Chapitre 2

Préparation et fouille des données

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

2.1 Données de trafic « centrées-utilisateur »

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

2.1.1 Technologies de recueil de données

Continuité ergonomique, ruptures techniques

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

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

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

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

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

---

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

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

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

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

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

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

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

Plusieurs choix technologiques possibles

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

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

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

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

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

---

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

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

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

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

Technologies de recueil de données

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

---

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

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

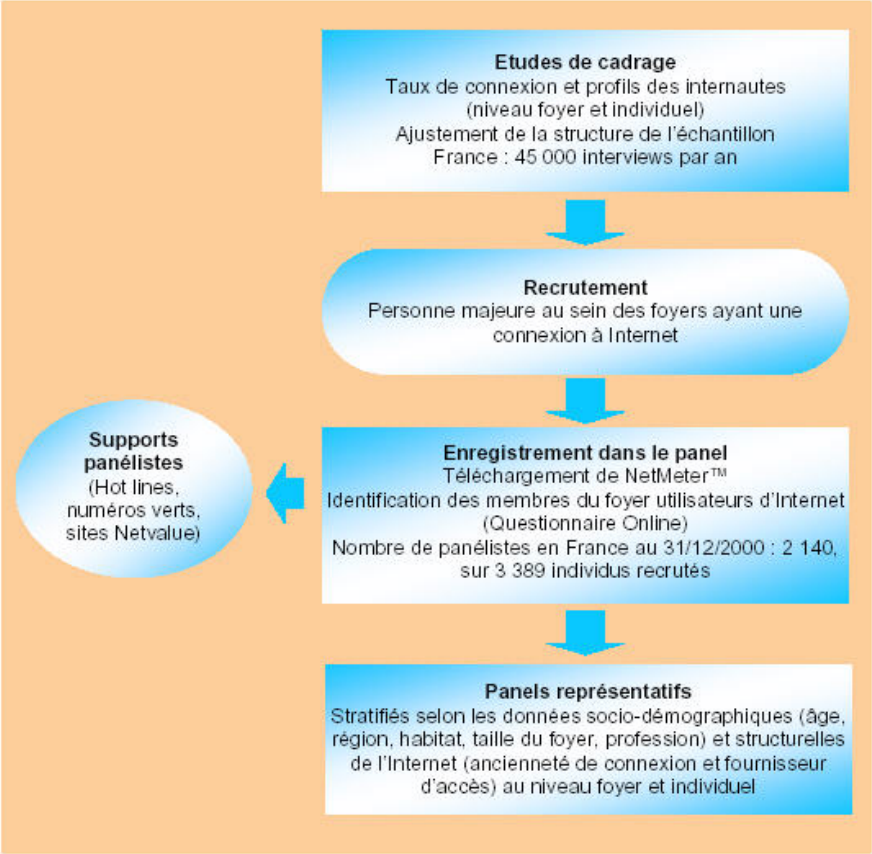

Figure 2.1. Constitution et le suivi du panel NetValue (source : NetValue, 2002)

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

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

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

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

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

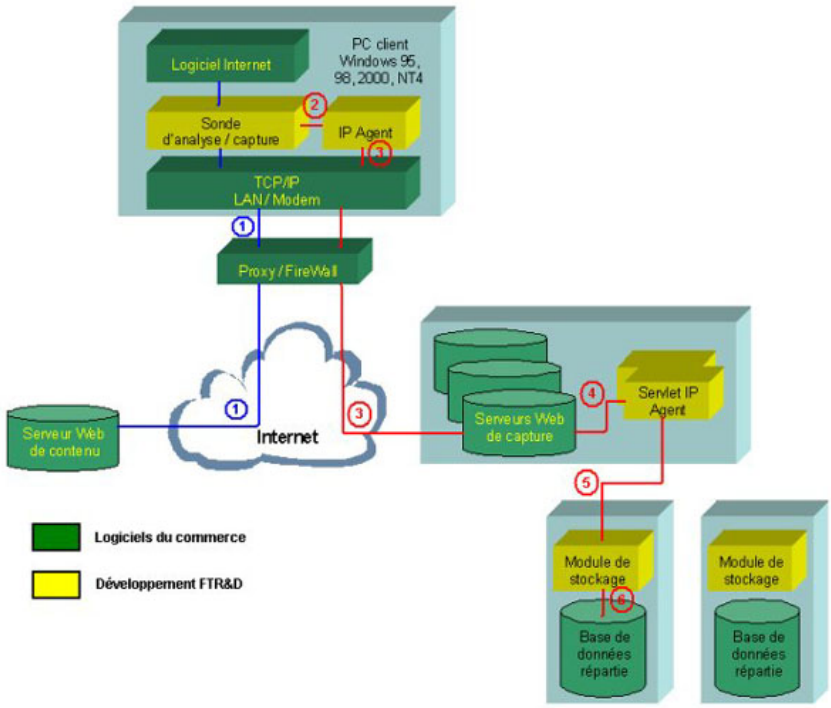

Figure 2.2. Architecture logicielle d'Audinet (source : L. Rabret, FTR&D)

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

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

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

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

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

2.1.2 Format des données

Informations recueillies

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

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

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

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

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

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

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

Figure 2.3. Extrait de données de trafic Web

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

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

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

Formalisme pour le stockage

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

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

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

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

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

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

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

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

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

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

2.2 Formatage des données pour l'analyse de trafic

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

2.2.1 Identifier les sessions

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

---

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

Sessions multi-protocoles

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

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

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

Figure 2.4. Exemple de session Internet multiprotocoles

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

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

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

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

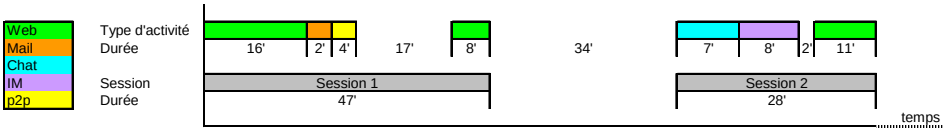

Figure 2.5. Identification des sessions sur la base de 30 minutes d'inactivité

Limitations

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

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

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

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

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

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

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

2.2.2 Traitement des URL

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

Découpage des URL

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

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

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

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

Plus précisément, ces éléments sont :

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

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

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

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

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

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

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

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

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

Qu'est-ce qu'un site ?

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

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

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

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

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

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

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

Identifier les « sites éditoriaux »

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

---

[1] Observé en mars 2002.

[2] Observé en mars 2002.

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

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

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

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

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

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

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

---

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

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

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

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

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

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

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

Tableau 2.1. Exemples de regroupement en sites éditoriaux

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

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

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

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

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

2.2.3 Recomposer les pages

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

La page Web, unité ergonomique inaccessible

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

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

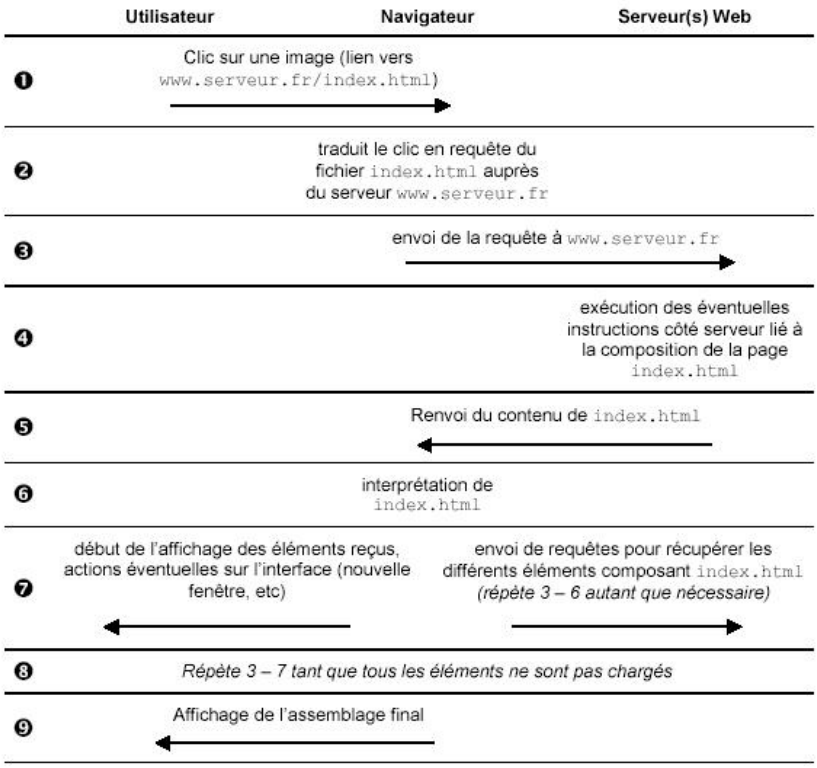

Figure 2.6. Requête d'une page auprès d'un serveur Web

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

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

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

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

Solutions partielles

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

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

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

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

---

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

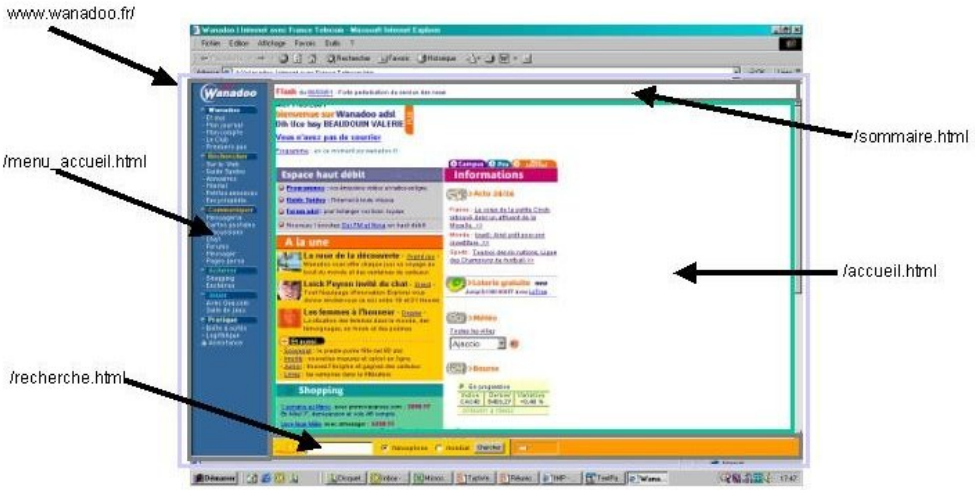

Figure 2.8. Les quatre frames *composant la page d'accueil de Wanadoo (2002)*

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

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

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

Conclusion

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

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

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