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).

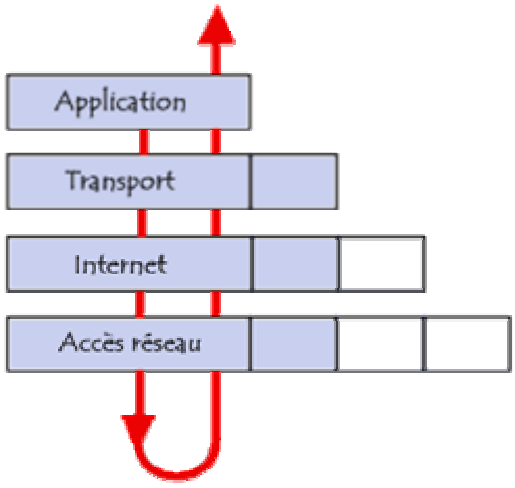
dire qu'il permet à deux machines qui communiquent de contrôler l'état de la transmission.

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

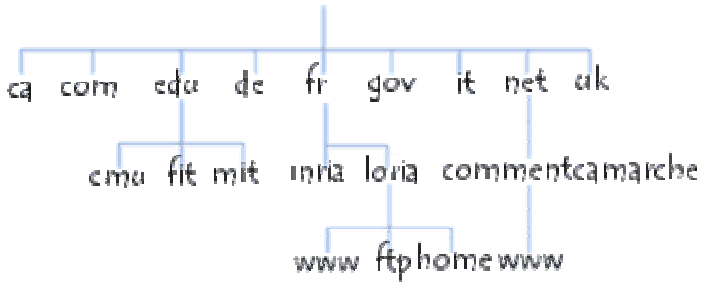
machine réceptrice, lors du passage dans chaque couche, l'en-tête est lu, puis supprimé.

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.

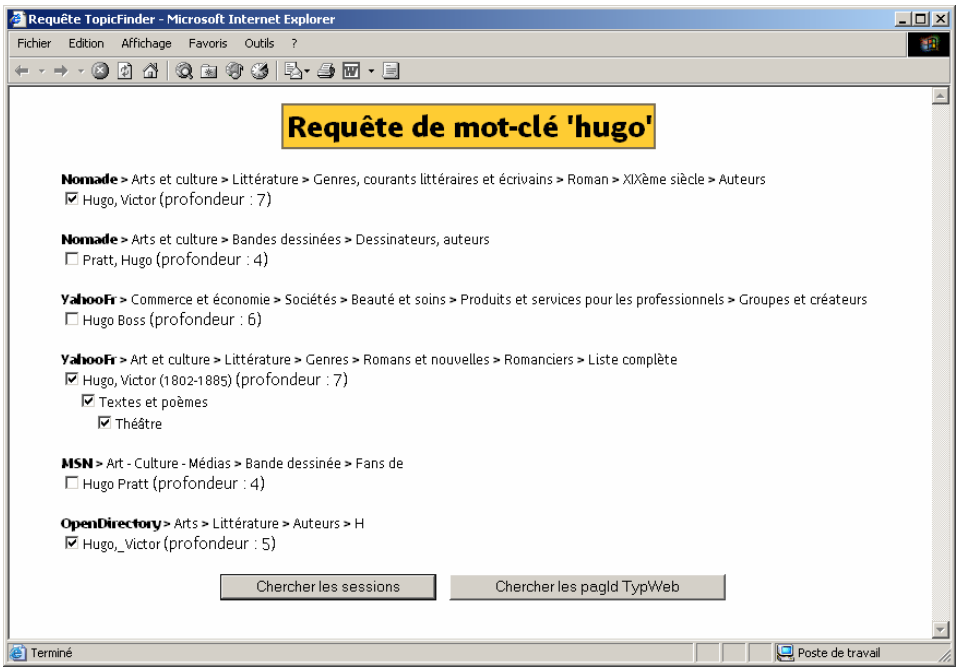


Figure 3.1. Interface de validation des catégories d'annuaires de TopicFinder

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

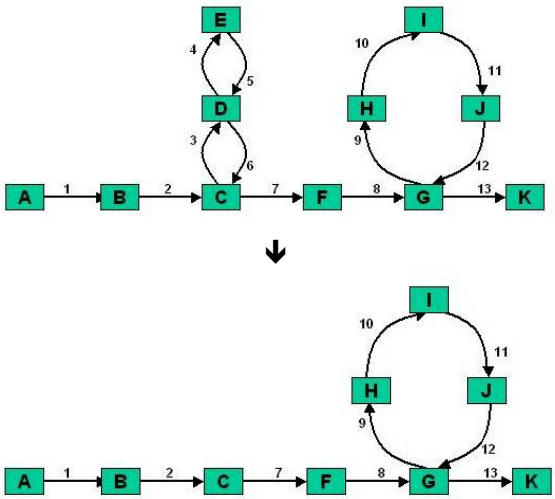
```
INPUT : $host           # le serveur
INPUT : $path           # le chemin vers la ressource (répertoires)
INPUT : $service        # le type de service de la page
INPUT : $format_path    # la syntaxe employée pour les sites personnels

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)
```