez-vous sur Gallica ?
(question ouverte)

Q9. Parmi les rubriques suivantes, quelles sont celles que vous avez consultées au cours de vos dernières visites ?
(plusieurs réponses possibles)
1. Gallica Découverte : Thèmes
2. Gallica Découverte : Chronologies
3. Gallica Découverte : Iconographies, monnaies
4. Gallica Découverte : Dictionnaires
5. Gallica Découverte : Mode texte
6. Gallica Recherche
7. Les dossiers de Gallica : Classique
8. Les dossiers de Gallica : Utopie

9. Les dossiers de Gallica : Proust
10. Les dossiers de Gallica : La voix
11. Sociétés Savantes
12. Voyages en France
13. Aide : Les questions/réponses
14. Aide : Assistance
15. (Je suis resté(e) uniquement sur la page d'accueil)

Q10. (Si Q9 pas 15) **Etes-vous d'accord avec les phrases suivantes à propos du site Gallica ?**

(une seule réponse par ligne)

	Tout à fait d'accord	Plutôt d'accord	Plutôt pas d'accord	Pas du tout d'accord	Ne sait pas
Ce site a un contenu de qualité	1	2	3	4	5
En général, on trouve l'information que l'on cherche	1	2	3	4	5
Le « look » de ce site est réussi	1	2	3	4	5
La présentation des pages est claire	1	2	3	4	5
L'information est présentée de façon attractive	1	2	3	4	5
On trouve FACILEMENT l'information que l'on cherche	1	2	3	4	5
On sait à tout moment où l'on se trouve dans le site	1	2	3	4	5
Le temps de chargement des pages est acceptable	1	2	3	4	5
Le site donne envie de revenir	1	2	3	4	5

Q11. Que souhaiteriez-vous trouver sur le site Gallica ?

(question ouverte)

Q12. Avez-vous enregistré l'adresse du site parmi vos favoris ou bookmarks ?

1. Oui
2. Non

PROFIL INTERNET DES VISITEURS

Q13. Depuis quand utilisez-vous, vous-même, Internet?

1. Depuis 2002
2. Depuis 2001
3. Depuis 2000
4. Depuis 1999
5. Depuis 1998
6. Depuis 1997 et avant

Q14. A quelle fréquence utilisez-vous, vous-même, Internet ?

1. Tous les jours
2. 2 à 5 fois par semaine
3. Environ une fois par semaine
4. 1 à 3 fois par mois
5. Moins souvent

Q15. D'où avez-vous l'habitude de vous connecter à Internet ?
(plusieurs réponses possibles)
1. De chez vous
2. De votre lieu de travail
3. De l'Université (pour les étudiants) ou d'une école supérieure d'ingénieurs, de commerce, …
4. D'un lycée ou collège
5. D'un lieu public tel qu'une bibliothèque, un cybercafé …
6. D'un autre lieu

Q16. (Si Q15=6) **De quel autre lieu ?**
(question ouverte)

Q17. (Si Q15=1) **Quel type de connexion Internet avez-vous à domicile ?**
(Une seule réponse possible)
1. Une connexion par modem
2. Une connexion Numéris
3. Une connexion Haut-débit ADSL
4. Une connexion Haut-débit Câble
5. Autres connexions
6. (vous ne savez pas)

Q18. (si Q17=5) **Quel autre type de connexion Internet avez-vous?**
(question ouverte)

Q19. Quel usage avez-vous d'Internet ?
(plusieurs réponses possibles)
1. Recherche d'information
2. Communication : chat, groupes de discussion, messagerie, emails…..
3. Achat en ligne
4. Opérations et consultations bancaires ou boursières
5. Téléchargement de musique et/ou de vidéo
6. Téléchargement de logiciels
7. Jeux en ligne
8. Autre usage

Q20. (si Q19=8) **Que faites-vous d'autre(s) sur Internet ?**
(question ouverte)

Q21. Quel(s) sont vos PRINCIPAUX centres d'intérêt sur le Web ?
(plusieurs réponses possibles)
1. Actualités
2. Banque et finances

3. Economie et entreprise
4. Institutions et service public
5. Emploi, stage (recherche ou offre)
6. Autres informations économiques ou institutionnelles
7. Sciences Humaines et sociales
8. Art et Littérature
9. Recherche documentaire ou bibliographique
10. Sciences et technologies
11. Informatiques et multimédia
12. Autres informations culturelles
13. Sorties, divertissements
14. Voyages, tourisme
15. Sports
16. Jeux
17. Autres Loisirs
18. Communication
19. Autres centres d'intérêt

Q22. Consultez-vous des sites Web appartenant aux catégories suivantes ?
(plusieurs réponses possibles)

1. Sites d'Université et/ou centres de Recherche
2. Sites de bibliothèques
3. Sites d'établissements culturels (musée, galerie ….)
4. Sites de journaux, de magazines en ligne
5. Sites de e-commerce de biens culturels (tels que la fnac, amazon, alapage ….)
6. Aucun site appartenant à ces catégories

Q23. Consultez-vous le site Gallica pour un usage principalement … ?
(plusieurs réponses possibles)

1. Personnel
2. Professionnel
3. Dans le cadre de vos études

Q24. Quel(s) type(s) d'ordinateur utilisez-vous habituellement pour vous connecter à Internet ? S'agit-il…
(Plusieurs réponses possibles)

1. D'un PC
2. D'un Mac
3. D'une station de travail

Q25. Quel est le système d'exploitation de votre/vos ordinateurs ?
(plusieurs réponses possibles)

1. Windows
2. Mac OS X version 10 ou supérieure
3. Mac OS Système 9 ou antérieur
4. Linux
5. Un autre système Unix
6. (vous ne savez pas)

Q26.A (Si Q15=1) **Partagez-vous votre ordinateur à domicile avec d'autres personnes ?**

1. oui
2. non

Q26B. (si Q15=2) **Partagez-vous votre ordinateur avec d'autres personnes sur votre lieu de travail ?**

1. oui
2. non

Q26C. (Si Q15=3) **Partagez-vous votre ordinateur avec d'autres personnes à l'Université ou dans votre école d'ingénieurs, de commerce … ? oui/Non**

1. oui
2. non

Q26D. (Si Q15=4) Partagez-vous votre ordinateur avec d'autres personnes au lycée ou au collège ?

1. oui
2. non

PROFIL SOCIO-DEMOGRAPHIQUE DES VISITEURS

Q27. Etes-vous :

1. Un homme
2. Une femme

Q28. Quel est votre âge ?

(question quantité)

Q29. Quelle est votre situation familiale ?

1. Célibataire
2. Vit maritalement
3. Marié(e) ou remarié(e)
4. Divorcé(e) ou séparé(e)
5. Veuf(ve)
6. (Vous ne souhaitez pas répondre)

Q30. Avez-vous une activité professionnelle rémunérée ?

1. Oui
2. non

Q31. (Si Q30=1) **Quelle est votre activité professionnelle?**

1. Agriculteur exploitant
2. Commerçant, artisan,
3. Chef d'entreprise, cadre dirigeant
4. Profession libérale
5. Cadre du secteur privé

6. Cadre de la fonction publique (catégorie A)
7. Technicien, agent de maîtrise, contremaître, catégorie B de la fonction publique
8. Employé, personnel de service
9. Ouvrier
10. Etudiant ayant une activité rémunérée

Q32. (Si Q30=1) **Quel est votre secteur d'activité?**
(une seule réponse possible)

1. Agriculture, chasse, exploitation forestière, pêche
2. Industries mécaniques, électroniques, chimiques, agro-alimentaires, production et distribution d'énergie et d'électricité, Imprimerie et autres industries
3. Batiments, Travaux publics
4. Commerce et distribution
5. Transport (terrestres, eau, aériens) et Télécommunications
6. Hotellerie, restauration
7. Etude, conseil, services aux entreprises
8. Informatique
9. Banque, assurance, immobilier
10. Santé et action sociale
11. Arts, spectacles
12. Professions de l'information
13. Professions des Bibliothèques, musées et archives
14. Ecrivain
15. Métiers du livre
16. Enseignement du Primaire
17. Enseignement du Secondaire
18. Enseignement du Supérieur
19. Recherche
20. Administration publique
21. Services aux personnes (blanchisserie, coiffure, soins de beauté, pompes funèbres, activités thermales et de thalassothérapie…)
22. Autres services (Assainissement, voierie et gestion des déchets, services domestiques ….)

Q33. (Si Q30=2 ou Q31=10) **Quelle est votre situation ?**

1. A la recherche d'un emploi
2. Femme/homme au foyer
3. Elève ,lycéen ou étudiant de 1er cycle
4. Etudiant de 2ème cycle
5. Etudiant de 3ème cycle en DEA/DESS
6. Etudiant de 3ème cycle en Doctorat/Thèse
7. Service militaire
8. Clergé, religieux
9. Membre d'une association
10. Retraité
11. Autre situation

Q34. (Si Q30=2 ou Q33=3 ou 4 ou 5 ou 7) **Quelle est l'activité professionnelle du chef de famille ?**

1. Agriculteur exploitant
2. Commerçant, artisan,

3. Chef d'entreprise, cadre dirigeant
4. Profession libérale
5. Cadre du secteur privé
6. Cadre de la fonction publique (catégorie A)
7. Technicien, agent de maîtrise, contremaître, catégorie B de la fonction publique
8. Employé, personnel de service
9. Ouvrier

Q35. (Si Q30=2 ou Q33=3 ou 4 ou 5 ou 7) **Quel est le secteur d'activité du chef de famille ?**

(une seule réponse possible)

1. Agriculture, chasse, exploitation forestière, pêche
2. Industries mécaniques, électroniques, chimiques, agro-alimentaires, production et distribution d'énergie et d'électricité, Imprimerie et autres industries
3. Bâtiments, Travaux publics
4. Commerce et distribution
5. Transport (terrestres, eau, aériens) et Télécommunications
6. Hôtellerie, restauration
7. Etude, conseil, services aux entreprises
8. Informatique
9. Banque, assurance, immobilier
10. Santé et action sociale
11. Arts, spectacles
12. Professions de l'information
13. Professions des Bibliothèques, musées et archives
14. Ecrivain
15. Métiers du livre
16. Enseignement du Primaire
17. Enseignement du Secondaire
18. Enseignement du Supérieur
19. Recherche
20. Administration publique
21. Services aux personnes (blanchisserie, coiffure, soins de beauté, pompes funèbres, activités thermales et de thalassothérapie…)
22. Autres services (Assainissement, voierie et gestion des déchets, services domestiques ….)

Q36. Quel est votre niveau d'études ?

1. Aucun diplôme / certificat d'études primaires
2. BEPC, Brevet des collèges
3. CAP, BEP
4. Baccalauréat, Capacité en droit
5. Diplôme universitaire de 1er cycle (Bac+2) : Deug, BTS, DUT ….
6. Diplôme universitaire de 2ème cycle (Bac+3 ou Bac+4) : Licence,Maîtrise
7. Diplôme universitaire de 3ème cycle (Bac+5) : DESS, DEA
8. Diplôme d'Ingénieur
9. Diplôme d'une Grande Ecole autre qu'une Ecole d'Ingénieur
10. Doctorat, Thèse

Q37. (Si Q33=3) **Quels sont les revenus annuels de votre foyer ?**

Moins de 15 245 euros (moins de 100 000 FF)
De 15 245 à moins de 30 490 euros (de 100 000 à moins de 200 000 FF)
De 30 490 à moins de 45 735 euros (de 200 000 à moins de 300 000 FF)
45 735 euros ou plus (300 000 FF ou plus)
(vous ne savez pas ou ne souhaitez pas répondre)

Q38. Résidez-vous …..
(une seule réponse possible)
1. En France métropolitaine
2. Dans les Dom-Tom
3. Dans un pays francophone
4. Dans un autre pays

Q39. (Si Q38=3 ou 4) **Le français est-il votre langue maternelle ?**
1. Oui
2. Non

Q40. (Si Q38=1) **Dans quel département vivez-vous ?**
(question quantité)

Q41. (Si Q38=1) **Habitez-vous…..**
1. Paris ou agglomération parisienne
2. Dans une agglomération de **plus de 200 000** habitants : *Angers, Avignon, Béthune, Brest, Bordeaux, Clermont-Ferrand, Dijon, Douai-Lens, Grenoble, Le Havre, Lille, Lyon, Marseille/Aix-en-Provence, Metz, Montpellier, Mulhouse, Nancy, Nantes, Nice, Orléans, Rennes, Reims, Rouen, St-Etienne, Strasbourg, Toulon, Toulouse, Tours, Valenciennes*
3. Dans une agglomération de **100 000 à 200 000** habitants : Annecy, *Amiens, Angoulème, Genève/Annemasse, Bayonne/Biarritz, Besançon, Caen, Calais, Chambéry, Dunkerque, Limoges, Le Mans Lorient, Montbéliard, Nîmes, Pau, Perpignan, Poitiers, La Rochelle, St-Nazaire, Thionville, Troyes, Valence*
4. Dans une agglomération de **moins de 100 000 habitants**
5. Dans une **commune rurale** (moins de 2000 habitants)

Q42. (Si Q38=2) **Habitez-vous…..**
Items 1 ET 2 : A Conserver mais à Filtrer systématiquement
1. Paris ou agglomération parisienne
2. Dans une agglomération de plus de 200 000 habitants
3. Dans une agglomération de plus de **100 000** habitants : Pointe-à-Pitre/Les Abymes (Guadeloupe), Saint-Denis ou Saint-Pierre (la Réunion), Fort de France (Martinique)
4. Dans une autre agglomération
5. Dans une **commune rurale** (moins de 2000 habitants)

PARTICIPATION AU PANEL D'UTILISATEURS GALLICA

Conditions pour poser la Question Q43 sinon fin du questionnaire

-> CONNEXION INTERNET A DOMICILE/SUR LIEU DE TRAVAIL (Q15= 1, 2, 3 ou 4)
ET
-> TRAVAILLER SUR UN SYSTEME D'EXPLOITATION WINDOWS (Q25= 1)
ET
-> DOMICILIES EN France métropolitaine et DOM-TOM (Q38= 1 ou 2)

Q43. La Bibliothèque nationale de France souhaite constituer, en partenariat avec France Télécom, un panel d'utilisateurs du site « gallica.bnf.fr ». France Télécom fournira un logiciel SECURISE à installer sur votre ordinateur, permettant à la BnF de suivre la manière dont vous utilisez le site Gallica et d'autres sites comparables : les types de recherche que vous effectuez, vos parcours de consultation sur le site. Cette étude est destinée à améliorer le contenu et les performances de Gallica en fonction de vos usages et de vos attentes. Les informations recueillies resteront strictement confidentielles.

Si vous souhaitez des informations complémentaires concernant cette étude, n'hésitez pas à nous contacter à l'adresse suivante: bibusages@voila.fr .

1. Je souhaite faire partie de ce panel
2. Je ne souhaite pas faire partie de ce panel

Q44. (Si Q43=1 ; sinon, fin du questionnaire) Nous vous remercions de votre participation à ce panel d'utilisateurs Gallica. Merci de nous laisser vos coordonnées afin de vous contacter d'ici quelques semaines pour la mise en place du logiciel. Elles resteront confidentielles, conformément à la loi informatique et libertés.

Votre Nom :
Prénom :

Au moins l'un des 2 numéros suivants pour vous contacter :

Téléphone personnel :
Téléphone professionnel :

Adresse e-mail** (information obligatoire) **:
Merci d'avoir accepté de participer à cette étude.

4.2 Grille d'entretiens BibUsages

Les entretiens semi-directifs menés dans le cadre de BibUsages, dont nous reproduisons ci-dessous la grille, sont élaborés autour de trois axes :

1. Une première partie centrée sur l'utilisation générale d'Internet permet de mieux cibler le profil général de l'internaute dans ses pratiques : durée, motivations, contexte de l'utilisation, modalités des recherches et traitement de l'information. En outre, cela permet d'avoir des renseignements sur les usages hors Web (*chat*, mail, forums, *peer-to-peer*…) que la sonde Audinet, dans la version utilisée pour cette étude, n'enregistre pas.
2. La deuxième partie de la grille se concentre sur l'usage des bibliothèques électroniques et de Gallica en particulier. Il s'agit de connaître les contextes d'utilisation des fonds numériques, les méthodes de recherche et les modalités

de traitement de l'information. Dans la discussion, on cherche également à pressentir des difficultés et à obtenir des propositions d'amélioration dans la conception du site Gallica.

3. La troisième partie de l'entretien se concentre sur les pratiques « off-line » : il s'agit ici de relier l'utilisation des bibliothèques électroniques à celle des bibliothèques classiques et, plus largement, aux pratiques de lecture et aux pratiques culturelles des interviewés.

Pour chaque entretien, une fiche descriptive du panéliste a été élaborée à partir de ses réponses au questionnaire présenté sur Gallica, et des statistiques de son trafic Internet déjà recueilli *via* Audinet.

Utilisation générale d'Internet

Usages et Fréquence

Sur votre utilisation générale d'Internet…
Bref rappel sur l'équipement

Comment êtes-vous arrivé à Internet ?
- Date de première utilisation
- Dans quel cadre ?
- Dans quel objectif ?
- Modalités d'apprentissage

Vous servez-vous beaucoup d'Internet ? Pour y faire quoi ?
- Fréquence d'utilisation – régularité d'utilisation – usages courants, besoins ponctuels – partage mail/Web/chat/jeu/autres

- Distinguer les différents type d'usages personnels et professionnels ?
  Quelle fréquence ?
  Interactivité entre les deux usages (pro et personnel) ?

Comment utilisez-vous le mail et les autres moyens de communication ?
- Nb adresses – mail classique, WebMail
- listes de discussion –
- fréquence de consultation
- nombre de correspondants
- chat, forums

Avez-vous un site Web / participez-vous à la conception d'un site ?
- Date de création du site – Contenu et but du site – Fréquence des modifications
- Evolution du site (contenu, présentation, public visé) –
  Connaissance de la fréquentation du site – Motivation et objectifs pour la création du site

Modalités des usages

Qu'allez vous voir sur le Web ?
Informations – Utilisations particulières (achat en ligne, réservation…)
téléchargement de programmes

Avez-vous des sites privilégiés que vous visitez régulièrement ?
Fréquence et mode d'utilisation
Favoris (nombre, organisation thématiques…)

Y a-t-il des sites que vous avez fréquentés intensément dans des contextes précis ?
Evénements particuliers (préparation de vacances, recherche d'appart, programme de cinéma, événement particulier, actualités … etc)

Comment trouvez-vous de nouveaux sites ?
Utilisation des moteurs, des annuaires
Utilisation de liens de sites vers d'autres sites
Adresses recommandées par d'autres gens (ami, mailing list, …)

Quels moteurs de recherche utilisez-vous ?
Si vous en avez un moteur privilégié, pourquoi l'avez vous choisi ?
Si plusieurs, comment les utilisez-vous ?

Quelle méthode employez-vous pour effectuer votre recherche à partir d'un moteur ?
Réflexion préalable sur un mot clé, combinaison, affinage, requêtes successives…

Que faites-vous quand vous trouvez une page intéressante ?
Sauvegarde locale – bookmark – comment la retrouver – conseil du lien à d'autres gens
Modalités du traitement de l'information. (Téléchargement, impression…)

Site de la BNF

Nous allons maintenant en venir aux bibliothèques électroniques...
Vous connaissez sans doute le site de la BnF
Comment l'avez-vous connu ?

Quel est la motivation première de votre visite ?
- La curiosité : intérêt culturel, suivre l'actualité de la BnF, recherche au gré du surf sur les différents dossiers
- Intérêt professionnel : recherche déterminée (utilisation du catalogue, Gallica, dossier pédagogique)
- Intérêt personnel : recherche personnelle sur un sujet particulier/ préparation et réservation des documents pour votre prochaine visite à la BnF.

Quelle est votre fréquence d'utilisation du site de la BnF ?
- Fréquence en fonction des différents types de navigation.
- Quotidien, hebdomadaire, plusieurs fois par mois.

Quelle rubrique visitez-vous le plus ?
Catalogue en ligne, Collection, Informations pratiques, service aux lecteurs, Gallica, programme culturel,
dossier pédagogique …

Plus généralement, vous direz que vous êtes satisfait, plutôt satisfait ou peu satisfait du site ?
Navigation, lisibilité des pages, se repérer dans l'espace, code couleur.
Points positifs et points négatifs

Gallica

Comment avez-vous découvert Gallica ?
Date de découverte –
Mode : par un moteur, un annuaire, par quelqu'un, par une publicité, par hasard - Via le site de la BNF

De manière générale comment vous connectez-vous à Gallica ?
Il est dans vos bookmarks.

Vous passez par le site de la BnF.
Vous tapez directement l'adresse dans votre navigateur.
Variation de fréquences en fonction des différents besoins ponctuels (recherche particulière sur tel thème, tel auteur…)

Pourquoi passez-vous par le site de la BnF?
Parce qu'il est enregistré dans vos bookmarks
Vous ne connaissez pas l'adresse directe de Gallica
Parce qu'il correspond à une stratégie de recherche, du catalogue à Gallica ou vice et versa.

Quelle est la motivation première de votre visite ?
Vous avez une recherche précise à effectuer
Vous cherchez une documentation sur un domaine
La curiosité, vous vous laissez guider au fil des pages

Quel procédé de recherche utilisez-vous ?
Catalogue, dossiers thématiques …..

Modalités d'utilisation des documents de Gallica

Comment utilisez-vous les documents trouvés ?
Lecture en ligne, Téléchargement, Impression.
Pourquoi ?
Vous les stockez pour les lire ultérieurement
Vous les imprimez pour un meilleur confort de lecture
Vous recherchez une information précise que vous détectez à l'écran

Une fois téléchargés, comment utilisez-vous les documents ?
Impression, lecture, réutilisation d'extraits

Combien de temps les gardez-vous en mémoire ?

Feuilletez-vous les documents en lignes ?
A quoi correspond ce feuilletage ?
Vous ne trouvez pas l'information recherchée, vous recherchez une information précise, ce mode-là vous convient, le feuilletage permet d'évaluer l'importance du document pour votre recherche avant téléchargement ou impression.

Si vous les lisez en ligne, cette lecture vous paraît-elle satisfaisante ?

D'un point de vue ergonomique, la navigation sur Gallica vous paraît-elle aisée ?
Repérage dans l'espace, code couleur, facilité d'accès au document…

Bibliothèques numériques, bibliothèques classiques

Connaissez-vous d'autres bibliothèques numériques ?

Les utilisez-vous ?
Quelles en sont vos utilisations ?

Que vous apportent-elles ?

Quels autres sites avez-vous l'habitude de fréquenter pour vos recherches ? Pourquoi ?

Achetez-vous (quand c'est possible) les ouvrages que vous consultez sur bibliothèque numérique ?

Lisez-vous sur écran ?
Gardez-vous une préférence pour le support papier (dans quels cas) ? Quelles différences percevez-vous entre les deux supports ?

Quelles sont vos habitudes en matière de fréquentation des bibliothèques classiques ?
Fréquence de visite
Consultation vs. emprunt
Types d'ouvrages consultés/empruntés
Nombre et type de bibliothèques fréquentées (universitaires, municipales, BPI, BNF…)

Achetez-vous beaucoup de livres ? Quel type ?
Posséder vs. Consulter – Collectionneurs

Que vous apporte l'usage des bibliothèques numériques par rapport aux bibliothèques classiques ?
Le développement des bibliothèques numériques a-t-il changé votre pratique des bibliothèques classiques et vos pratiques professionnelles

Annexe 5

Programmation

Cette section présente quelques exemples notables de développements informatiques que nous menés, concernant la mise en forme des URL brutes, l'identification des sites à partir des url, et la reformulation des parcours sans les mouvements de *Back*. Nous décrivons ces modules en termes fonctionnels et sous forme de pseudo-code.

5.1 Découpage des URL

Objectif

Le module de découpage des URL vise, à partir d'une URL quelconque, à en identifier et en séparer les différents constituants, sachant qu'une URL correspond formellement au schéma :

```
[protocole]://[utilisateur]@[serveur]:[port]/
      [répertoire]/[fichier]?[arguments]#[ancre]
```

Les champs obligatoires sont *protocole* et *serveur* ; les autres champs peuvent apparaître ou non dans l'URL. Quelques exemples d'URL rencontrées dans les données de trafic dont nous disposons :

- `http://www.sncf.fr`
- `http://fr.search.yahoo.com/search/fr?o=1&zw=1&p=escargots+bourgogne&d=y&za=and&h=c&g=0&n=20`
- `http://194.51.10.18:8080/enoviewer/servlet/GetGeoData?rset=WNOAA2000&ps=3000.0&pq=50.62500000000001,16.875`
- `https://www.lbmicro.com:443/cgi-bin/emcgi?session=eyRCVCC0`
- `ftp://ftp.schneeberger.fr/schneeberger/Pub/dc2/dc2nc20a.txt`
- `ftp://mp3@007mp3.dyndns.org:21/%3D%3DFULL%20ALBUMS%20002%3D%3D/Tom%20Jones%20-%20Reload/`
- `aol://aol.prop/4344:3873.d1_res.35914016.591441539`

Nous souhaitons donc disposer d'un module qui, à partir d'une URL quelconque, renvoie la liste du contenu des sept champs : *protocole*, *utilisateur*, *serveur*, *port*, *répertoire*, *fichier*, *arguments* et *ancre*.

Réalisation

Le découpage des URL se base sur l'application d'expressions régulières pour la reconnaissance des champs qui la constituent *protocole*, *utilisateur*, *serveur*, *port*, *argument* et *ancre* ;

Le découpage des URL se base à la fois sur l'application d'expressions régulières pour la reconnaissance des différents champs, ainsi que de règles pour repérer certains cas complexes :

1. Un premier traitement isole les champs protocole, utilisateur, serveur et port du reste de l'URL, en se basant sur la première occurrence de '://' et le '/' suivant. Si aucun '/' ne suit le motif '://', aucune ressource n'est spécifiée sur le serveur.
2. Pour le reste de l'URL, toutes sortes de cas peuvent se présenter :
   a. format classique d'un nom de fichier, précédé ou non d'un répertoire, et suivi ou non de paramètres ou d'une ancre, du type : `/chemin/fichier.html` ou `/search.php?var=toto&var2=titi`.
   On se base sur le fait que le nom de fichier contient un point et qu'il est précédé d'un '/' ; les caractères '?' et '#' servent à identifier le passage de paramètres et l'appel à des ancres.
   b. la séparation entre le fichier et les arguments est matérialisée par un ';' dans le cas de fichiers de type 'jsp', contre un '?' dans le reste des cas, ce qui nécessite un traitement particulier.
   c. l'URL pointe vers un nom de répertoire seul : si celui-ci se termine par un '/', le cas est non ambigu. Sinon, on postule que l'absence de point dans la chaîne de caractères qui suit le dernier '/' implique qu'elle désigne un répertoire et non un fichier, par exemple `http://zor.org/LeoGetz`.
   Ce choix laisse de côté un cas particulier où, semble-t-il, un nom de fichier est suivi de paramètres comprenant des '/', par exemple :
   `www.genhit.com/popup.php/carine83/kayash`
   ou `www.qxl.com/cgi-bin/qxlhome.cgi/FR/QXL/PR/U1010`
   Dans ce cas particulier, il apparaît que ce sont les scripts `popup.php` et `qxlhome.cgi` qui sont appelés, la suite de l'url étant prise en charge par ces scripts. Ce cas étant numériquement rare, et attestant une configuration de serveur particulière, nous ne le traitons pas en tant que tel.

Le module opère également un travail de normalisation de la partie répertoire, qui est modifiée afin de terminer toujours par un '/', et de ne pas contenir plusieurs '/' consécutifs.

Pseudo-code

```
INPUT : $url

$proto REÇOIT chaîne vide    # protocole
$user REÇOIT chaîne vide     # nom d'utilisateur
$host REÇOIT chaîne vide     # serveur
$port REÇOIT chaîne vide     # numéro de port
$path REÇOIT chaîne vide     # chemin
$file REÇOIT chaîne vide     # fichier
$query REÇOIT chaîne vide    # arguments
$ref REÇOIT chaîne vide      # ancre
```

```
# Première extraction : protocole, utilisateur, serveur, port
$url VÉRIFIE
       /^(.+?):\/\/(([^@\/]+)\@)?([^\/]+?)(:([0-9]+))?(\/(.*))?$/
$proto REÇOIT $1
$user REÇOIT $3
$host REÇOIT $4
$port REÇOIT $6
$suite REÇOIT $8

# Ensuite, traitement du reste de l'URL contenu dans $suite


# Si l'url ne pointe pas vers une ressource non spéficiée
SI $suite VAUT-PAS vide :

  | # Si l'adresse contient une ancre, on se base sur le "#"
  | # pour faire le découpage, et on l'extrait de $suite
  | SI $suite VERIFIE /(.*)(\#.*)/ :
  |    $suite REÇOIT $1
  |    $ref REÇOIT $2
  | FIN SI.

  | # Si l'adresse pointe vers un répertoire explicite (avec
  | # un '/' à la fin, avec ou sans ancre
  | SI $suite VERIFIE /(.*\/)(#[^#]+)?/ :
  |    $path REÇOIT $1
  |    $ref REÇOIT $2
  |
  | # Si $suite contient un appel à un script jsp, on se
  | # base sur ".jsp;" pour faire le découpage
  | SINON-SI $suite VERIFIE /(.*\/)([^\/]+\.jsp)(;[^#]+)(#[^#]+)?/ :
  |    $path REÇOIT $1
  |    $file REÇOIT $2
  |    $query REÇOIT $3
  |    $ref REÇOIT $4
  |
  | # si l'adresse contient un fichier avec un '.'
  | SINON-SI $suite VÉRIFIE
  |    /(.*\/)([^\/.]+\.[^\/.#?]+)(\?[^#]*)?(#[^#]+)?/ :
  |    $path REÇOIT $1
  |    $file REÇOIT $2
  |    $query REÇOIT $2
  |    $ref REÇOIT $2
  |
  | FIN-SI.

  # normalisation du $path
  | SI $path VERIFIE-PAS /\/$/ :
  |    $path REÇOIT CONCATENER($path, '/')
  | FIN-SI
  $path VERIFIE-SUBSTITUE /\/{2,}/\//g

SINON :
   $path REÇOIT '/'
FIN-SI.

RENVOIE ($proto, $user, $host, $port, $path, $file, $query, $ref)
```

5.2 Identification des sites

Objectif

Le module d'identification des sites a pour objectif, à partir des différents champs qui constituent une URL, d'identifier le site auquel se rattache cette URL. La définition d'un site est loin d'être évidente : plusieurs critères de regroupement des pages peuvent être mis en avant – capitalistique, technique, auctorial (voir 2.2.2 « Traitement des URL » p. 57 pour une discussion de ce problème). D'un point de vue technique, réduire le site au nom du serveur inclus dans l'URL peut poser, selon les cas, un problème de réduction (par exemple l'agrégation de tous les sites personnels hébergés par Wanadoo sous `perso.wanadoo.fr`) ou d'éclatement (par exemple la scission des différents sites de `m6.fr` en `www.m6.fr`, `m6kid.m6.fr`, `bac2004.m6.fr`, etc.).

Pour répondre à ces problèmes, nous mettons en avant la notion de *site éditorial* : il s'agit d'identifier un site comme un ensemble de contenus dépendant du même auteur (individu, entreprise, institution, etc.), qui en a la responsabilité. Deux cas sont distingués :

- pour les sites personnels, on identifie ce qui correspond à la racine du site pour l'auteur et non pour l'hébergeur, par exemple `perso.wanadoo.fr/french.roads/` ;
- pour les autres sites, on identifie le nom de domaine tel qu'il peut être acheté auprès des centres d'enregistrement.

Étant donné que dans le cas de pages personnelles, le *site éditorial* peut inclure des éléments du chemin vers la ressource sur le serveur, le module d'identification du site éditorial d'une URL doit également procéder à un redécoupage du chemin, et calcule donc un *chemin éditorial*.

Réalisation

Le module se base sur le découpage des URL présenté ci-dessus, dont il mobilise les champs *serveur* et *répertoire*, ainsi que sur l'identification, au sein des URL, de celles relevant de sites personnels. Deux cas se présentent :

- sites personnels : chaque URL sur un site personnel est rattaché à un hébergeur ; pour chaque hébergeur, on a identifié la syntaxe utilisée pour la désignation de la racine. Quatre cas de *format de chemin* sont distingués :
  1. type *host* : le site personnel est présenté comme un sous-domaine de l'hébergeur, par exemple : `restaurefour.free.fr`. Dans ce cas, *site éditorial* et *chemin éditorial* valent *serveur* et *répertoire*.
  2. type *path* : le site personnel est présenté comme un sous-répertoire dans un domaine spécifique de l'hébergeur, par exemple : `perso.wanadoo.fr/french.roads/`. Dans ce cas, *site éditorial* est la concaténation de *serveur* et du nom du premier répertoire de *répertoire*, et *chemin éditorial* vaut *répertoire* moins le nom du premier répertoire de *répertoire*.
  3. type *geocities* : réservé à l'hébergeur Geocities, les site hébergés sont de la même syntaxe que le type *path*, soit comprennent un nombre variable de

noms de villes avant le répertoire identifiant le site éditorial, qui est sous la forme d'un identifiant numérique, par exemple : `www.geocities.com/Tokyo/Palace/7574/`. Dans ce deuxième cas, le *site éditorial* est la concaténation de *serveur* et, dans le champ *répertoire*, de la liste des noms de lieux *plus* un répertoire composé de chiffres uniquement, et *chemin éditorial* vaut le reste de *répertoire*.

- autres sites : on cherche à identifier le domaine enregistré auprès des centres d'enregistrement. Dans les domaines de premier niveau (TLD, Top Level Domaine), certains sous-domaines sont réservés (*.asso.fr*, *.co.uk*, etc.) : on identifie donc ces sous-domaines, pour retenir le domaine situé en dessous dans la hiérarchie, par exemple : `crimlangueso.asso.fr`. Pour les autres noms de serveurs, applique une règle de type `TLD-1` : on retient le domaine juste sous le domaine de premier niveau, par exemple : `01net.com`.

Pseudo-code

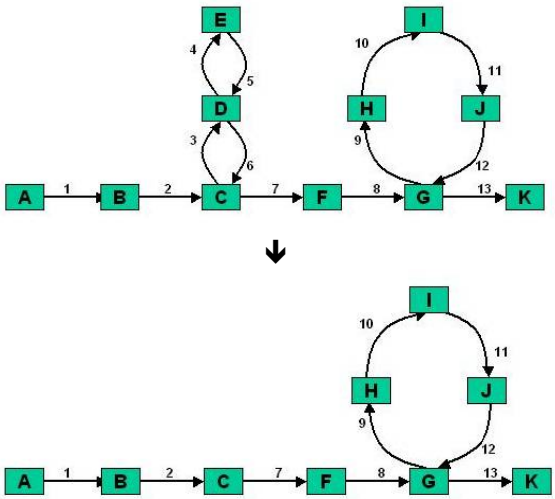
```
INPUT : $host           # le serveur
INPUT : $path           # le chemin vers la ressource (répertoires)
INPUT : $service        # le type de service de la page
INPUT : $format_path    # la syntaxe employée pour les sites personnels

RESSOURCE : @sous_domaines_reserves  # liste de TLD-2 réservés

$editorial_host REÇOIT chaîne vide   # le site éditorial
$editorial_path REÇOIT chaîne vide   # chemin éditorial

$host REÇOIT MINUSCULISE($host)

# Si c'est un site personnel
SI $service VAUT 'Page perso' :

   # Premier cas : format de type 'host'
   SI $format_path VAUT 'host' :
      $editorial_host REÇOIT $site ;
      $editorial_path vaut $path

   # Deuxième cas : format de type 'path'
   # On procède au découpage de path.
   SINON-SI $format_path VAUT 'path' :
      $path VERIFIE /^(\/[^\/]+)(\/.*)/ ;
      $editorialSite REÇOIT CONCATENER($site, $1)
      $editorialPath REÇOIT $2

   # Troisième cas : format de type 'geocities'
   SINON-SI $format_path VAUT 'geocities' :
      SI $path VERIFIE
           /^(\/(Area51\/|Athens\/|Augusta\/|Baja\/|BourbonStreet\/
           |Broadway\/|CapeCanaveral\/|CapitolHill\/|CollegePark\/
           |Colosseum\/|EnchantedForest\/|Eureka\/|FashionAvenue\/
           |Heartland\/|Hollywood\/|HotSprings\/|MadisonAvenue\/
           |MotorCity\/|NapaValley\/|Nashville\/|Paris\/|Pentagon\/
           |Petsburgh\/|Pipeline\/|RainForest\/|ResearchTriangle\/
           |RodeoDrive\/|SiliconValley\/|SunsetStrip\/|SoHo\/
           |SouthBeach\/|TelevisionCity\/|TheTropics\/|TimesSquare\/
           |Tokyo\/|Vienna\/|WallStreet\/|WestHollywood\/
           |Yosemite\/)([A-Z][A-Za-z]+\/)?[0-9]{4})(\/.*)/i) :
         $editorial_host REÇOIT CONCATENER($site, $1)
         $editorial_path REÇOIT $4
      SINON :
```

```
            $path VERIFIE /^(\/[^\/]+)(\/.*)/
            $ editorial_host REÇOIT CONCATENER($site, $1)
            $ editorial_path REÇOIT $2
         FIN-SI.
      FIN-SI.


   # Si ce n'est pas un site personnel
   SINON
      $editorial_path REÇOIT $path ;

      # Cas d'une adresse IP : on la conserve telle quelle
      SI $host VERIFIE /^(\d+\.\d+\.\d+\.\d+)$/) :
         $editorial_host REÇOIT $1
      SINON :

         # Si le serveur est au moins au niveau TLD-3
         SI $host VERIFIE /([^.]+)\.([^.]+\.[^.]+)$/
            $tld_moins_1 REÇOIT $2
            $tld_moins_2 REÇOIT CONCATENER ($1, $2)

            # Test de sous-domaine réservé
            SI $tld_moins_1 EXISTE-DANS @sous_domaines_reserves :
                $editorial_host REÇOIT $tld_moins_2
            SINON
                $editorial_host REÇOIT $tld_moins_1
            FIN-SI.

         # Serveur en TLD-1, local ou erroné
         SINON :
            $editorial_host REÇOIT $host
         FIN-SI.
      FIN-SI.
   RENVOIE ($editorial_host, $editorial_path)
```

5.3 Séquences de *back*

Objectif

Ce module a pour objectif d'identifier les séquences de *back* d'un parcours et de les isoler du reste de la session. Pour une session donnée, le module en produit une nouvelle représentation qui correspond au parcours sans les mouvements de *back*. Par exemple, une session "A→B→C→D→E→D→C→F" sera transformée en "A→B→C→F", la séquence "C→D→E→D→C" étant réduite à "C". La figure ci-dessous illustre ce travail de réécriture des sessions, et montre en particulier la différence entre les séquences de type *back* et les boucles, non exclues dans ce traitement.



Dans la mesure où l'on peut souhaiter disposer d'informations attachées aux pages vues (la durée de visite, le type de contenu, etc.) pour analyser ces éléments une fois les séquences de *back* supprimées, le programme doit être capable de conserver ces informations lors du traitement. En effet, si l'on ne garde pas trace de ce matériau au cours du traitement, il est impossible de le reconstituer par la suite, certaines pages pouvant apparaître plusieurs fois dans un parcours.

Réalisation

Une session est représentée, dans un tableau indexé, sous la forme d'une suite ordonnée de symboles, correspondant à la succession de l'accès aux pages ou aux sites : ce peut être les URL, les sites éditoriaux, leurs identifiants numériques, ou tout autre identifiant unique. Parallèlement, un autre tableau indexé prend en charge les informations disponibles sur les éléments du parcours ; les deux tableaux ont, pour un index donné, des éléments correspondants.

Tableau des éléments du parcours	Index commun	Tableau d'informations sur les éléments : durée (exemple)
page A	← 1 →	12 sec.
page B	← 2 →	5 sec.
page C	← 3 →	8 sec.
…	…	…
page F	← n →	9 sec.

Les deux tableaux sont traités conjointement, de manière à ce que la correspondance entre les deux soit conservée ; le tableau des éléments sert de référence pour identifier les *back*, le second est modifié lorsque des *back* sont repérés.

L'application examine un à un les éléments du parcours dans l'ordre de visite :

- l'élément est comparé à l'élément *n-2* :
    i. s'il est différent, on n'est pas dans un *back* ;

ii. s'il est identique, on est dans un mouvement de *back*, et on cherche à voir si ce mouvement se poursuit : on compare n+1 à n-3, n+2 à n-4, etc. :
   1. tant que la comparaison est vraie, les éléments vus sont dans un mouvement de *back*, on les parcourt un à un ;
   2. dès que la comparaison est fausse à l'indice *n+i*, on n'est plus dans un *back* :
      a. on remplace l'ensemble des pages du mouvement de *back* (de l'indice *n-2-i* à l'indice *n+i*)par l'élément de l'indice *n-2-i*, qui correspond à la visite de la page qui a initié la séquence de *back*.
      b. on incrémente le compteur de séquences de *back*.

- le processus continue jusqu'au dernier élément du parcours.

La suppression des mouvements de *back* opérée sur le tableau contenant les représentations des éléments du parcours est également opérée sur le tableau contenant les informations sur ces éléments ; ceci se fait sur la base des indices des tableaux, qui correspondent un à un.

Pseudo-code

```
INPUT : @pages
INPUT : @infos

$nbBoucles REÇOIT 0
@sb_elements REÇOIT vide
@sb_infos REÇOIT vide


SI INDICE-MAX(@pages) > 1 :

   $etat REÇOIT 'hors_boucle'

   AJOUTE-A(@sb_elements, $pages[0]) ;
   AJOUTE-A(@sb_elements, $pages[1]) ;

   AJOUTE-A(@sb_infos, $pages[0]) ;
   AJOUTE-A(@sb_infos, $pages[1]) ;


  POUR $i DE 2 A INDICE-MAX(@pages) :

    SI $etat VAUT 'hors_boucle' :

      SI $pages[$i] VAUT $sb_elements[INDICE-MAX(@sb_elements)-1] :
         $etat REÇOIT 'boucle'
         INCREMENTE $nb_boucles
         SUPPRIME DERNIER-ELEMENT(@sb_elements)
         SUPPRIME DERNIER-ELEMENT(@sb_infos)
      SINON :
         AJOUTE-A(@sb_elements, $pages[$i])
         AJOUTE-A(@sb_infos, $infos[$i])
      FIN-SI.

    SINON-SI $etat VAUT 'boucle' :

      SI $pages[$i] VAUT $sb_elements[INDICE-MAX(@sb_elements)-1]):
```

```
|  |  |  |     SUPPRIME DERNIER-ELEMENT(@sb_elements)
|  |  |  |     SUPPRIME DERNIER-ELEMENT(@sb_infos)
|  |  |  |  SINON :
|  |  |  |     $etat REÇOIT 'hors_boucle'
|  |  |  |     AJOUTE-A(@sb_elements, $pages[$i])
|  |  |  |     AJOUTE-A(@sb_infos, $infos[$i])
|  |  |  |  FIN-SI.
|  |  |  FIN-SI.
|  |  FIN-POUR.

|  SINON :
|     @sb_elements RECOIT @pages
|     @sb_infos RECOIT @infos
|  FIN-SI.

|  RENVOIE (@sb_elements, @sb_infos, $nb_boucles)
```

Glossaire

Ce glossaire contient les termes techniques et de spécialité les plus employés dans cette thèse. Il emprunte, pour certaines d'entre elles, les définitions proposées par www.themanualpage.org (© themanualpage.org) soumises à la licence GNU FDL.

ADRESSE IP. Adresse de 32 bits utilisée par le protocole IP pour identifier de manière unique les hôtes (machines) sur réseau IP.
L'adresse IP est généralement présentée sous forme décimale « pointée » du type '127.0.0.1'. Elle comporte une partie identifiant le réseau (network ID), et une partie identifiant le numéro d'équipement dans le réseau (host ID).
Voir aussi : TCP/IP.

ANNUAIRE WEB. Site Web proposant une sélection de sites classés par thème. Les annuaires Web généralistes (ex : Yahoo, Voila, Nomade) ont vocation à couvrir l'ensemble des contenus disponibles sur le Web, et à proposer des liens pertinents pour chaque thème représenté.

APPLET. Application Java invoquée par une page Web. Les *applets*, contrairement aux vrais programmes Java ne peuvent pas accéder aux ressources système locales, telles que des fichiers ou périphériques. Les *applets* ne peuvent également pas communiquer avec d'autres serveurs que ceux dont elles sont issues. L'intérêt d'une *applet* est d'avoir des pages Web dynamiques, animées et interactives, en utilisant toute la puissance du Java. Les *applets* sont portables et offrent un bon niveau de fiabilité et de sécurité.

ASPIRATEUR WEB. Logiciel permettant de faire une copie locale d'une page ou d'un site Web.

CENTRE-SERVEUR. Les méthodes d'analyse des usages d'Internet opposent les approches centrées-serveur, qui reposent sur l'analyse de données collectées sur les serveurs (serveur Web notamment) à l'approche centrée-utilisateur, où les données sont recueillies au niveau de l'utilisateur.
Voir aussi : *DONNEES DE TRAFIC*, *LOGS*.

CENTRE-UTILISATEUR. Voir *CENTRE-SERVEUR*.

CLIENT/SERVEUR. Modèle fonctionnel logiciel dans lequel plusieurs programmes autonomes communiquent entre eux par échange de messages.
Le modèle est à l'origine dissymétrique, c'est toujours le client qui fait appel aux services du serveur. En aucun cas celui-ci ne peut fournir des services de sa propre initiative.
Voir aussi : *SERVEUR*, *REQUETE*.

COHORTE. Ensemble d'individus étudiés sur une période de temps donnée. Une cohorte permet de suivre de manière longitudinale les comportements de la population observée ainsi que sa réaction à un ou plusieurs événements donnés.

DNS. *Domain Name System/Server*
Système d'annuaire distribué sur l'Internet qui contient principalement les noms et les adresses IP des stations. Il sert à faire la conversion nom de machine-adresse IP.
Exemple : l'adresse symbolique www.themanualpage.org est convertie en l'adresse IP

numérique *205.206.106.50*.

DOMAINE. Le domaine identifie un groupe d'ordinateurs hôtes ou de réseaux locaux qui, sous une même entité administrative, sont branchés sur le réseau Internet. Le nom des domaines se compose de sections séparées par des points, qui définissent une arborescence (de droite à gauche) : au sein du domaine *.fr*, on trouve les sous-domaines *.lemonde.fr*, *.asso.fr*, etc. ; au sein du domaine *.asso.fr*, chaque association dispose d'un sous-domaine, etc.
L'achat d'un nom de domaine revient à disposer d'une entité au niveau -1 ou -2 des domaines de premier niveau. L'acheteur gère ensuite l'ensemble des sous-domaines du nom de domaine qui lui appartient (ex : *sport.tf1.fr*, *info.tf1.fr* font partie du domaine *tf1.fr*, géré par TF1).
Lors d'une communication entre deux ordinateurs du réseau Internet, les noms des ordinateurs et des domaines sont traduits en adresses numériques par un serveur de noms de domaine (ou Domain Name Server ou DNS).

DONNEES DE TRAFIC. Données techniques basées sur l'enregistrement de la communication sur un réseau.
Dans le cas du trafic Internet, selon l'endroit du réseau où est fait cet enregistrement, les données peuvent centrée-utilisateur, centrées-serveur, ou recueillies à un endroit intermédiaire (routeurs, *proxy*, DSLAM, etc.).
Voir aussi : *CENTRE-SERVEUR*, *CENTRE-UTILISATEUR*, *LOGS*.

ECHANTILLON. Sous-ensemble caractéristique d'une population, dont l'analyse permet de décrire l'ensemble de la population.

FAI. Fournisseur d'Accès à Internet (ex : Wanadoo, Free, AOL, etc.).

FORUM. Ou *newsgroup*, *groupe de discussion*, *Usenet*.
Système de discussion écrite asynchrone, sorte de salon électronique d'accès libre où se tiennent les discussions relatives à un ou plusieurs thèmes définis lors de la création du groupe.
Généralement, le nom du forum renseigne sur son contenu. Par exemple, le nom *fr.comp.sys.windows* désigne un forum francophone (*fr*) traitant d'informatique (*comp* pour 'computer'), plus précisément des systèmes d'exploitation (*sys*) et plus spécifiquement du système Windows.
Il existe plusieurs types de forums, selon la façon dont les messages sont transmis : certains se trouvent sur le Web, d'autres sur l'Usenet et les serveur de news.

FRAME. Concept inventé par Netscape, consistant à diviser la fenêtre d'un navigateur Web en plusieurs sous-fenêtres, dans chacune desquelles on affiche une page Web différente (page HTML ou autre). Chaque *frame* possède sa propre URL, la position et le contenu des *frames* étant définis par une page HTML appelée *frameset*.

FTP. *File Transfer Protocol* (Protocole de Transfert de Fichiers)
Protocole orienté vers le transfert de fichiers, en envoi comme en réception, fonctionnant sur les protocoles TCP/IP.

HTML. *Hypertext Markup Language*
Langage utilisé pour écrire les pages Web d'Internet.
C'est une version simplifiée de la norme SGML (Standard Generalized Markup Language), langage de document structuré, avec liens hypertexte, utilisé en gestion documentaire. Il a été inventé dans les années 80 par Tim Berners-Lee qui cherchait à l'époque un moyen simple et efficace pour mettre à disposition sur réseau la document du CERN.

HTTP. *Hyper Text Transfer Protocol*
Protocole inventé par Tim Berners-Lee à la même époque que le HTML et spécialement conçu pour accéder aux documents HTML. Par extension, ce protocole peut être utilisé pour accéder à presque tout type de ressource Web *via* Internet.

INTERNET. Réseau de portée mondiale interconnectant des centaines de réseaux spécifiques et auquel sont reliés quelques centaines de millions d'utilisateurs individuels et professionnels.
Ce réseau est le support de multiples activités et protocoles : consultation de sites Web (HTTP), messagerie (POP/SMTP), *peer-to-peer*, *chat*, jeux, etc.

IRC. *Internet Relay Chat*
Protocole permettant de dialoguer en mode texte en direct avec plusieurs personnes.

LOGS. Enregistrement technique de traces d'activité. Dans le cas des serveurs HTTP, les fichiers de *logs* conservent la liste de tous les accès aux ressources du serveur.

NAVIGATEUR. Logiciel de navigation sur le Web. Formellement, il s'agit d'un client pour le protocole HTTP, qui a comme fonctionnalité centrale l'interprétation et la visualisation en local ou à distance du langage HTML.
C'est en grande partie la réalisation du premier navigateur Web graphique Mosaic par un étudiant américain de NCSA (National Center for Supercomputing Applications, Université d'Illinois) nommé Marc Andreessen et la commercialisation des logiciels qui s'en sont inspiré, qui explique l'explosion de l'Internet à au milieu des années 90. À partir de cet instant, le navigateur devient de plus en plus le client Internet universel.
Les navigateurs possèdent en outre la faculté d'exécuter du code javascript et des *applets* Java. D'autres types d'applications peuvent être exécutées grâce à des *plugins* (composants logiciels additionnels).
Parmi les navigateur les plus connus, on retrouve Internet Explorer, Netscape Navigator ou Mozilla/Firefox.

NTIC. Nouvelles Technologies de l'Information et de la Communication.

PAGE WEB. Unité ergonomique élémentaire d'un site Web, désignée par une URL unique.
Une page contient souvent plusieurs éléments (images, *frames*, etc.), qui occasionnent chacun une requête auprès d'un serveur Web, et qui sont assemblés par le navigateur.
Une page peut être dite statique (le serveur Web renvoie le contenu d'un fichier) ou dynamique (le serveur Web renvoie le résultat d'un traitement particulier).
Voir aussi : *FRAME*, *SITE*, *SERVEUR*.

PANEL. Échantillon de personnes représentatif d'une population. Suivi dans le temps, un panel est corrigé régulièrement pour conserver sa représentativité.

POP. *Post Office Protocol*
Protocole permettant l'accès aux messages E.Mail se trouvant sur un serveur de messagerie ; il est utilisé conjointement avec SMTP(envoi de messages).
Voir aussi : *SMTP*.

PROTOCOLE. Description formelle de règles et de conventions à suivre dans un échange d'informations. Les protocoles peuvent définir ces échanges au niveau des couches Internet (IP), transport (TCP, UDP, etc.) ou applicatives (HTTP, FTP, Telnet, etc.).
Voir aussi : *HTTP*, *FTP*, *POP*, *SMTP*, *IRC*.

REQUETE. Dans un modèle client-serveur, envoi d'une instruction d'un client vers un serveur.
Dans le cas des serveurs Web, une requête HTTP demande à un serveur l'envoi du contenu d'un document (ex. : une page HTML) ou du résultat de l'exécution d'un traitement par le serveur (ex. : rechercher dans une base de données).

Voir aussi : *CLIENT/SERVEUR*, *SERVEUR*.

SERVEUR. Ressource informatique (machine ou programme) capable de délivrer une information où d'effectuer un traitement à la requête d'autres équipements.
Voir aussi : *CLIENT-SERVEUR*, *REQUETE*.

SESSION. Période d'activité cohérente d'un utilisateur sur Internet.
À partir des données de trafic centrées-utilisateur, on identifie les sessions sur la base d'une période d'inactivité (aucune trace d'activité enregistrée) de plus de trente minutes.

SITE WEB. Ensemble de pages et de services Web sous la responsabilité d'une même entité éditoriale (individu, société, organisme, etc.).
Il est identifié par un nom de domaine, par exemple *tf1.fr*. La plupart du temps, le site correspond à l'adresse du serveur Web sur le domaine spécifié (ex : *www.globz.net*), mais il peut être réparti sur plusieurs (*sport.tf1.fr*, *info.tf1.fr*, etc. pour le domaine *tf1.fr*). Dans le cas de sites personnels, le site, en tant qu'entité éditoriale, correspondre à un sous-domaine (comme chez Free, ex : *restaurefour.free.fr*) ou a un sous-répertoire sur un domaine dédié (comme chez Wanadoo, ex : *perso.wanadoo.fr/4pat/*).

SMTP. *Simple Mail Transfer Protocol*
Protocole de messagerie permettant d'envoyer des mails à partir d'un logiciel client vers un serveur de messagerie. Il est utilisé conjointement avec POP (réception des messages).
Voir aussi : *POP*.

TCP/IP. *Transmission Control Protocol / Internet Protocol*
Les deux protocoles de communication qui forment les fondements de l'Internet. TCP assure le transfert des données, IP l'adressage.
Voir aussi : *ADRESSE IP*.

TIC. Technologies de l'Information et de la Communication.

URL. *Uniform Resource Locator*
Identifiant de ressource à travers le réseau Internet permettant la détermination du protocole à utiliser pour transférer le document, l'adresse du serveur sur lequel se trouve le document, et le chemin d'accès du document sur le serveur.
La structure d'un URL est la suivante :
`[protocole]://[Nom du serveur][:numéro de port]/[chemin d'accès]`

W3C. *World Wide Web Consortium*
Consortium crée début 1995 dont le principal objectif est la mise au point de normes et de protocoles ouverts et libres pour le Web, dans un souci d'interopérabilité maximale. Il est géré conjointement par le MIT aux États-Unis, l'INRIA en France et l'université Keio au Japon. Son directeur est Tim Berners-Lee, père du Web.

WEB. Ou *World-Wide-Web* : littéralement « toile d'araignée mondiale ». Système d'information réparti, basé sur des documents en hypertexte (au format HTML). Service d'informations sur le réseau Internet, crée au CERN (Centre Européen de la Recherche Nucléaire) à Genève en 1993 par Tim Berners-Lee, et mettant à la disposition des utilisateurs un ensemble distribué de documents multimédias composites reliés entre eux par des liens hypertextes. Chaque document est identifié par une adresse appelée URL. C'est ce service qui explique en grande partie le succès récent d'Internet, et son utilisation par le grand public. Les sites et les documents sont accessibles et visualisables grâce à des logiciels clients appelés navigateurs.
Voir aussi : *NAVIGATEUR*, *HTTP*, *URL*.

WEBCHAT. Accès à des services de *chat* à l'aide d'un navigateur Web, au lieu d'un logiciel spécifique.

WEBMAIL. Accès à des services de messagerie à l'aide d'un navigateur Web, au lieu d'un logiciel spécifique. Certains services de WebMail permettent d'accéder à un compte de messagerie « classique » (c'est le cas pour la plupart des FAI), d'autres n'existent que sur le Web.

WEB MINING. Champ de recherche apparu dans les années 90 se centrant sur l'analyse et la fouille de données relatives au Web. Il se divise en trois catégories : analyse de la structure du Web (*Web Structure Mining*), fouille de son contenu (*Web Data Mining*) et étude de ses usages (*Web Usage Mining*).

WEB USAGE MINING. Champ de recherche apparu dans les années 90, dont l'objectif est l'analyse des usages du Web. Il se concentre principalement sur l'étude soit centrée-serveur, soit centrée-utilisateur.
Voir aussi : *WEB MINING*, *CENTRE-UTILISATEUR*, *CENTRE-SERVEUR*.

WWW. Voir *WEB.*
Placé au début d'une URL : convention désignant l'adresse d'un serveur HTTP sur un domaine donné.