CENTRE DE RECHERCHE DMSP

Mesures d'audience sur Internet : A la croisée des chemins entre approche publicitaire et marketing direct

de PECHPEYROU* Pauline,
GOUDEY Alain, DESMET Pierre

Cahier n°324
Septembre 2003

Pauline de Pechpeyrou et Alain Goudey

Doctorants au centre de recherche DMSP
Université Paris Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16

Pierre Desmet est professeur à l'université Paris-Dauphine
Place du Maréchal de Lattre de Tassigny
75775 Paris Cedex 16

* auteur à contacter : paulinedepechpeyrou@yahoo.fr

La communication sur Internet : A la croisée des chemins entre approche publicitaire et marketing direct

Résumé
Principalement sous forme de bannières, la communication sur Internet offre des possibilités de communication attractives mais ne représente encore qu'une très faible partie des investissements publicitaires tous média considérés. La mesure de l'audience et sa qualification sont essentielles tant pour le site qui cherche à connaître son trafic et ses performances, que pour l'annonceur qui vise à mettre en œuvre une politique de communication *online*. Or, du fait de la spécificité et de la jeunesse du média Internet, les outils de mesure sont encore balbutiants et leur divergence est un frein au développement de la communication publicitaire.
Cet article présente les caractéristiques et les différences entre les mesures de fréquentation sur le site (*site centric*) et celles effectuées à partir de panels (*user centric*). Sont ensuite abordés les obstacles, tant méthodologiques que techniques, qui restent encore à surmonter pour harmoniser les méthodes de mesure d'audience.

Mots-clés : Communication, Internet, mesure, audience, performance, panels

Internet traffic : at the crossroads of advertising approach and direct marketing

Summary
Mainly based on banners, online communication is very attractive but consists in a small part of overall advertisement investments. Audience measurement and its qualification are primordial for both sites which want to measure their traffic and performance, and also for announcers who want to plan online communication. As the Internet is still a new media, measuring tools are slowly becoming mature and their diversity is slowing down online investment in communication.

This article deals with characteristics and differences between site centric approaches and user centric ones. The authors try to underline methodological and technical difficulties for normalizing audience measurement.

Key words : Communication, Internet, measurement, audience, performance, panels

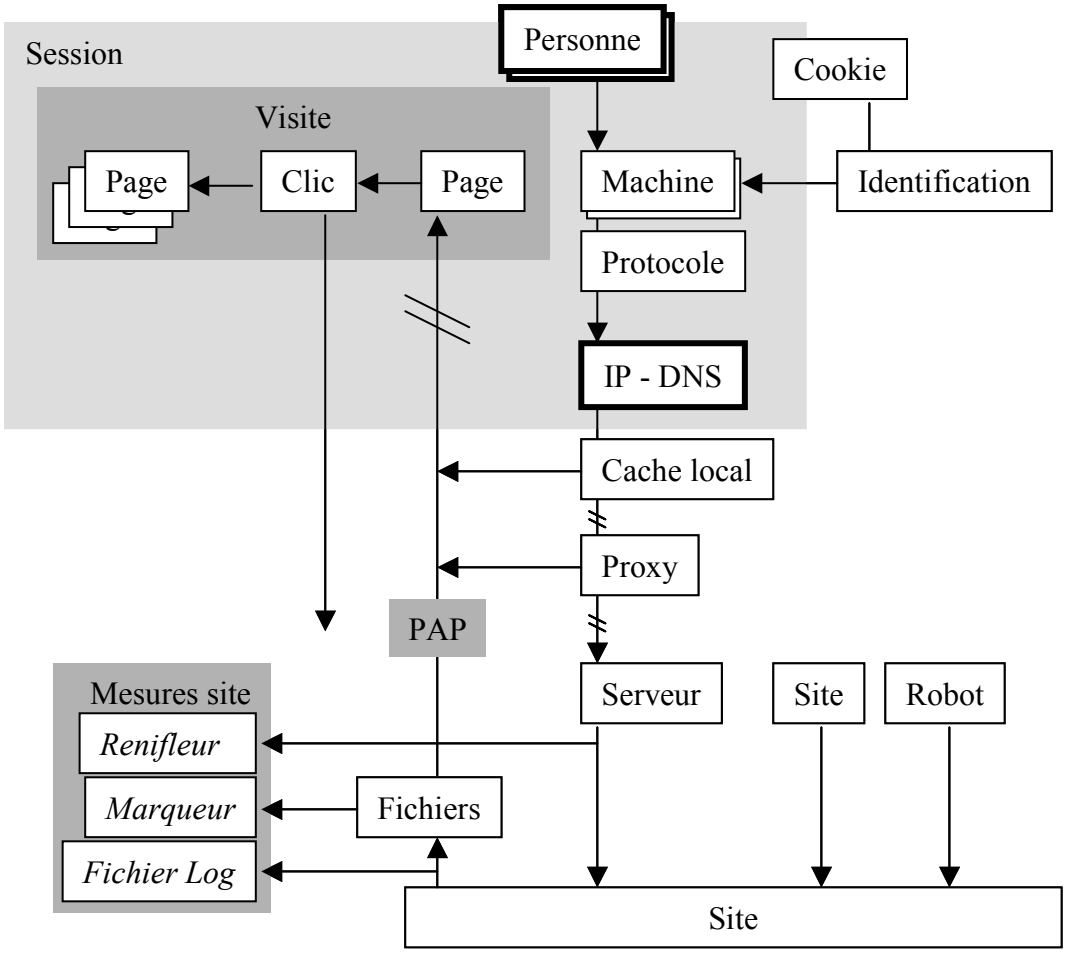
Introduction

La communication sur Internet, principalement sous forme de bannières, est évaluée à 309 millions d'euros en 2002 en France soit à peine 2% des investissements publicitaires plurimédia [TNS-Secodip]. Cependant, les prévisions lui promettent une part nettement plus significative avec 8% des dépenses publicitaires aux USA dès 2005 [Jupiter]. La couverture étant en progression rapide (37% d'internautes en France en 2002 et 62% aux USA selon TNS), le frein principal à l'utilisation d'Internet pour la communication publicitaire est l'absence de standards et la qualité inégale des mesures. Le marché représenté par ces études, évalué à 4 ,5 milliards d'euros pour 2004, est d'ailleurs très important.
La mesure d'audience et sa qualification sont essentielles aussi bien pour le site qui cherche à connaître son trafic et ses performances, que pour l'annonceur qui vise à mettre en œuvre une politique de communication en ligne. Il s'agit ici de décrire les mesures d'audience, les objectifs et les méthodes utilisées. Cette approche va permettre de mettre en évidence les sources des écarts ainsi que les solutions apportées et les difficultés persistantes.
Dans la première section, le schéma de la communication sur Internet est rappelé avec les différentes mesures. La seconde section aborde la problématique générale de la mesure d'une communication sur Internet qui, du fait de son interactivité, est située au croisement de deux problématiques, le contact pour la communication et le comportement pour le marketing direct. La troisième section permet de comparer les caractéristiques, les avantages et les limites des deux approches de la mesure d'audience sur Internet, la mesure côté site (*site centric*) et la mesure côté utilisateur (*user centric*). Les solutions proposées pour réduire les divergences des mesures ainsi que les problèmes encore en suspens sont finalement discutés.

Un schéma général de la communication sur Internet

Si les problématiques de mesure sur Internet sont similaires à celles des médias traditionnels, leur mise en application diffère du fait des particularités de ce nouveau média. Il est donc nécessaire de cadrer le schéma de communication sur Internet, afin de voir comment les concepts traditionnels de mesure d'audience peuvent être transposés sur Internet. Contrairement aux autres médias publicitaires dont le message est figé (presse) ou reçu de manière identique (TV), une communication sur Internet est un processus interactif qui trouve sa source dans une requête. La figure 1 présente les différentes étapes de cette communication et sa relative complexité en utilisant les définitions retenues par le CESP (Centre d'Etude des Supports de Publicité).

Figure 1 : Schéma des éléments d'une communication sur Internet



Une personne peut se connecter sur Internet (a) de différents lieux, privés (domicile, travail) ou publics (université, cybercafés,…), (b) en utilisant différentes machines ou hosts (ordinateur, assistant personnel, téléphone mobile,…) (c) pour des tâches différentes (site, courrier électronique, chargement de fichiers, forums de discussion,…) avec différents protocoles (http, ftp, pop,…) (d) en passant (éventuellement) par un fournisseur de connexion avec un identifiant de connexion (IP ou DNS). Plusieurs points peuvent ainsi faire l'objet de mesure : les visites, qui représentent le trafic du site, les requêtes, qui représentent l'interaction de l'internaute et du site, et les sessions, qui permettent d'étudier le comportement de navigation.

La visite et la session

Une *session* est une interaction personne-machine qui peut comporter une ou plusieurs visites sur un ou plusieurs sites. La succession des pages vues sur un même site constitue une *visite*. Elle débute par une connexion ou une phase d'identification et est déclarée terminée si la période d'inactivité excède un seuil (fixé à 30 minutes par le CESP), par un changement de navigateur ou par un changement d'identifiant (IP-DNS). Sur une période déterminée, par exemple un mois, un *visiteur* est une personne ayant consulté au moins une fois le site et un *internaute* a utilisé au moins l'un des protocoles sur le réseau.
L'identification d'un poste et le suivi de la connexion sur un site sont possibles grâce à un témoin de connexion (*cookie*), petit fichier texte placé par le serveur du site ou un tiers autorisé, sur le

disque dur du poste connecté à l'occasion d'une consultation. Il permet notamment de recueillir et de stocker des données sur le comportement de navigation à partir du poste connecté.

Les requêtes

Une requête est un clic d'appel d'une page directement à partir de l'adresse ou en transitant par d'autres sites proposant des liens en fonction des demandes (moteur de recherche) ou des liens fixes. Pour une communication publicitaire, la requête est effectuée par un clic sur une bannière. L'habitude du double clic ou l'impatience suite à l'absence de réaction visible peuvent générer plusieurs clics pour la demande d'une même page, et donc, plusieurs requêtes.
Outre les visiteurs, des requêtes automatiques sont effectuées par des logiciels (robots ou *bots*) qui explorent les pages proposées par les sites soit pour les indexer soit pour d'autres motifs comme des vérifications légales (droits d'auteur). Enfin, des requêtes quotidiennes sont effectuées par les administrateurs du site.

La réponse à une requête

Une requête d'un utilisateur peut être satisfaite (a) *sur site* si la connexion a lieu avec le site lui-même, ou *hors site* soit (b) directement par la machine utilisée dont le navigateur a gardé en mémoire la page déjà consultée auparavant, (c) par un serveur relais (*proxy*) qui a chargé les pages fréquemment demandées. De nombreuses consultations hors-site sont ainsi ignorées mais il existe des solutions informatiques à base de scripts qui éliminent la consultation hors-site en rendant obligatoire l'appel d'une mise à jour. Malgré son envoi par le site, la page appelée, ou l'un des fichiers qui la compose, peut ne pas être vue lorsque le visiteur a paramétré son navigateur pour filtrer certains fichiers, annulé ainsi le téléchargement ou terminé sa session.

La page et sa consultation

Lors de la requête, le site envoie un ensemble de fichiers (*hits*) qui forment une page, réelle (statique) ou virtuelle (dynamique) c'est-à-dire constituée en fonction des spécifications de la requête. La page peut être envoyée par un serveur ou faire appel à un serveur spécifique de bannières. Cette possibilité d'interaction est une caractéristique spécifique d'Internet qui permet l'envoi d'une page standard ou adaptée selon des critères liés à la cible ou à la répétition du message. Si celle-ci contient des éléments publicitaires (bandeaux,…), elle constitue une *impression* ou *page vue avec publicité* (PAP).
Une page peut contenir un marqueur (*tag*), code informatique (image, script CGI, applet Java, etc.) qui déclenche, si la machine est connectée, l'envoi d'une information lors de la consultation de la page à un serveur de comptabilisation. Si le téléchargement de la page est arrêté, seuls les marqueurs des fichiers envoyés en premier seront actifs.
Un fichier répertorie l'activité de réponse du serveur aux requêtes (journal de connexion ou fichier *log*). Chaque enregistrement identifie pour chaque fichier envoyé (*hit*), la date et heure, les référentiels (site, page d'origine et page demandée, la séquence ayant conduit à la requête), les informations sur l'équipement de l'utilisateur (navigateur,…), le résultat de la demande (erreurs,…) et les caractéristiques du fichier (taille,..).

Approches publicitaire et marketing direct sur Internet

Les nombreuses mesures qui existent sur la communication Internet diffèrent selon leur objectif et selon la perspective de l'utilisateur : annonceurs, fournisseur d'accès, agence-conseil, sites. Bhat (2002) recense ainsi de nombreux indicateurs adaptés aux contextes d'utilisation : mesure d'exposition, de qualité de la relation avec le site, d'utilité des pages du site, de réussite d'actions de co-marketing et d'efficacité de ciblage.

Les approches publicitaire et marketing direct

En tant que média pour un annonceur, Internet est à l'intersection d'une approche publicitaire et d'une approche marketing direct.

a) Approche publicitaire

En tant que support publicitaire, Internet présente des avantages vis-à-vis des médias classiques : possibilité d'analyse de retour sur investissement, faible coût, optimisation en temps réel, influence online & offline (notamment sur les achats offline). Cependant, il doit aussi satisfaire les objectifs traditionnels d'un support publicitaire, en particulier engendrer une influence durable sur la notoriété et sur l'attitude vis-à-vis d'une offre. Le comportement n'est pas directement recherché et, en tout cas, aucune mesure comportementale n'est directement associée à la performance de la communication. Les objectifs tactiques sont l'émission d'un message sur une audience qualifiée et une rémunération du support en fonction de la taille et de la qualité de l'audience.
Les choix sont structurés selon plusieurs niveaux hiérarchiques : médias, supports et emplacements, chacun ayant une problématique propre. Les choix entre les médias se font sur la qualité de la communication, la capacité à transmettre des éléments particuliers et donnent lieu à des mesures d'efficacité spécifiques (compréhension, mémorisation, attention,…). Le choix entre les supports privilégie la dimension économique et la couverture de la cible (ou la puissance). Pour le choix des meilleurs emplacements, il est nécessaire d'étudier le comportement de fréquentation du support, l'exposition, la circulation à l'intérieur d'un magazine par exemple. Shamdasani et al. (2001) montre que l'audience de sites spécifiques est maintenant suffisante pour permettre une communication plus ciblée sur des produits à forte implication et que le critère d'affinité peut être préféré à celui de puissance du fait du coût élevé du contact sur les sites généralistes à forte fréquentation.

b) Approche marketing direct

L'approche marketing direct est plus orientée vers des objectifs commerciaux. La communication recherche alors directement un comportement, de l'émission d'un signe d'intérêt à la fréquentation voire l'achat. L'obtention d'un contact avec un prospect (*lead*) est le critère de choix et de rémunération en fonction d'un objectif de recrutement lié à la capacité disponible pour la réponse. L'analyse est effectuée en regard d'un comportement, des coûts de recrutement d'un client, voire de sa valeur vie entière (*Life time value*). Le comportement est décomposé en différentes étapes avec des taux associés : avoir vu, avoir cliqué, avoir demandé de l'information, avoir commandé, avoir payé, ne pas avoir renvoyé ou annulé. Les taux peuvent être corrélés de manière négative, par exemple un impact fort d'une bannière entraîne un fort effet de curiosité et un faible taux de concrétisation.

Les mesures d'audience et de fréquentation

Les mesures associées à ces deux problématiques sont spécifiques, mesure de l'audience pour la communication publicitaire et mesure de fréquentation pour l'approche marketing direct.

(a) L'audience

Répondant à une préoccupation publicitaire, l'étude de l'audience consiste à dénombrer les internautes et à étudier leur comportement pour replacer la communication sur Internet au sein des mesures déjà disponibles sur les autres médias (mémorisation, attention,…). Pour un site comme pour un support, l'objectif est de caractériser les visiteurs sur des critères socio-économiques et de centres d'intérêt pour pouvoir utiliser des critères d'affinité.

Il est par ailleurs tout particulièrement intéressant d'étudier les indicateurs classiques d'audience, notamment la couverture brute (pourcentage d'internautes ayant eu au moins une ODV, occasion de voir) ou pourcentage de visiteurs uniques sur une période (*reach*). Les courbes utilisées pour les autres médias ne sont pas adaptées à Internet qui permet d'atteindre un pourcentage maximal beaucoup plus élevé que ceux des autres médias (Wood, 1998).

(b) La fréquentation

Elle permet de dénombrer les visiteurs d'un site pour fixer la fréquentation attendue sur les échelles de puissance et d'économie. Cette dernière échelle peut correspondre à différents niveaux de comportement selon les objectifs : nombre d'impressions pour une exposition, clic voire au delà.

Pour le responsable du site, l'information est détaillée de manière statique au niveau des pages et des rubriques (agrégats de pages) ainsi que des origines de la requête. Elle permet, de manière dynamique, de suivre la trajectoire de fréquentation au sein du site (pages d'entrée et de sortie,…).

Les modèles économiques liés à ces mesures

Du fait des différents acteurs et problématiques, Shen (2002) relève la coexistence de trois modèles principaux de fixation des prix pour l'espace publicitaire sur Internet : un modèle basé sur l'*exposition* avec le nombre d'impressions, un modèle basé sur l'*interaction* avec les clics des internautes et un modèle basé sur la *performance* avec le nombre d'achats induits.

Les modèles basés sur l'interaction ou la performance sont préférés aux Etats-Unis, avec pour indicateurs principaux le *Cost-per-Click* (ou CPC) où l'annonceur sera facturé sur la base du nombre d'internautes qui cliquent effectivement sur sa publicité, et le *Cost-per-Sale* (ou CPS) où la base est le nombre d'internautes qui, attirés par la publicité, passent commande en ligne (au sens large : commande d'un produit, inscription à une lettre d'information,…).

Cette approche conative de la mesure de l'efficacité publicitaire sur Internet à travers les actions des internautes exposés à la publicité est critiquée du fait de la non prise en compte des effets attitudinaux. En effet, dans le cadre de la théorie de la simple exposition à la publicité (" mere exposure "), des études (Briggs et Hollis, 1997) montrent que la visualisation de la publicité, même en l'absence d'action de la part de l'internaute, contribue à augmenter significativement les indicateurs traditionnels utilisés pour la mesure de l'efficacité du message publicitaire, notamment la notoriété de la marque, le souvenir publicitaire spontané, la reconnaissance et l'attribution. D'après une étude de Hussherr et Rosanvallon (2001), le taux d'attribution d'une bannière publicitaire à une marque est de 18% des internautes exposés et le taux de mémorisation spontané de 11% alors que le taux de clic ne sera que de 1%.

Par conséquent, mesurer l'efficacité d'une bannière publicitaire sur Internet seulement à partir du nombre de clics, assez faible, tend à sous-estimer l'impact global de la publicité. Comme pour les autres médias de masse, l'utilisation d'une bannière se révèle peu adaptée, et en tout cas génère un coût de recrutement élevé, pour une campagne de prospection en marketing direct. Ce constat

justifie l'existence du dernier modèle économique, basé sur l'exposition. La mesure de l'exposition (audience), principalement utilisée en France et en Europe, repose sur l'indicateur du *Coût Pour Mille impressions* (ou CPM). Cet indicateur est considéré comme le plus adéquat pour les annonceurs qui visent à faire des campagnes d'image ou de notoriété.
L'utilisation de ces modèles est différente selon que l'objectif correspond à la mesure d'efficacité de la campagne ou à la fixation d'une grille tarifaire. L'étude de Shen (2002) révèle que le CPM arrive en première position dans les modèles de prix, avec 90,2% des agences interrogées déclarant y recourir fréquemment, mais qu'il n'arrive qu'en deuxième position (après le taux de clic) dans les modèles de mesure de l'efficacité de la campagne publicitaire. L'échantillon composé uniquement d'agences américaines, montre bien la préférence évoquée aux Etats-Unis pour la mesure de l'efficacité à travers les comportements.

Les mesures actuelles et leurs limites

Les mesures proposées sont spécifiques à chaque problématique : la mesure de l'audience repose sur des analyses externes d'échantillons représentatifs d'internautes (*user centric*) et s'appuie sur un nombre de visiteurs uniques, alors que la fréquentation est obtenue par le traitement, interne ou par des tiers de confiance, des journaux de connexion (*log*) ou des marqueurs émis par la consultation des pages (*site centric*) et s'appuie sur un nombre de requêtes, sans possibilité d'identifier de façon fiable les visiteurs uniques. Nous présenterons plus en détail ces deux approches et leurs limites.

Les mesures centrées sur l'utilisateur

Pour décrire le comportement de l'internaute, différentes études statistiques sont menées sur des échantillons par des instituts de sondages et des sociétés d'études. Elles fournissent des informations riches pour mieux décrire et comprendre les comportements des internautes et leurs évolutions. En plus de leur coût élevé, ces études sont sujettes à une incertitude directement reliée à la taille de l'échantillon et à la faible fréquence de certains comportements.
Les études de cadrage sur des échantillons de très grande taille, ont pour objectif de connaître la population mère (l'univers des internautes) en vue de fixer les caractéristiques des échantillons représentatifs. Ces enquêtes sont elles-mêmes soumises à un biais de représentativité, leurs résultats pouvant être redressés pour correspondre à une information publique, délivrée par exemple par les statistiques des abonnements aux fournisseurs d'accès.
Des échantillons de taille plus restreinte sont utilisés pour étudier des comportements spécifiques soit lors de sondages qui répondent aux objectifs d'études *ad hoc* menées par les sociétés d'études, soit dans le cadre de panels qui permettent le suivi du comportement d'un échantillon d'internautes identifiés stable dans le temps.

Les mesures centrées sur le site

Les mesures centrées sur le site, généralement des mesures de fréquentation, se font par les témoins de connexion, les marqueurs et l'analyse de fichiers de connexion.
Un témoin de connexion (*cookie*) permet au gestionnaire du site de distinguer un poste connecté d'un autre poste, voire même sur un poste donné, deux comptes utilisateurs différents et donc d'améliorer la précision de la mesure de fréquentation des sites. Leur usage pour connaître la fréquentation d'un site est sujet à controverse de par leur caractère intrusif, même si la plupart d'entre eux ne contiennent pas d'informations nominatives. De plus, les internautes peuvent à tout moment, durablement ou temporairement, s'opposer à l'implantation de tels fichiers. Dans ce cas, la précision de la mesure de fréquentation basée sur cette méthode s'en trouve largement altérée.

Pour pallier ces inconvénients, les sites se sont de plus en plus tournés vers la méthode des marqueurs. A la différence du cookie, activé sur le poste par le site à chaque connexion, le marqueur est inséré dans les pages du site et permet de renseigner une base de données sur l'affichage effectif du marqueur et, donc, de la page. Le site peut choisir de marquer les pages dont il souhaite connaître l'activité, ou l'ensemble pour se comparer aux sites concurrents dans les classements effectués pour éclairer les choix médias. Cette méthode nécessite de prévoir à l'avance les marqueurs, leur enchaînement et leur signification. Elle apparaît particulièrement performante au niveau des statistiques fournies dans la mesure où elle a été réfléchie en terme marketing au moment du développement technique du site. Elle permet ainsi de répondre aux objectifs de mesure de l'efficacité d'une campagne publicitaire, de la mise en place de partenariat, de mesurer l'impact d'une newsletter...
L'analyse des fichiers de connexion constitue sans aucun doute l'outil de mesure *site centric* le plus utilisé. En effet, contrairement aux marqueurs, elle ne nécessite aucune intervention sur le contenu du site. A l'inverse des cookies, elle ne présente aucun caractère intrusif. Les fichiers *logs* représentent donc une solution peu coûteuse et néanmoins riche en informations.
L'analyse des fichiers de *logs* est complexe (Galan, 2002) et connaît de nombreux développements visant à apporter une composante qualitative aux informations traditionnellement fournies en termes de nombre de visites et de pages visualisées. L'analyse réseau développée par Ferrandi et Boutin (1999) permet ainsi de visualiser les pages fortement connectées entre elles et de différencier les visiteurs en fonction de leur utilisation du site, ou de segmenter par exemple les acheteurs et les non-acheteurs d'un site marchand.

Les limites actuelles et les solutions proposées

Malgré la rapidité de développement d'indicateurs à la signification partagée, des différences importantes existent encore entre les résultats fournis non seulement par des approches différentes, mais aussi par des prestataires ayant des approches identiques. Plusieurs difficultés sont actuellement en voie de résolution. Il s'agit des écarts dus à des différences de définition et des risques liés à une auto-évaluation de leur audience par les sites.

a) Les différences dues aux échantillons

Des différences statistiques importantes existent entre les résultats d'études statistiques qui ne définissent pas de la même manière la population mère. Comment choisir entre un univers de 4,5 millions d'internautes (Médiamétrie) et de 6,8 millions (Netvalue) ? De telles différences proviennent de multiples sources dont les effets combinés conduisent à des écarts, parfois considérables, de classement, y compris dans le classement des dix sites les plus fréquentés.
Les lieux et les modes de consultation ne sont pas couverts de la même manière. Au départ limités en France à la consultation au domicile, les panels d'internautes suivent maintenant la consultation sur le lieu de travail depuis 2002 (MegaPanel et NetValue) mais négligent encore d'autres lieux privés (écoles et universités) ou publics (cybercafés) et toute consultation nomade (assistants personnels, téléphones mobiles).
Les échantillons ne recouvrent pas les mêmes tranches d'âge. Ainsi, Nielsen Netratings fixe l'âge minimum d'un internaute à 2 ans, Médiamétrie et CSA TMO à 11 ans et NetValue à 15 ans. Ces différences de définition conduisent non seulement à des mesures d'audience différentes mais également à des comportements de navigation différents. En effet, les adolescents présentent un comportement de navigation spécifique : alors que leur navigation dans le cadre scolaire est axée sur des recherches précises et spécifiques, leur pratique d'Internet à domicile est plus soutenue et sophistiquée [Les jeunes et Internet, février 2002, enquête menée en France par le Clemi].
Les enquêtes de calage représentent la première étape dans la constitution du panel d'internautes, puisqu'à partir d'un premier ensemble de personnes contactées, elles vont permettre de sélectionner

les plus représentatives. La représentativité de ces enquêtes de calage est donc essentielle et la taille déjà importante des échantillons, 22 000 pour Jupiter MMXI et 24 000 pour Nielsen/Netratings, ne semble pas encore suffisante pour obtenir une définition précise de la population des internautes.
Enfin, les résultats varieront selon la taille et la représentativité des panels d'internautes. Traditionnellement, la détermination de la taille d'un panel est fonction de trois critères : la taille de la population globale, la variance existant au sein de cette population et la précision souhaitée dans l'extrapolation. Kalyanam et MacEvoy (1999) soulignent ainsi que la diversité des comportements de navigation sur Internet induirait des panels d'un million d'internautes pour atteindre une précision semblable à celle des autres médias alors que la mesure se fait à partir de panels d'une dizaine de milliers d'internautes seulement. Les problèmes résultants sont particulièrement cruciaux pour la validité et la précision des résultats fournis et pour les sites à faible trafic dont on ne peut déterminer l'audience.

b) Les différences dues au choix de paramètres à fixer

La détermination du nombre de visites et par conséquent du nombre de visites répétées est fonction du choix de la durée d'inactivité conduisant à considérer qu'une visite est terminée. Alors que l'IAB (International Advertising Bureau) et le CESP recommandent une durée de 30 minutes, les sociétés de panel fixent le plus souvent ce seuil à 10 minutes, sauf si, entre temps, l'internaute a continué sa navigation de manière continue, mais sur un ou plusieurs autres sites. Le choix de ce seuil, facilement paramétrable, n'est pas toujours spécifié dans les statistiques alors que son influence sur les résultats est déterminante, le raccourcissement de cette durée d'inactivité conduisant à augmenter le nombre de visites aux dépens du nombre de pages vues par visite.
La définition de la visite et du temps passé sur le site est aussi variable. Une des caractéristiques du comportement des internautes est de réaliser plusieurs actions en même temps tels que le copier/coller entre deux applications. Il est parfois difficile de savoir s'il poursuit sa recherche ou s'il l'abandonne. Ce cas de figure est pris en compte de manière différente par les instituts de sondage : certains détectent immédiatement le fait que l'internaute a arrêté son surf, bien qu'il soit toujours connecté au site, et stoppent le temps de comptage; d'autres ne détectent qu'un arrêt de la navigation et mettent donc 10 minutes avant de stopper le décompte du temps sur le site (Hussherr, 2002). Ces différents choix entraînent évidemment des nombres et des durées de visites différents.
Enfin, le choix de la période de mesure influence les mesures du trafic, notamment le reach et la fréquence (Lee, 1999). Ainsi, sur la base des 25 sites les plus visités pendant un mois, le reach augmente de 71% si la période de référence passe du jour à la semaine ou de la semaine au mois, la fréquence augmentant quant à elle de 55% si la période de référence n'est plus la semaine, mais le mois. En revanche, les auteurs démontrent qu'en utilisant un standard unique (reach ou fréquence), les classements des premiers sites ne sont pas modifiés par le choix de la période de référence. L'annonceur souhaitant communiquer en ligne devra donc s'interroger sur le critère de mesure d'audience le plus pertinent pour lui dans le choix de son support.

c) La fiabilité des mesures internes et la certification

Dans la plupart des cas, ce sont les sites eux-mêmes qui réalisent leur propre mesure d'audience, à partir des méthodes *site centric* exposées ci-dessus. Il leur est donc facile de gonfler artificiellement leurs chiffres (en comptant des hits et non des pages ou sous la forme de robots qui génèrent des requêtes automatiquement), afin d'influencer les annonceurs dans leur choix de support publicitaire et d'augmenter leur base de tarification. La certification par un tiers est donc un élément central dans la fiabilité des mesures présentées par un site.
La mesure d'audience des médias traditionnels s'est stabilisée dans le temps et bénéficie d'organismes certifiés qui servent de référence pour l'annonceur souhaitant connaître le rendement attendu de ses investissements publicitaires (Hong et Leckenby, 1996) : AC Nielsen pour la

télévision, Simmons Market Research Bureau (SMRB) et Mediamark Research, Inc (MRI) pour les magazines. De même, les magazines et les journaux possèdent des agences d'audit indépendantes qui vérifient la circulation de la presse.
Pour Internet, nouveau média, la certification de l'audience n'a pas encore atteint la même maturité. Un premier pas a été réalisé en mars 2001 par la publication par Diffusion Contrôle, l'association qui certifie la diffusion des médias presse sur le marché français, d'une liste des outils de mesure dits *site centric* auxquels il accorde son label. Cependant, le processus de certification sur Internet est encore loin d'être normalisé et généralisé, et ce déficit est souligné à la fois par les sites et par les annonceurs.

Les questions encore en suspens

Les limites évoquées jusqu'à présent reposent essentiellement sur la jeunesse du média qui se traduit par une absence de consensus sur les définitions et les périmètres de mesure. A terme, il est possible d'envisager une homogénéisation des concepts et donc un rapprochement des différents résultats. Cependant, avec ces problèmes de maturité coexistent des problèmes de nature durable, liés aux caractéristiques propres du schéma de communication sur Internet.

a) L'identification

L'identification des visiteurs uniques dans le cadre des mesures *site centric* constitue l'un des problèmes les plus difficiles à résoudre pour améliorer la précision des mesures. Identifier un visiteur à partir de son adresse IP peut conduire à plusieurs cas de figure, recensés par Drèze et Zufryden (1998) : un visiteur accédant à Internet depuis deux ordinateurs différents, un visiteur effectuant plusieurs requêtes traitées par des serveurs proxy différents, un visiteur se voyant attribuer des adresses IP différentes au cours de visites successives, une même adresse IP attribuée successivement à deux visiteurs différents. L'identification du visiteur au moyen des adresses IP conduira donc à des nombres de visites et de visites par visiteur unique erronés.
Afin d'estimer l'ampleur des erreurs liées à cette approximation, il est nécessaire de se placer sur un site permettant d'identifier les visiteurs avec certitude, au moyen d'un identifiant ou d'un mot de passe par exemple. Il est alors possible de comparer les résultats obtenus en termes de visites et de visiteurs à partir des identifiants et de leur approximation à travers les adresses IP. Cette étude menée par Drèze et Zufryden (1998) révèle ainsi un nombre moyen de 2,1 visiteurs par adresse IP, ce qui conduit à une sous-estimation du nombre de visiteurs (-39%) et de visites (-35%), et donc à une surestimation du nombre de pages par visite (+64%). Cette approximation conduit également à faire des erreurs sur les mesures d'efficacité publicitaire communiquées aux annonceurs. Il y aurait ainsi un écart positif de 25% en moyenne sur le *reach*, et un déficit moyen de 1,1% de la fréquence. Globalement, l'erreur moyenne sur le GRP (Gross Rating Point) est de 22,7%. Ces erreurs rendent plus difficile toute tentative de comparaison de l'efficacité publicitaire d'Internet vis-à-vis des médias traditionnels. Deux solutions se présentent pour éviter cette approximation. La première consiste à s'appuyer sur les profils déclaratifs notamment dans le cadre des panels ou de sites demandant un mot de passe pour accéder aux données. L'autre alternative consiste à passer par un nombre moyen d'utilisateurs par ordinateur. Cependant, ce nombre moyen d'utilisateurs par ordinateur, évalué aux alentours de 1,4 par Hussherr et Rosanvallon (2001), ne prend pas en compte, comme toute moyenne, la spécificité de la population des visiteurs de chaque site Internet.

b) Le nettoyage des connexions parasites

Les mesures *site centric* englobent toutes les requêtes effectuées sur les pages du site, y compris celles provenant d'administrateurs du site ou de robots. Le trafic interne exclu correspond à la consultation du site web effectuée en interne par l'éditeur et/ou par l'hébergeur du site web le plus souvent pour des motifs d'administration, de développement, ou de maintenance du site web. Le

trafic externe exclu relève quant à lui de la fréquentation générée par l'activité automatique d'exploration des robots (*spiders*) servant, par exemple, à l'indexation des pages pour les moteurs de recherche ou la pige publicitaire. Il convient alors de retirer dans la comptabilisation des visites ces requêtes. Dans les deux cas, cette opération s'appuie sur la définition d'une liste d'adresses IP et l'exclusion de l'activité générée à travers elles. Cependant, les listes ne sont jamais totalement exhaustives et conduisent inévitablement à des imprécisions de mesure.

c) La non prise en compte des consultations hors-site

Une caractéristique importante du schéma de communication sur Internet réside dans l'existence de caches et serveurs *proxy* destinés à réduire le flux des échanges d'information sur Internet. Lorsque la requête est traitée par le cache, elle n'est pas comptabilisée au niveau du site, ce qui conduit au phénomène connu sous le nom de " paradoxe de la mesure de fréquentation d'Internet " : plus une page est demandée, plus elle est stockée dans des mémoires intermédiaires, et donc moins elle sera comptée dans les mesures effectuées sur le site lui-même. Autrement dit, plus une page est demandée, plus elle a de chance de voir sa fréquentation sous-estimée (Brignier et al., 2002). Une solution employée par certains sites est d'introduire une petite image dynamique qui va forcer le serveur à se reconnecter sur le site pour rafraîchir l'image, introduisant ainsi une distorsion entre les sites par le gonflement du nombre de requêtes.

L'intégration des informations fournies par les deux approches

Les approches *site centric* et *user centric* répondent à des objectifs différents et fournissent des résultats qui ne sont pas directement comparables : les mesures *site centric* procurent des indications sur le volume de pages demandées, alors que les mesures *user centric* cernent les caractéristiques qualitatives de navigation d'un ensemble d'internautes. Du point de vue de l'annonceur, ces deux types d'informations sont importants dans le choix de son support publicitaire. Cependant, la réconciliation des résultats fournis par les deux méthodes se heurte à trois difficultés.

a) Le périmètre de l'audience

Les mesures sur site comptent l'ensemble du trafic d'où qu'il vienne, du pays d'origine du site ou de n'importe où dans le monde, quel que soit le lieu de connexion de l'internaute. Les panels en revanche ne s'intéressent généralement qu'au trafic d'un pays, ou au mieux à la somme des trafics des pays dans lesquels l'opérateur est présent. De plus, dans la majorité des cas, les panels ne représentent que la population connectée depuis son domicile, en omettant les autres contextes de connexion. Or, selon Jupiter MMXI, l'accès à Internet en pourcentage de la population internaute en France en 2001 se fait à 51% depuis le domicile, à 39% depuis d'autres lieux et à 30% depuis l'ordinateur au travail. Enfin, les opérateurs de panels produisent des résultats calculés sur un sous-échantillon des panélistes recrutés, à savoir ceux qui ont effectivement utilisé Internet au cours du mois ou des deux derniers mois. La précision des résultats s'en trouve affectée, puisqu'ils sont calculés sur un nombre d'individus inférieur à celui de l'échantillon total du panel.

Un travail de purification des données fournies par les outils *site centric* est donc indispensable pour concilier les deux périmètres. Ces ajustements peuvent être assez conséquents, comme le révèle l'étude de Kalyanam et McEvoy (1999). Ainsi, sur un total de 10 000 enregistrements d'un fichier *logs*, 56% seulement correspondent à des demandes réelles d'internautes nationaux. En effet, il a été nécessaire de supprimer les requêtes non abouties et, pour s'ajuster au périmètre des panels, les requêtes de pages provenant d'utilisateurs étrangers (21%), des doublons (17%), de l'activité des robots (0,2%) et des sites éducatifs (6%).

b) La méthode d'extrapolation des données de panel

Pour passer des informations issues des navigations des internautes du panel à des statistiques nationales, l'approche la plus fréquemment utilisée est de multiplier le pourcentage des panélistes ayant fréquenté le site par le nombre d'internautes en France. Pour les sites bénéficiant d'une forte fréquentation, les classements concordent généralement avec les volumes de trafic enregistrés par les outils *site centric*. En revanche, des différences apparaissent pour les sites plus marginaux. Cependant, les caractéristiques de la population des internautes sont encore mal connues aujourd'hui. C'est pourquoi, dans le cadre de la mesure de l'efficacité publicitaire notamment, les chercheurs (Drèze et al., 1999) ont suggéré d'extrapoler les données de panel, non pas à partir de caractéristiques socio-démographiques hypothétiques, mais à partir des comportements de navigation des panélistes et de l'ensemble des visiteurs du site.

Ce problème d'extrapolation des données de panels est soulevé par les professionnels eux-mêmes. Ceux-ci expriment leurs réticences à utiliser les services proposés par les sociétés d'études car ils relèvent d'importantes contradictions entre les données fournies par les panels et celles observées en interne avec leurs propres outils. Ces écarts peuvent provenir aussi bien de la constitution et du renouvellement de leur panel que des techniques d'extrapolation. Face à ces incertitudes, les professionnels se tournent de plus en plus vers des outils développés en interne, qui leur permettent d'avoir une mesure précise et fiable du nombre de visiteurs uniques et d'obtenir ainsi le taux de transformation, critère déterminant dans le management d'un site. Les sociétés de panels doivent donc à l'avenir intégrer ces problématiques spécifiques des sites marchands, afin de leur communiquer des données fiables et pertinentes.

c) L'objet mesuré

De par leur nature et objectif, les deux outils mesurent deux objets différents : les panels fournissent des résultats concernant le nombre d'internautes venus sur le site dans le mois, alors que les outils *site centric* communiquent des nombres de visites sur le site. Tant que les outils *site centric* ne sauront pas différencier avec exactitude les individus qui se connectent, ces deux grandeurs mesurées ne seront pas comparables. A moins d'envisager de comparer les chiffres générés par un autre outil de type site-centric, qui mesurerait lui aussi des visites, l'autre source d'information accessible à des tiers reste les chiffres des panels.

L'exemple du site Tchache.com est révélateur de cette difficulté à comparer des mesures qui n'appréhendent pas le même phénomène. Ainsi, pour le mois de février 2001, le site compte 5.268.348 visites dans le classement Cybermétrie, ce qui le situe dans les 10 premiers sites de ce classement juste devant le site Tf1.fr. Paradoxalement, il ne ressort pas dans les panels Netvalue et Nielsen NetRatings du même mois, ce qui signifie que moins de 1% des membres de ces panel ont visité le site au cours du mois considéré, soit moins de 80.000 visiteurs uniques [Journal du Net]. En effet, Cybermétrie mesure physiquement le nombre de visites enregistrées par le site, tandis que les chiffres des panels mesurent le nombre d'utilisateurs uniques chaque mois. Ainsi un internaute qui a visité trois fois le site Tchatche.com représente un visiteur pour les panels mais trois visites pour Cybermétrie.

Enfin, la notion de représentativité d'un panel d'internautes déjà abordée précédemment apparaît cruciale lors du rapprochement des données et classements issus des deux types de mesure. En effet, si les classements pour les principaux sites du marché concordent généralement avec les volumes de trafic enregistrés par les outils site-centric, la fiabilité des résultats est beaucoup plus aléatoire pour les sites moins connus.

Cependant, l'apport des panels aux mesures quantitatives issues de l'analyse des fichiers de *logs* est évident : les panels permettent de qualifier l'audience, de comprendre ses motivations, ses priorités, son mode de cheminement à travers le site. Ainsi, le gestionnaire de site est en mesure d'organiser

son site – contenu et design – en fonction de ses visiteurs et de sa cible, et l'annonceur peut choisir un support de communication adapté à sa cible.

Conclusion

Aujourd'hui, deux approches coexistent pour mesurer l'audience sur Internet : d'une part, les mesures *site centric*, qui utilisent une information secondaire issue de l'outil informatique (les fichiers log) et, d'autre part, les mesures *user centric*, qui sont basées sur une démarche de panels avec une précision statistique variable et des coûts d'acquisition élevés.

De nombreux obstacles s'opposent à la comparaison, voire à la conciliation, de ces deux méthodologies de mesure d'audience : la définition du périmètre étudié, le retraitement et certification des *logs*, la normalisation des concepts mesurés,.... Par ailleurs, Internet étant un support de communication mondial, les panels d'internautes prennent de plus en plus une dimension internationale, qui accroît la nécessité d'harmoniser les normes de mesure et de qualification d'audience.

Déjà recensés par Costes (1999), ces problèmes de mesure d'audience restent un frein au développement d'Internet comme support de communication publicitaire. Depuis cette date, des efforts ont été réalisés afin d'améliorer la fiabilité des chiffres communiqués, ainsi que d'établir des bases communes pour la mesure d'audience. Notamment, la certification de l'audience est plus fréquemment effectuée par des tiers extérieurs et peut s'appuyer sur des outils eux-mêmes labellisés par Diffusion Contrôle.

Cependant, au-delà de cette problématique de la mesure d'audience, les annonceurs désireux de communiquer en ligne doivent s'interroger sur l'efficacité de cette publicité en ligne. Dans la plupart des cas, la publicité en ligne est facturée au nombre d'impressions du message publicitaire (CPM). Un moyen de maximiser le retour sur investissement est donc d'améliorer le ciblage du message grâce à une meilleure connaissance de l'audience des sites, en termes qualitatifs (Hussherr, 1999). L'apport des panels est essentiel et complète les informations sur les volumes de trafic communiquées par les sites. La réconciliation des données des deux types de méthodes représente donc un enjeu pour la recherche future.

Par ailleurs, les coûts des messages publicitaires sur Internet sont élevés, d'où la nécessité pour les annonceurs de connaître les performances relatives de la communication sur ce nouveau média par rapport aux médias traditionnels. Les études empiriques sur l'efficacité publicitaire du média Internet constituent une voie de recherche importante. Dans ce contexte, les problématiques de mesure d'audience évoquées dans cet article prennent toute leur ampleur afin de déterminer le *reach* et la fréquence, indicateurs usuels de l'efficacité média (Leong et al., 1998). Enfin, les études sur l'impact relatif selon la forme de l'annonce et sa localisation dans le site (accueil ou en profondeur) constituent une deuxième piste de recherche pertinente pour les annonceurs souhaitant optimiser leur budget média (Shamdasani et al., 2001).

Bibliographie

Bhat S., Bevans M., Sengupta S. (2002), Measuring Users' Web Activity to Evaluate and Enhance Advertising Effectiveness, *Journal of Advertising*, 31, 3, 97-106.

Briggs R., Hollis N. (1997), Advertising on the Web: Is There Response Before Click-Through? , *Journal of Advertising Research*, 37, 2, 33-45.

Brignier J.-M., Chavenon H., Dupont-Ghestem F., Dussaix A.-M., Haering H. (2002), *Mesurer l'audience des medias : Du recueil de données au média-planning*, Paris, Dunod.

Costes Y (1999), La mesure d'audience sur Internet : Terminologie, technologie et méthodologie cahier de recherche N° 278, Université Paris-Dauphine, http://www.dmsp.dauphine.fr/dmsp/CahiersRecherche/CR278.pdf

Drèze X., Kalyanam K., Briggs R. (non daté), Increasing Panel Data Accuracy : An Application to Internet Panels, work in progress, http://www.xdreze.org/Publications/panels3.pdf

Drèze X., Zufryden F. (1998), Is Internet Advertising Ready for Prime Time ?, *Journal of Advertising Research,* 38, 3, 7-18.

Ferrandi J.-M., Boutin E. (1999), Un outil de mesure de l'audience d'un site Internet : L'analyse réseau, *Actes du 15ème Congrès International de l'Association Française du Marketing (AFM)*, Strasbourg, 669-696.

Galan J.-PH. (2002), L'analyse des fichiers log pour étudier l'impact de la musique sur le comportement des visiteurs d'un site Web culturel*, Actes du 18ème Congrès International de l'Association Française du Marketing (AFM)*, Lille, http://www.recherche-marketing.com/alt/20020523.pdf

Hong J., Leckenby J. (1996), Audience Measurement and Media Reach/Frequency Issues in Internet Advertising , *Proceedings of the American Academy of Advertising Annual Conference*,15-27,Vancouver, http://www.ciadvertising.org/resource/rf_models/drl_AAA97.pdf

Hussherr F-X, Rosanvallon J. (2001), *@-communication*, Paris, Dunod.

Hussherr F-X (1999), La publicité sur Internet : un modèle économique dépendant de l'efficacité publicitaire, Thèse de doctorat en Economie des systèmes d'information, Paris, ENST

Kalyanam K, Mac Evoy B. (1999), Data reconciliation : Reducing Discrepancies in Audience Estimates from web servers and Online panels , *IAB/ARF/FAST Summit Measurement Committees*, http://lsb.scu.edu/~kkalyanam/data-rec.pdf.pdf

Lee S., Leckenby J. (1999), Impact of Measurement Periods on Website Rankings and Traffic Estimation : A user-centric approach, *Journal of Current Issues and Research in Advertising*, 21, 2.

Leong E., Huang X., Stanners P. (1998), Comparing the Effectiveness of the Website with Traditional Media, *Journal of Advertising Research*, 38, 5, 44-51.

Shamdasani P., Stanaland A., Tan J. (2001), Location, location, location : Insights for Advertising Placement on the Web*, Journal of Advertising Research*, 41, 4, 7-21.

Shen F. (2002), Banner Advertising : Pricing, Measurement and Pretesting Practices : Perspectives from Interactive Agencies, *Journal of Advertising Research*, 31, 3, 59-68.

Wood L. (1998), Internet Ad Buys – What Reach and Frequency Do They Deliver ?, *Jo**urnal** of Advertising Research*, 38, 1, 21-28.

Les jeunes et Internet : Représentations, usages et appropriations (2001). Enquête menée en France par le Clemi, http://www.clemi.org/jeunes_internet.html