La mesure d'audience sur Internet : Terminologie, technologies et méthodologie

Y. Costes

Cahier n°278
Octobre 1999

Yseulys Costes
Allocataire de Recherche, DMSP, Université Paris Dauphine
17, Boulevard de Rochechouart
75009 Paris
E-mail : yseulys@filnet.fr

LA MESURE D'AUDIENCE SUR INTERNET : terminologie, technologies et méthodologies

Résumé :

Cet article, après avoir présenté la terminologie de la mesure d'audience sur Internet, fait un état des lieux des technologies et méthodologies utilisées. Il analyse également les limites des solutions logicielles et des services disponibles et aborde les problèmes encore irrésolus que sont la qualification et la certification de l'audience sur Internet.

Mots-clés : Mesure d'audience, Caches, Certification, Fichiers de logs, Internet, Terminologie

Remerciements : L'auteur remercie les personnes qui ont relu cet article pour leurs commentaires.

AUDIENCE MEASUREMENT ON THE INTERNET: Terminology, Techniques and Methodologies

Summary:

This article provides a taxonomy of audience measurement on the Internet and an inventory of the techniques and methodologies used. It also analyses the limits of the available measurement software and services and the unsolved problems of qualitative data collection and audience certification on the Internet.

Key words : Audience measurement, Proxies, Cetification, Logs files, Internet, Terminology

Introduction

Internet présente de multiples particularités en terme de mesure d'audience. C'est un média qualifié de traçable, même si cette traçabilité comporte de nombreuses limites, à la fois mesuré et outil de mesure, et chaque éditeur de contenu, support ou non de publicité, peut mesurer son audience par ses propres moyens.

L'objectif de cet article est de faire un état des lieux des technologies et des méthodologies utilisées pour mesurer l'audience des sites ainsi que des campagnes publicitaires sur Internet. En effet, sur ce média, les données de mesure d'audience sont utilisées à la fois dans un contexte éditorial pour mesurer l'audience des sites et dans un contexte publicitaire pour mesurer l'audience des campagnes menées sur le réseau (Costes, 1998b). Les limites des technologies et des méthodologies ainsi que des solutions logicielles et des services disponibles sur le marché français seront également étudiées.

Mais avant tout, il est nécessaire de savoir nommer ce que l'on prétend mesurer. Le problème de la terminologie, qui sera abordé dans la première partie de cet article, est un problème récurrent quel que soit le média. Mais Internet étant un outil que l'on peut qualifier de "transnational" (Baumard et Forgues, 1989), ce problème en est encore accentué puisque l'harmonisation de la terminologie ne doit pas seulement se faire à l'intérieur de chaque pays mais également à un niveau international.

L'intérêt de cet l'article réside en premier lieu dans la nouveauté du sujet. Les problématiques liées à la mesure d'audience sur Internet sont apparues quelques mois après l'arrivée de la publicité sur Internet en 1994 (Onnein-Bonnefoy, 1997) et très peu d'articles ont encore été publiés sur ce sujet. Faire un état des lieux mettant à jour les limites des technologies et des méthodologies mises en oeuvre permet de dégager de nombreuses pistes de recherches restant encore à explorer. De plus, cet article a pour objectif de permettre aux personnes conduisant des recherches ayant trait à des problématiques marketing sur Internet de se servir des particularités méthodologiques de ce média en terme de récupération des données d'audience et comportementales. Cet état des lieux mettant en lumière des limites méthodologiques et technologiques, il devrait également permettre d'évaluer plus précisément la qualité des données de mesure d'audience à traiter dans le cadre d'une recherche.

La terminologie de la mesure d'audience sur Internet

Les problèmes d'harmonisation de la terminologie prennent, sur Internet, une dimension internationale pour les raisons suivantes :

- dans un contexte publicitaire, l'achat d'espace se fait sur des sites dont la localisation géographique réelle n'est pas limitée à un seul pays; des données de mesure d'audience de sites localisés dans différents pays doivent donc pouvoir être comparées. La réalité du marché français modère cependant la portée de cette remarque. En effet, les annonceurs français investissent encore très majoritairement sur des supports nationaux car la cible de leur campagne est encore bien souvent exclusivement hexagonale.

- des outils de mesure d'audience utilisés par exemple par certains sites localisés en France sont développés aux États-Unis et proposent donc les résultats en anglais. Les sites localisés dans un même pays peuvent donc, en fonction de l'outil de mesure d'audience qu'ils ont choisi d'utiliser, publier des résultats dans des langues différentes.

Le manque d'harmonisation de la terminologie est un frein à la fois au développement du marché de la publicité sur Internet, comme le souligne Rich Le Furgy, Président de l'IAB[1] et Vice Président de la publicité du groupe Buena Vista (Ginburg, 1998), et à la viabilité à long terme du modèle publicitaire sur Internet (Murphy, 1996). En effet, pour pouvoir être utilisées de façon pertinente, les données de mesure d'audience doivent être comparables. Il est donc "nécessaire de normaliser la terminologie utilisée, afin de comparer des indices rigoureusement semblables" (Reboul et Xardel, 1997). Les acteurs du marché, conscients de la nécessité d'une telle harmonisation, se sont réunis en collèges ou groupes de travail au sein d'associations professionnelles. Aux États-Unis, la réflexion a lieu au sein d'un groupe de travail de l'IAB intitulé *Media Measurement*, tandis qu'en France un collège Internet créé au sein du CESP[2] travaille en collaboration avec l'IAB France qui lui apporte une perspective internationale sur cette question.

Les définitions proposées dans le Tableau 1 résultent de la confrontation de la terminologie française élaborée au sein du collège Internet du CESP et publiée en septembre 1997, et de la terminologie américaine élaborée au sein du groupe de travail *Media Measurement* de l'IAB.

L'ordre de présentation reprend celui adoptée par le CESP car il a pour principal avantage de mettre en évidence les contextes d'utilisation des indicateurs d'audience présentés :

- indicateurs d'audience éditoriale
- indicateurs d'audience publicitaire
- indicateurs d'efficacité publicitaire (bilan de campagne).

---

[1] L'IAB (Internet Adverstising Bureau) est une association Interprofessionnelle créée aux États-Unis en 1996 et qui regroupe aujourd'hui plus de 300 acteurs du marché de la publicité sur Internet (http://www.iab.net). Des IAB locaux s'ouvrent en Europe, et notamment en France (http://www.iabfrance.com), Pays-Bas, Angleterre et Allemagne

[2] Le CESP (Centre d'Étude de Supports de Publicité) est une association interprofessionnelle qui regroupe les acteurs du marché publicitaire concernés par l'étude de l'audience des médias : annonceurs, agences et conseils en communication, centrales d'achat d'espace, médias et régies publicitaires.

Tableau 1. - La terminologie de la mesure d'audience sur Internet

Indicateur	Concept	Mesure
Indicateurs d'audience éditoriale		
Pages vues sur site	Pages vues par les visiteurs directement sur le serveur du site.	Nombre de fois où une page est téléchargée et comptabilisée dans les fichiers de logs du serveur du site étudié.
Pages vues hors site	Pages vues par les visiteurs sur les caches des serveurs proxies ou des navigateurs. C'est un concept qui se rapproche de celui des occasions de voir.	La mesure des pages vues hors site pose des problèmes techniques dont des solutions (insertion d'une image dynamique, applet Java...) sont exposées dans cet article.
Visite	Ensemble de pages vues sur site au cours d'une même session.	Une absence de consultation de nouvelles pages sur le site étudié dans un délai de 30 minutes vaut pour fin de la visite. La mesure du nombre de visites n'est pas triviale car il est difficile de déterminer si les pages vues l'ont été par le même individu ou non, sachant que les requêtes sont gérées au "fil de l'eau", c'est à dire lorsqu'elles arrivent au serveur.
Visiteur	Individu qui consulte un même site au cours d'une période définie en tenant compte de la déduplication du nombre de visites.	Mesurer le nombre total de visiteurs suppose que l'utilisateur respecte une procédure d'identification.
Nombre de pages vues par	Nombre moyen de pages vues	Nombre de pages

visite	par visite sur un site et pour une période définie.	vues/Nombre de visites pour une période définie.
Origine géographique des consultations	Lieu physique réel de connexion des utilisateurs d'un site.	L'origine géographique réelle est difficile à mesurer car les adresses des machines-hôtes finissant par exemple en .com, .net ou .edu ne renseignent pas sur l'origine géographique réelle. Aucune solution réellement efficace n'a encore été trouvée à ce problème.

Indicateurs d'audience publicitaire		
Pages avec publicité vues sur site (PAP)	Pages vues sur lesquelles figurent l'offre de l'annonceur (bandeau publicitaire, objet, icône).	Nombre de fois où une page avec publicité est téléchargée et comptabilisée dans les fichiers de logs du serveur du site étudié. La non comptabilisation des pages vues hors site aboutit à sous estimer le nombre réel de pages avec publicité vues.
Coût au mille pages avec publicité vues sur site (CPM)	Coût d'achat de l'espace publicitaire d'un site ramené à une base de mille pages avec publicité vues sur site.	1000*Côut d'achat de l'espace/Nombre de pages avec publicité vues sur site.

Indicateurs d'efficacité publicitaire (Bilan de campagne)		
Nombre de clics constatés	Nombre de fois où les visiteurs auront cliqué sur la publicité.	Nombre de demandes de transfert sur l'adresse de la page à laquelle renvoie la publicité inscrites dans les fichiers de logs.
Nombre de pages avec publicité constatées	Nombre de pages avec publicité vues sur le site pendant une période donnée pour un annonceur donné.	La mesure est identique à celle du nombre de pages avec publicité vues sur site, mais pour un annonceur donné.
Taux de clic	Pourcentage de réponse à l'incitation publicitaire.	Nombre de clics constatés/Nombre de pages avec publicité constatées.

Le principe de base de la création de fichiers de logs

En terme de mesure d'audience, une des principales caractéristiques d'Internet par rapport aux médias traditionnels c'est sa traçabilité, c'est à dire le fait que les échanges de données entre le client et le serveur sont enregistrés dans des fichiers texte appelés fichiers de logs (Costes, 1998a).

Les fichiers de logs, encore appelés fichiers de traces ou journaux de connexions, sont des fichiers texte qui stockent, les unes à la suite des autres, les lignes d'informations générées par le serveur. Comme le montre la figure 1, lorsqu'un utilisateur d'Internet, que l'on appelle alors un client, demande une page qui peut contenir du texte, du son, des images ou encore de la vidéo à un site Web hébergé sur un serveur, on parle alors de requête, une ligne de texte s'inscrit dans les fichiers de logs du serveur.

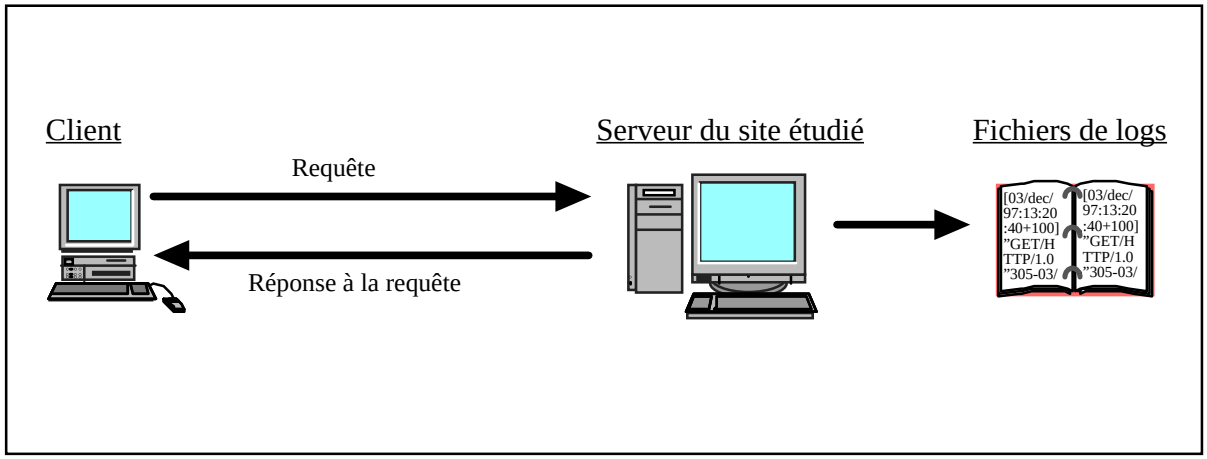

Figure 1. - Réponse directe du serveur du site étudié à la requête du client

Il semble cependant nécessaire de souligner que la fonction première des fichiers de logs est de permettre l'exploitation du serveur, c'est à dire son administration informatique comme par exemple le dimensionnement des lignes ou le diagnostic des problèmes techniques. L'utilisation marketing de ces fichiers visant à mesurer l'audience grâce à leur retraitement n'est donc qu'un détournement de la fonction première de ces données.

Les actions de chaque client laissent donc des traces dans les fichiers de logs du serveur. Ces traces sont, suivant leur nature, laissées dans l'un des quatre types de fichiers présents sur un serveur Web. Nous verrons dans la suite de cet article que ce fonctionnement de base est perturbé par des problèmes à la fois technologiques et méthodologiques.

La typologie des informations contenues dans les fichiers de logs

Chaque fichier de logs enregistre des informations différentes et complémentaires :

- les logs de transfert
- les logs d'erreur
- les logs référentiels
- les logs d'agent

Les logs de transfert

Ils enregistrent tous les transferts de fichier résultant d'une requête d'un client à un serveur. Leur analyse révèle notamment le moment (date et heure) de la connexion, le nom ainsi que le type (texte, image, son ou vidéo) du fichier demandé et le nom de la machine-hôte, c'est à dire de la machine du client. C'est grâce au suffixe du nom de la machine-hôte qu'il est possible d'extrapoler l'origine géographique du visiteur. Nous verrons que cette extrapolation n'est cependant pas exempte de défauts.

Trois principaux formats de logs de transfert existent : le *Common Log Format*, l'*Extended Common Log Format* et le *Harvest*. L'exemple qui suit porte sur le format le plus courant.

Une ligne de *Common log Format* dans un fichier de logs s'écrit de la façon suivante :
cache-1.www.anonyme.com--[01/Sep/1998:18:39:52+0100]“GET/image/image3.gif”
L'expression "cache-1.www.anonyme.com" désigne le nom de la machine-hôte du client qui a fait une requête; "01/Sep/1998:18:39:52+0100" indique la date (jour/mois/année) et l'heure (heure:minute:seconde+différentiel par rapport au méridien de Greenwich) de la requête. Enfin, "GET/image/image3.gif" signifie que le client a demandé (GET) une image appelée *image 3* en format *gif*[3].

Les logs d'erreurs

Ils conservent la trace des incidents survenus lors d'une transaction entre le client et le serveur étudié. Il peut s'agir d'une erreur dans l'écriture de l'adresse du site, appelée URL[4] sur Internet, ce qui empêche le serveur de trouver le fichier demandé par le client, d'une tentative d'accès à un répertoire protégé ou de la demande d'un fichier qui a été supprimé sur le site par exemple. Le principal intérêt des fichiers de logs d'erreurs, en terme de mesure d'audience, est de détecter sur quel fichier le téléchargement a été interrompu et donc de repérer les pages sur lesquelles les utilisateurs quittent le site Internet. Cependant, l'analyse de cette information demeure complexe et encore peu pratiquée.

Les logs référentiels

Ils facilitent l'identification à la fois du site depuis lequel le client est arrivé (par exemple un moteur de recherche ou un site contenant un lien avec le site étudié) et de la page du site étudié sur lequel le client est arrivé.

Par exemple, la ligne "http://sitededépart.fr/lien -> /accueil.html" signifie que le client est arrivé depuis la page *lien* d'un site appelé *sitededépart* sur la page appelée *accueil* du site étudié. L'analyse de ces fichiers permet donc d'analyser les parcours de navigation des internautes.

Les logs d'agent

Ils archivent les informations portant notamment sur l'équipement informatique et sur le navigateur[5] utilisé par chaque client. Ces fichiers de logs sont surtout utilisés par les logiciels de gestion de campagnes publicitaires qui exploitent ces variables dites d'environnement afin de délivrer des bannières adaptées en fonction de ces variables. Enfin, ils sont un outil utile dans la détection des robots dont l'activité fausse les résultats de mesure d'audience.

Mesurer l'audience d'un site grâce aux fichiers de logs revient donc à classer les informations contenues dans ces fichiers avec des algorithmes de tri. Mais les données obtenues ne sont pas exemptes de défauts car des contraintes technologiques inhérentes au fonctionnement du réseau Internet rendent certains résultats imprécis.

---

3 Le format gif est le format de compression des fichiers images le plus utilisé sur Internet.

4 L'URL (Uniform Ressource Locator) est l'adresse d'un site web sur Internet.

5 Un navigateur, ou browser en anglais, est une application logicielle permettant d'aller sur Internet. Les navigateurs les plus connus sont Netscape et Microsoft Explorer.

Les limites de l'analyse des fichiers de logs

Outre les problèmes de sous-estimation de l'audience dus aux caches des serveurs proxies et des navigateurs, l'analyse des fichiers de logs se heurte à une détermination incertaine des origines géographiques des visiteurs et à une perturbation des données due aux incessantes visites des robots.

Les problèmes inhérents à l'utilisation des caches sur le réseau

Deux types de caches coexistent sur le réseau : les caches des serveurs proxies et les caches des navigateurs. Ces caches ont été mis en place essentiellement dans le but de fluidifier le trafic en allégeant la masse de données transférées.

Les caches des serveurs proxies

Un serveur proxy est un serveur relais permettant à un fournisseur d'accès ou à une Université par exemple de stocker les pages Web qui font l'objet des requêtes les plus fréquentes.

Comme le montre la figure 2, lorsqu'un client fait une requête, cette dernière peut transiter par un serveur proxy. Si le serveur proxy détient déjà les pages correspondant à la requête dans son cache, il ne va pas transférer cette requête au serveur étudié, aucune ligne n'apparaîtra donc dans les fichiers de logs et les résultats obtenus ne comptabiliseront pas ce client.

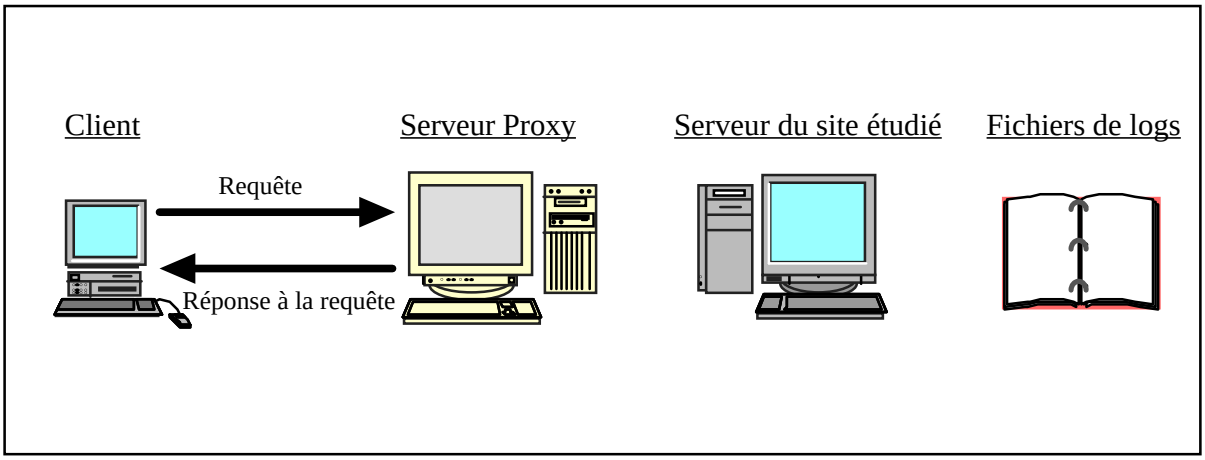

Figure 2. - Réponse directe du serveur proxy à la requête du client

Le fonctionnement de base est celui décrit dans la figure 2, mais il semble nécessaire d'affiner cette analyse. En effet, les informations diffusées sur les sites étant modifiées régulièrement, mis à part le système des pages dynamiques exposé précédemment, les serveurs proxies font des requêtes au serveur étudié afin de vérifier si les composantes de la page dont ils disposent dans leur cache n'ont pas été modifiées. Il y a donc une génération de ligne dans les fichiers de logs du serveur étudié. Cependant, aucune information sur le client réel ne sera disponible, comme par exemple son origine géographique.

Les caches des navigateurs

Une mémoire-cache est un espace disque adjoint au navigateur, qui remplit, sur l'ordinateur du client, le même office que le cache d'un serveur proxy.

Comme le montre la figure 3, lorsque un client fait une requête, si le navigateur détient déjà les pages correspondants à la requête dans son cache, il ne va pas transférer cette requête au serveur étudié, aucune ligne n'apparaîtra donc dans les fichiers de logs et les résultats obtenus ne comptabiliseront pas ce client. Lorsque le client fait des requêtes, le navigateur stocke les informations en deux temps : d'abord dans la mémoire vive, ensuite sur le disque dur de sa mémoire cache.

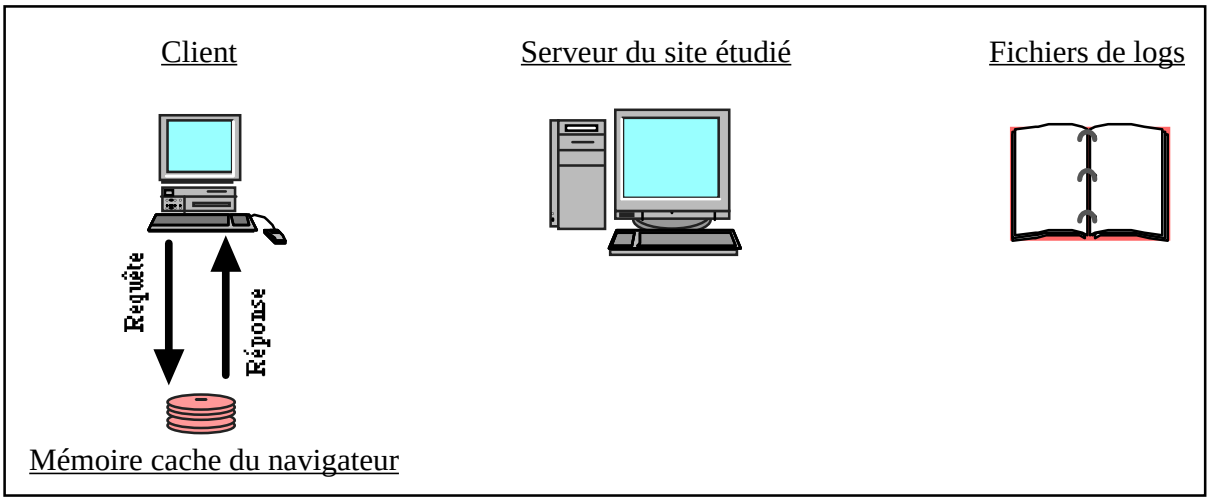

Figure 3. - Réponse directe du navigateur à la requête du client.

La difficile détermination des origines géographiques réelles des visiteurs

Comme cela a été exposé précédemment, les fichiers de logs de transfert conservent les noms des machines-hôtes des clients ayant visité le site étudié. Mais le retraitement de cette information ne permet cependant pas une détermination exacte des origines géographiques réelles des visiteurs. Par exemple, l'extrait suivant, cache-2.www.anonyme.net, indique qu'un serveur proxy du domaine ".net" s'est connecté au serveur étudié. Mais cette information ne donne aucun renseignement sur l'origine géographique réelle de ce serveur proxy. En effet, une machine-hôte dont le nom se termine par ".net" peut se trouver à n'importe quel endroit de la planète. Ce problème se pose pour tous les noms de domaines suivants : commercial (.com), network (.net), organisation (.org) et éducation (.edu).

Ce problème ne va pas aller en s'améliorant avec la création des sept nouveaux noms de domaines mondiaux : entreprises (.firm), magasins (.store), entreprises sur le net (.web), activités artistiques (.arts), divertissement (.rec), information (.info) et pages personnelles (.nom).

Les solutions apportées à ce problème sont encore assez inefficaces. Pourtant, pour contrôler l'effectivité des qualités internationales de ce nouveau média, des données de mesure d'audience fiables sur l'origine géographique réelle sont indispensables.

L'audience générée par les robots

Des robots scannent en permanence les sites présents sur le World Wide Web et ceci avec plusieurs objectifs. Les robots des moteurs de recherche visent à détecter les nouvelles adresses afin de les intégrer à leur base de données, d'autres robots génèrent automatiquement des pages faites à partir d'informations collectées sur d'autres sites, ce problème est significatif pour des sites comme celui de l'Agence France Presse par exemple, ou encore, des robots scannent le World Wide Web afin de constituer des corpus d'étude. Les requêtes générées par les robots ne doivent pas être intégrées aux données de mesure d'audience car elle ne correspondent pas à des visites d'individus ayant été exposés aux pages du site étudié.

La solution actuellement adoptée pour résoudre ce problème est de collecter les noms des machines-hôtes des robots et de retirer les lignes de fichiers de logs contenant ces adresses. Cette méthode est la plus efficace aujourd'hui, mais, rien ne permet de contrôler l'exhaustivité des bases de données recensant les adresses des machines-hôtes.

Un autre problème similaire mais encore plus difficile à résoudre est celui posé par les logiciels de capture de sites qui permettent de consulter les sites *off-line*. En effet, le rapatriement de l'intégralité d'un site par un client ne signifie pas forcément que ce site sera consulté ou encore qu'il sera consulté par une seule personne. Lors du retraitement, il est donc impossible de déterminer si les pages enregistrées par ces robots ne doivent pas être comptées, doivent être comptées une fois, ou plusieurs fois. La mise en place d'un protocole de recherche visant à déterminer le taux de consultation des sites rapatriés à l'aide de ces logiciels permettrait d'améliorer les résultats obtenus.

S'il n'existe pas encore de solutions réellement efficaces aux problèmes de la détermination de l'origine géographique et de l'audience générée par les robots, en revanche, plusieurs solutions sont utilisées pour résoudre les problèmes de non comptabilisation de l'audience générés par les caches.

Les solutions aux problèmes générés par les caches

Trois principales solutions sont utilisées par les concepteurs de logiciels et de services de mesure d'audience pour résoudre le problème de mesure dû aux caches :

- l'insertion d'une image dynamique invisible
- la redirection vers un autre serveur
- la mise en place d'applets Java

L'insertion d'une image dynamique invisible

Sur Internet, il est possible de construire des pages de façon dynamique. Cette fonction interdit le stockage de la page contenant des informations très périssables, comme par exemple les formulaires, dans les caches des serveurs proxies et des navigateurs. Cela permet donc une comptabilisation exhaustive des pages dynamiques vues.

Il est nécessaire de noter dès à présent que cette solution ne peut pas être appliquée à tous les documents mis en ligne. En effet, si toutes les informations étaient traitées comme des documents dynamiques, l'encombrement du réseau rendrait son utilisation impossible. Il ne faut en effet pas oublier que malgré leurs inconvénients en ce qui concerne la mesure d'audience, les

caches des serveurs proxies et des navigateurs sont indispensables pour limiter l'encombrement et réduire les délais de transmission sur le réseau Internet (Costes, 1997).

Comme nous venons de le voir, il serait dangereux pour le réseau de rendre tous les documents dynamiques. Pour éviter cet inconvénient tout en profitant des avantages de cette technique en terme de mesure d'audience, il est possible d'insérer une toute petite image dynamique dans la ou les pages dont on souhaite mesurer l'audience avec précision. Cette petite image (généralement de un pixel sur un pixel) est invisible, car transparente, et donc ne modifie en rien l'aspect visuel de la page pour l'internaute.

Cette solution, dont l'avantage principal est la simplicité, a cependant quelques inconvénients :

- certains navigateurs d'ancienne génération mémorisent l'image dynamique. Cette méthode n'est donc pas totalement fiable, même si ce problème devient de plus en plus marginal avec la modernisation du parc de navigateurs.
- un certain nombre d'internautes consultent les sites sans charger les images afin d'accélérer le temps de chargement, le pixel ne sera donc pas chargé et aucune ligne ne sera inscrite dans les fichiers de logs. La mise en place d'une recherche visant à mesurer le pourcentage de la population des utilisateurs d'Internet qui consultent les sites sans charger les images donnerait une indication précieuse quant à la pertinence de cette technique.

La redirection vers un autre serveur

Cette seconde solution aux problèmes générés par les caches consiste à diffuser aux utilisateurs l'adresse d'un serveur qui ne détient aucune information; ce serveur ne peut donc pas répondre directement aux requêtes et les caches ne peuvent mémoriser aucune donnée. Les requêtes sont redirigées vers un autre serveur sur lequel les informations sont effectivement stockées. Lorsque le serveur redirige la requête, une ligne de logs est ajoutée dans ses fichiers. Ce système est largement utilisé par les logiciels de gestion de campagnes publicitaires qui seront abordés dans la suite de cet article.

Cette solution n'est cependant pas sans inconvénients :

- certain anciens proxies et navigateurs mémorisent quand même l'information dans leurs caches.
- cette technique a une incidence sur la charge globale du réseau. L'importance de cette incidence est difficile à apprécier, elle est cependant théoriquement non nulle.
- les temps d'accès à l'information sont augmentés. Si une redirection est effectuée sur tous les documents, chaque document est pénalisé, même si l'information demandée par le client est mémorisée dans le cache du navigateur ou du serveur proxy. Aucune étude précise ne permet à ce jour de connaître ce temps avec exactitude.

La mise en place d'applets Java

Les applets Java sont des programmes insérés afin d'être chargés et exécutés par le navigateur. Dans le contexte de la mesure d'audience, les applets Java sont insérés dans le seul but d'aller générer une requête au niveau des serveurs. L'avantage majeur de cette méthode est de

ne pas utiliser le protocole HTTP[6]. Aussi, lorsque le navigateur charge le fichier demandé par le client, que l'applet Java soit ou ne soit pas dans un cache, le navigateur exécute le programme et donc génère une ligne de texte dans les fichiers de logs du serveur étudié.

Cette méthode a cependant quelques inconvénients :

- les utilisateurs ont la possibilité d'interdire au navigateur d'exécuter les applets Java. En effet, ces derniers sont souvent utilisés pour faire des animations à l'écran ce qui indispose un certain nombre d'internautes. Là encore, la mise en place d'une recherche visant à mesurer le pourcentage de la population des utilisateurs d'Internet qui consultent les sites sans charger les applets Java donnerait une indication précieuse quant à la pertinence de cette technique.
- tous les navigateurs ne supportent pas le langage Java, c'est-à-dire ne sont pas capables d'en exécuter les programmes. C'est le cas des versions de Netscape antérieures à la version 2 et de certains navigateurs de petits éditeurs qui n'ont pas acquitté le prix de la licence Java.
- cette technique pénalise l'affichage des pages sur l'ordinateur client car la page demandée n'est pas affichée avant le téléchargement complet de l'applet Java.

Les logiciels et les services disponibles

Analyser les fichiers de logs afin d'en extraire des résultats de mesure d'audience exploitables implique l'utilisation d'un logiciel (vendu dans le commerce ou développé en interne). Il est également possible d'avoir recours à une société de service spécialisée qui génère des fichiers de logs sur son propre serveur pour le compte du site étudié et en assure l'analyse. Le CESP a mis en oeuvre un audit des solutions logicielles et des services afin d'émettre des recommandations sur les méthodologies utilisées. Les tableaux 1, 2 et 3 regroupent les offres disponibles sur le marché français. Le grand nombre de sociétés américaines présentes confirme bien la dimension internationale des problèmes terminologiques soulevés dans la première partie de cet article dus à la présence d'une majorité de produits américains dont les résultats sont présentés en anglais.

Les logiciels disponibles

Il existe deux types de logiciels. Ils se différencient par leur contexte d'utilisation ainsi que par la nature des résultats qu'ils permettent de mesurer:

- les logiciels de mesure d'audience d'un site
- les logiciels de gestion de campagnes publicitaires

Les logiciels de mesure d'audience d'un site

Ces logiciels permettent principalement de mesurer les indicateurs d'audience éditoriale. Ils sont utilisés par les gestionnaires de sites, supports ou non de publicité, pour en améliorer l'ergonomie et le contenu éditorial. En effet, "l'écriture interactive" (Médiangles, 1997) est encore

---

[6] Le protocole HTTP (HyperText Transfert Protocol) est le protocole grâce auquel l'ordinateur d'un client et un serveur peuvent dialoguer sur le Web.

mal maîtrisée par les concepteurs de sites sur Internet. Ces données sont donc les seuls indicateurs qui puissent leur donner des directions pour faire évoluer leur site et l'adapter aux besoins des utilisateurs.

Tableau 2. - Les logiciels de mesure d'audience

Éditeur	Produit	Adresse du site Web	Commentaires
Accrue Software Inc.	Accrue Insight	http://www.accrue.com	Cette solution supporte de forts volumes de trafic et des configurations multi-sites et multi-serveurs.
Andromedia	Aria 2.0	http://www.andromedia.com	Cette solution permet de suivre l'audience d'un site en temps réel.
Cartel Informatique	Net@udience	http://www.cartel-info.fr	Seule société française sur le marché des logiciels de mesure d'audience. Solution testée par le CESP.
CQMinc	Web Tracker 2.1.42	http://www.CQMinc.com	L'interface graphique est agréable, l'utilisation en est simple et il permet de réaliser des analyses en profondeur.
Marketwave	Hit List Standard 3.5 Hit List Pro 3.5	http://www.marketwave.com	Son interface est à la fois riche et simple d'utilisation. Ce logiciel est une référence en matière d'outil dit "boîte noire".
net.Genesis	net Analysis Pro NT 3.1 net Analysis Pro UX	http://www.netgen.com	Malgré une représentation graphique médiocre des résultats, Net.Analysis est un logiciel simple d'utilisation et complet.
WebTrends Corporation	WebTrends Log Analyser v4.0a WebTrends Professional suite WebTrends Entreprise suite	http://www.webtrends.com	WebTrends est avant tout conçu comme un éditeur de rapports de synthèse. Ce logiciel paramétré par Compuserve a été testé par le CESP.

Les logiciels de gestion de campagnes publicitaires

En terme de mesure d'audience, ces logiciels ont pour principale fonction d'éditer des bilans de campagne pour les annonceurs. Ils mesurent donc les indicateurs d'efficacité publicitaire, c'est à dire le nombre de clics constatés, le nombre de pages avec publicité constatées et le taux de clic, et ceci pour chaque bannière et pour chaque site support.

Tableau 3. - Les logiciels de gestion de campagnes publicitaires

Société	Produit	Adresse du site Web	Commentaires
Accipiter	AdManager 4.0	http://www.accipiter.com	C'est le concurrent principal de NetGravity sur le marché américain.
Ad Knowledge	ClickWise v1,5	http://www.adknowledge.com	Il est gratuit jusqu'à 10 000 pages vues par jour.
Doubleclick	Doubleclick DART	http://www.doubleclick.com	C'est le logiciel utilisé sur le réseau Doubleclick.
Imaginet	Ad Manager	http://www.imaginet.fr	Première société française à s'être positionnée sur ce marché, son produit a au départ été développé pour les besoins de la régie publicitaire du même groupe, ROL.
NetGravity	AdServer 3.5	http://www.netgravity.com	Leader sur le marché américain. La nouvelle version 3.5 a été lancée sur le marché en septembre 1998.
RealMedia	Open AdStream	http://www.realmedia.com	Produit leader sur le marché français.

Les services disponibles

Les sociétés de service spécialisées dans la mesure de l'audience sur Internet n'utilisent pas les fichiers de logs générés sur le serveur étudié. Comme le montre la figure 4, ces sociétés mettent en place un système leur permettant de générer des fichiers de logs relatifs à l'audience du serveur étudié sur leur propre serveur.

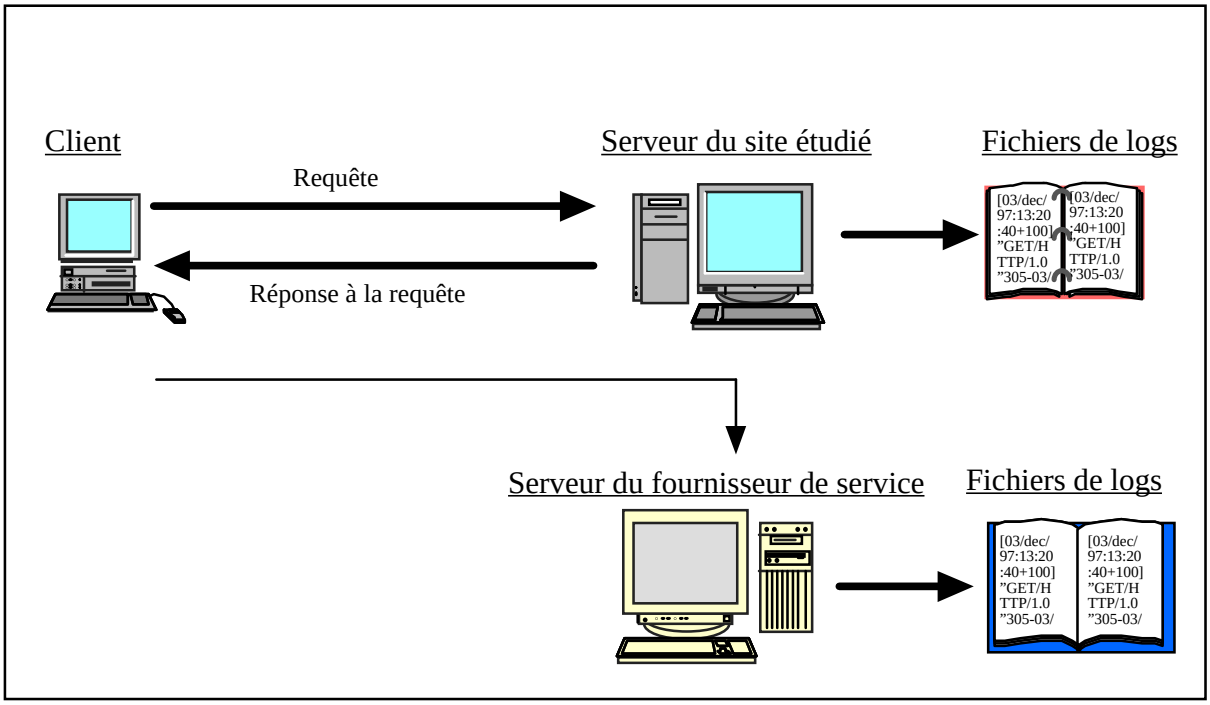

Figure 4. - Génération de fichiers de logs sur les serveurs des sociétés de service.

Les fichiers de logs générés chez le fournisseur de service ne sont donc en rien des fichiers d'administration, mais bien des fichiers ayant pour unique fonction de permettre la mesure de l'audience du site étudié.

Il est difficile de donner une idée des prix de ces solutions car les systèmes de tarification adoptés ne sont pas homogènes. Il semble cependant intéressant de citer le prix du produit Cybermonitor de Médiamétrie, leader incontestable du marché français, qui est de 14 770 francs hors taxe par an jusqu'à 30 000 pages vues par mois. Cette remarque a pour but de mettre en avant le fait que le développement des outils et services de mesure d'audience et le niveau des revenus sur Internet sont étroitement liés.

Tableau 4. - Les services de mesure d'audience

Société	Produit	Adresse du site Web	Commentaires
Services utilisant la technologie des marqueurs image			
AS/Tech	ActiveStats	http://www.astech.fr http://www.activestats.com	Son prix est remarquablement faible. Solution testée par le CESP.
Echo	eStats	http://www.estat.com	Solution testée par le CESP.
Médiamétrie	Cybermonitor.Pro	http://www.mediametrie.fr	Spécialiste de la mesure d'audience des médias traditionnels, Médiamétrie a mis longtemps à commercialiser une offre sur le marché de l'Internet. Aujourd'hui, son produit est le leader incontestable du marché et le seul véritable tiers de confiance assurant une certification des résultats. Solution testée par le CESP.
Services utilisant la technologie des marqueurs Java			
France Cybermédia	AdSuite Certifier	http://www.france-cybermedia.fr http://www.adsuite.com	La version 2.0 a été annoncée, mais elle n'est pas encore disponible sur le marché. Solution testée par le CESP.
IMR:Sofres	Webmeasure	http://www.sofres.com http://www.sofresimr.com	WebMeasure a fait partie des 15 produits de l'année 1997 d'Internet Professionnel. Solution testée par le CESP.

Le fait de sous-traiter la mesure de l'audience, qui nécessite la mise en oeuvre de compétences qui ne sont pas toujours présentes au sein de l'entreprise, permet de disposer de données émanant d'un tiers de confiance. Cependant, il n'existe pas d'organisme unanimement reconnu par le marché considéré comme un organisme de certification.

Le problème de la certification de l'audience

Sur Internet, il n'existe pas d'organisme de certification reconnu unanimement par tous les acteurs du marché. Or, comme chaque site peut mesurer lui-même sa propre audience, la véracité des résultats diffusés peut être sujette à caution. En France, le CESP a mené des vagues de tests sur les logiciels et les services de mesure d'audience, mais le résultat de leur audit n'avait pas pour objectif d'aboutir à la labélisation d'un seul et unique outil, mais plutôt à l'émission d'un jugement global sur les résultats de chaque outil. Les outils audités par le CESP ont un niveau de qualité

minimum mais le CESP ne garantit en aucun cas la véracité des résultats qu'ils permettent de mesurer.

Les sociétés proposant des services de mesure d'audience, et en premier lieu Médiamétrie, se sont positionnées sur ce marché de la certification de l'audience. Mais, même si ce service est incontestablement le leader français, ce n'est pas un organisme reconnu de façon unanime par les acteurs du marché. De plus, s'ajoute là encore une dimension internationale à ce problème. En effet, certains puissants acteurs du marché américain qui ont des filiales en Europe, comme Yahoo! par exemple, prônent la certification de l'audience par des audits de procédure des outils utilisés en interne. Yahoo! France a obtenu la certification de ses résultats par ABVS Interactive, une filiale de l'Audit Bureau of Circulation, organe de référence dans l'audit de la diffusion de la presse aux États-Unis. Dans le même temps, Ernst and Young, spécialiste de l'audit des systèmes d'information, a réalisé un audit des processus d'utilisation des logiciels de mesure d'audience et de gestion de campagne publicitaire par Yahoo!. Dans le numéro du trois juillet 1998 du magazine Stratégies, Yahoo! France annonce le trafic total de son site à 34 371 279 pages vues pour le mois de mars 1998 en précisant que cette mesure est certifiée. On peut cependant s'interroger sur la validité d'une telle procédure de certification ainsi que sur sa crédibilité vis à vis des annonceurs.

Outre le problème de la certification, se pose celui de la qualification de l'audience sur Internet.

Le problème de la qualification de l'audience

Les outils présentés ci-dessus visent à mesurer les indicateurs présentés dans le tableau 1. Mais, comme le soulignent Chandon et Schrameck (1998), "pour qualifier l'audience, connaître le profil socio-démographique du visiteur, ses habitudes de fréquentation, il faut mettre en place des panels représentatifs dont la gestion est complexe et coûteuse". Cette qualification de l'audience qui s'accompagne généralement d'une classification des sites est surtout utilisée dans un objectif de médiaplanning ou lors de la détermination de la stratégie média.

Aux États-Unis, de tels outils ont été mis en place avec notamment le PC Meter de Media Metrix[7] et le panel de Relevant Knowledge[8].

Pour prendre l'exemple de Media Metrix, le PC Meter est un logiciel fonctionnant comme une "boîte noire", enregistrant et datant tout les comportements de l'utilisateur d'un ordinateur, aussi bien *on line* que *off line*. Son fonctionnement est le suivant : lorsque l'utilisateur allume son ordinateur, une fenêtre lui demande quel individu il est parmi les personnes enregistrées. L'utilisateur clique alors sur une petite icône avec son prénom qui s'affiche à l'écran. Si l'utilisateur change, il doit à nouveau s'identifier, et chaque fois que l'économiseur d'écran se met en route, la personne doit s'identifier à nouveau. L'échantillon actuel est de 28 000 utilisateurs de PC aux États-Unis. Les principales critiques méthodologiques adressées à Media Metrix sont le fait que les PC Meter ne soient pas installés sur les lieux de travail et que leur échantillon ne soit

---

7 http://www.mediametrix.com

8 http://www.relevantknowledge.com

constitué que d'utilisateurs de PC. Pour ce qui est de la première critique, Media Metrix a créé un échantillon séparé d'utilisateurs de PC sur le lieu de travail, mais les résultats de ces tests ainsi que la taille de l'échantillon ne sont pas communiqués. Media Metrix a conduit des programmes pilotes en Allemagne, en France et en Grande-Bretagne, mais aucun échantillon n'a encore été effectivement recruté.

Relevant Knowledge quant à lui est un panel dont les membres, âgés de douze ans et plus, sont recrutés dans un échantillon d'utilisateurs localisés grâce à la méthode du *random-digit dialing sampling*. Les membres de leur panel sont des utilisateurs de PC et de Macintosh qui naviguent sur Internet depuis leur domicile, leur lieu de travail ou leur école. Les utilisateurs recrutés doivent télécharger le logiciel de Relevant Knowledge sur chaque ordinateur qu'ils sont susceptibles d'utiliser (aussi bien à domicile que sur le lieu de travail ou à l'école). Lorsque les utilisateurs accèdent à des sites Web, le logiciel enregistre les URL demandées et envoie les données à Relevant Knowledge. Les utilisateurs sont considérés comme faisant partie du panel et leurs données utilisées lorsque Relevant Knowledge reçoit la confirmation que le logiciel a été installé convenablement.

En France, ces investissements sont encore à l'étude. La Sofrès, qui avait installé des PC Meter sur le marché français, les a retirés, Relevant Knowledge est en phase de test sur le marché français et la société Net Value a annoncé le lancement d'un panel Européen pour le premier ou le second trimestre 1999. Ces différences dans le développement de ces outils sur les marchés français et américains tiennent à leur taille respective. En effet, si l'on compare les résultats de l'étude menée par PricewaterhouseCoopers pour le compte de l'IAB en France et aux États-Unis, on se rend compte que sur le marché américain, les revenus de la publicité sur Internet sont plus de trente fois supérieurs puisqu'ils sont de 907 millions de Francs contre 28 millions de Francs en France pour l'année 1997.

Conclusion

Bien que conçu pour un usage d'administration des serveurs, les fichiers de logs sont la matière première de la mesure d'audience sur Internet. Bien que permettant de calculer assez simplement une batterie d'indicateurs d'audiences éditoriale et publicitaire, cette technologie n'est pas exempte de défauts. Le principal problème est celui de la sous-estimation de l'audience comptabilisée due à la présence des caches des serveurs proxies et des navigateurs. Des solutions au problème cité précédemment, telles que l'insertion d'une image dynamique miniature, la redirection, ou la mise en place d'une applet Java, sont utilisées par les logiciels et les services de mesure d'audience. Mais aucune d'entre elles ne permet de résoudre parfaitement ce problème qui conduit à une imprécision des résultats de mesure, aussi bien pour les sites que pour les publicités. De plus, les erreurs induites par les visites des robots restent mal corrigées, et il est toujours impossible de déterminer l'origine géographique réelle d'une importante partie du trafic.

Aujourd'hui, pour des raisons dues principalement au niveau de développement du marché, les méthodologies utilisant les fichiers de logs sont, en France, favorisées par rapport à celles utilisant des panels. En effet, la mise en place de panels, si elle a déjà eu lieu aux États-Unis, représente en France un investissement important par rapport à la rentabilité que peut faire

escompter la taille actuelle du marché. Pourtant, les panels permettent de qualifier l'audience, ce qui fait encore défaut aux données de mesure d'audience sur Internet.

Pour ce qui est de la terminologie, elle est encore loin d'être stabilisée. Les évolutions technologiques et méthodologiques, ainsi que l'apparition de nouveaux outils, modifient aussi bien les concepts que les mesures, et la dimension internationale d'Internet complexifie encore le travail d'harmonisation. Pourtant, la comparabilité des données de mesure d'audience, aussi bien dans un contexte éditorial que publicitaire, passe par une harmonisation de la terminologie.

Enfin, sur un média où chacun peut mesurer sa propre audience avec des technologies et des méthodologies comparables mais différentes, l'adoption d'un organisme certificateur national ou international unanimement reconnu semble devoir être un processus long qui vient juste de commencer.

Bibliographie

Baumard Philippe et Bernard Forgues (1995), Internet : un outil transnational au service du commerce, *Décisions Marketing*, 5, 21-32.

Chandon Jean-Louis et Benjamin Schrameck (1998), Publicité sur le Net et mesure des audiences des sites, *La lettre Marketing*, 9, mars, 2-3.

Costes Yseulys (1997), *Comment mesurer l'audience d'un site Web* ?, Paris, A Jour-Groupe Tests.

Costes Yseulys (1998a), Que vaut la mesure d'audience sur le Net ?, *Internet Professionnel*, 18, mars, 72-77.

Costes Yseulys (1998b), La mesure d'audience sur Internet, *Décisions Marketing*, à paraître.

Ginburg Stu (1998), Online Ad Industry Will Develop Audience Measurement Standards, http://www.iab.news/measure-source.html

Médiangles (1997), *L'observatoire du commerce électronique en France*, Paris, A Jour-Groupe Tests.

Murphy Ian P. (1996), On-line Ads Effective? Who Knows For Sure?, *Marketing News*, 30, 20, 1 et 38.

Onnein-Bonnefoy Carole (1997), Les bandeaux publicitaires sur Internet : Mesures d'efficacité, *Décisions Marketing*, 11, 87-92.

Reboul Pierre et Dominique Xardel (1997), *Le Commerce Électronique*, Paris, Eyrolles.