

Avantages de la mesure d'audience par marqueur distant : le Tag

XiTi agit comme tiers de confiance pour la mesure d'audience des sites web, grâce à sa technologie de marquage utilisant le Tag.

Le serveur qui distribue le marqueur XiTi étant indépendant du site hébergé, XiTi mesure l'audience d'une manière plus fiable que les fichiers LOG.

1. Définition des outils

Le fichier Log

Fichier texte où est enregistré l'historique des communications entre un serveur et des postes clients. On retrouvera en particulier les requêtes demandées au serveur, les messages d'erreurs générés par l'application.¹

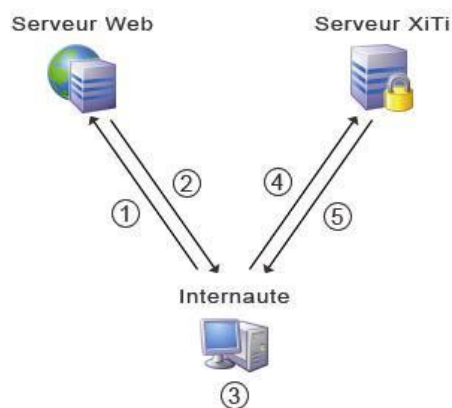
Le marqueur Distant : Tag

Un marqueur (*ou tag*) est un élément introduit dans chacune des pages à mesurer, pour témoigner de leur diffusion. Il est inséré dans le code source de la page. Il permet notamment de générer un journal de connexion sur le serveur de l'outil de mesure tiers.²

2. Technologie du marqueur distant

Les marqueurs ou « tags » XiTi s'insèrent uniquement sur le site, ce qui permet une étude quelque soit la plateforme d'accès (solutions fixes et mobiles).

Le logiciel de mesure des « logs » doit quant à lui s'installer sur tous les serveurs hébergeant le site web, ce qui alourdit le processus de récupération et d'analyse des données.



- (1) L'internaute demande une page du site web sur le serveur hébergeant le site.
- (2) Ce serveur renvoie le contenu de la page (images, textes, ... et le marqueur XiTi)
- (3) Le marqueur Javascript s'exécute sur l'ordinateur de l'internaute, récupérant quelques informations.
- (4) Le marqueur demande une image sur le serveur XiTi, passant en paramètres les informations.
- (5) Le serveur XiTi renvoie l'image demandée (1 pixel sur 1 pixel)

NB : aucune interaction entre le serveur Web et le serveur XiTi

¹ Définition du Journal du Net, www.journaldunet.com, ² Définition du CESP, www.cesp.org

3. Analyse des différences entre Logs et Tags

LOG	TAG
Installation, maintenance et coût	
L'installation d'un logiciel d'analyse de Log est nécessaire sur le serveur ou à distance avec récupération des Logs en FTP.	Aucune installation requise, seul un code Javascript est à rajouter sur les pages à auditer.
Maintenance du serveur hébergeant le logiciel, installation régulière des services packs et patches correctifs de bugs	Tout est géré par les équipes XiTi en totale transparence pour le site audité
Difficulté des mises à jour des bases de données de références (géolocalisation, liste des moteurs de recherche identifiés, webmails, navigateurs, systèmes d'exploitation, lecteurs RSS, etc.)	Tout est géré par les équipes XiTi en totale transparence pour le site audité
Evolutivité très limitée du logiciel : téléchargement de mises à jour ou achat de nouvelles licences régulièrement à chaque sortie de nouvelle version	Evolutivité importante car simplifiée. Aucune conséquence ni aucun changement pour le site audité qui profite immédiatement des améliorations. Aucun surcoût.
Achat ponctuel d'une licence d'un logiciel comportant peu d'évolutions. Achat à renouveler régulièrement. Problèmes de compatibilité, de réinstallation, etc.	Coût mensuel, principe locatif. Approche d'externalisation et contrôle budgétaire.

Forte dépendance au logiciel voire à l'hébergeur si le logiciel est fourni par celui-ci	Indépendance totale de l'hébergeur. Vision objective de la qualité d'hébergement.
Consultation des données	
Données statistiques difficilement exploitables et très techniques. Faible lisibilité et interprétation compliquée.	Consultation rapide et efficace à travers une interface web compatible tous navigateurs, mobile, rapports par emails.
Logiciel de reporting	Site Web sophistiqué (interface AJAX) permettant de calculer (sommés, moyennes, etc.), gérer les colonnes affichées dans les tableaux, choisir son type de graphique et les options d'affichage (comparaison, évolution, base 100, zoom, etc.), module de segmentation (technologie Cube)
Mémoires caches et Proxies	
20% du volume des pages consultées est mis en mémoire dans le proxy ou le cache du navigateur.	Analyse de toutes les informations incluant celles contenues dans le proxy ou le cache du navigateur.

Méthodologies

Analyse des adresses IP de connexion. Prise en considération des Firewalls, un groupe de personnes se connectant avec une même adresse IP est considéré comme un unique visiteur.	Les Firewalls ne sont pas pris en compte. Analyse des adresses IP et utilisation des cookies installés dans chaque navigateur
L'utilisation d'adresses IP dynamiques peut fausser le calcul des données visiteurs.	En plaçant un cookie dans le navigateur du visiteur, les internautes qui utilisent une même adresse IP seront considérés comme des visiteurs distincts.
Chaque appel serveur (image, frame, Popup...) est considéré comme une page.	Le tag s'insère dans la partie à contenu seulement, ainsi la page est comptabilisée précisément.

Supports et pages dynamiques

Mesure peu adaptée aux sites intégrant des supports interactifs comme le Flash, Streaming Vidéo, Wap...	Flexibilité et capacité d'adaptation à tous les types de médias ainsi qu'aux pages sécurisées.
Les pages ne sont reconnaissables que par leur URL.	Possibilité de nommer dynamiquement les pages avec des noms beaucoup plus parlant.
Handicap lourd pour les pages dynamiques (ASP, PHP, etc.) qui portent toutes le même nom de fichier.	Possibilité de différencier automatiquement des pages différentes pourtant générées par le même fichier dynamique.
Mesure des emails impossible car l'email ne déclenche aucun hit	Possibilité de placer un marqueur dans les emails au format HTML.

dans les Logs du serveur	
Réseau de sites	
Complexité d'analyse et de récupération de données grandissantes en fonction du volume audité. En cas de volume très important, impossible de répartir la charge sur plusieurs serveurs.	Système centralisé dimensionné pour gérer des milliards de pages par mois. La croissance du site est donc sans conséquence sur la mesure, les calculs, la restitution des rapports.
Difficulté d'analyse des sites basés sur différents systèmes d'exploitation, des clusters de serveurs ou encore des serveurs hébergés dans différents endroits.	Tous les supports quelque soit l'architecture réseau peuvent être supportés.
Impossible d'obtenir des informations globales sur l'ensemble des sites d'un même groupe.	Information mutualisée ou consolidée pour un groupe de sites. Analyse transversale disponible avec le marqueur Tag.
Robots et audience indésirable	
Exclusion des postes internes à l'entreprise très difficile à mettre en œuvre et impossible par cookies.	Possibilité très simple d'ignorer des postes par simple clic pour identification ou par adresses IP (classes complètes et sous-réseaux)
Surévaluation des données par la prise en compte des outils de monitoring, des logiciels d'aspiration et des robots d'indexation (crawlers).	Le code du marqueur empêche le calcul des visites réalisées par certains crawlers. Des bases de données de référence complètent cette distinction. Une analyse des comportements de navigation permet d'exclure les derniers robots non référencés et non ignorés par défaut.

4. XiTi labellisé Diffusion Contrôle

XiTi 7.0 a reçu le label OJD exploitable en vue de la certification de la fréquentation des sites Web.

Ce label signifie que XiTi 7.0 correspond aux quatre critères suivants :

- Pratiquer une mesure sur la base de la technique du marqueur (ou "tag")
- Se présenter comme "tiers mesureur"
- Appliquer strictement la définition internationale de la visite
- Surmonter les différents "biais" techniques inhérents à ce type de mesure (contournement des caches, translation d'adresses, interruption inopinée des navigateurs...)



Pour bénéficier de l'Option Certification OJD, nous vous invitons à [contacter un de nos conseillers](#).

5. Cas pratique

Prenons un exemple concret faisant la démonstration des différences qui résultent de l'utilisation de TAG ou de LOG. Un de nos clients a constaté des différences entre les chiffres fournis par un analyseur de fichiers LOG et ceux fournis par XiTi. Pour ce cas précis, seule la page d'accueil du site a été prise en compte.

La différence de chiffres peut être expliquée par 3 raisons :

- Les robots et aspirateurs détectés par leur USER AGENT (navigateur)
- Les aspirateurs détectés par comportements louches
- Les doublons ou les triplons enregistrés plusieurs fois et très rapidement sur la page d'accueil

Les robots et aspirateurs détectés par leur USER AGENT

Au total, 1220 hits ont été comptabilisés.

Sur l'URL <http://www.topix.net/topix/newsfeeds> on a dénombré 24 hits pour cet aspirateur de contenu qui ne charge pas le tag.

- GoogleBot : 20 hits comptabilisés
- Yahoo! Slurp : 17 hits comptabilisés
- VoilaBot : 1 hit comptabilisé
- West+Wind+Internet : 14 hits comptabilisés pour un environnement de développement (Visual FoxPro) pour aspirer des sources de pages
- MSNbot : 2 hits comptabilisés
- BecomeBot : 1 hit comptabilisé
- CFNetWork : 2 hits comptabilisés
- Ask Jeeves Bot : 2 hits comptabilisés
- eZ+publish+Link+Validator : 12 hits comptabilisés pour ce validateur de liens
- FindLinks : 4 hits comptabilisés

On trouve alors 92 hits (chargements de la page d'accueil) résultants de robots ou d'aspirateurs.

Les aspirateurs détectés par comportements louches

Le second cas est illustré par une adresse IP dont proviennent de nombreux hits.

On a noté deux séries d'aspirations par la même adresse IP dont chaque hit est espacé de 2 secondes du suivant :

- 2006-04-20 08:28:29 xx.xx.xx.xx Mozilla GET /Page/Home.htm - 80 - xxx.xx.x.xxx Mozilla Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+FunWebProducts) - - 200
- 2006-04-20 08:28:31 xx.xx.xx.xx GET /Page/Home.htm - 80 - xxx.xx.x.xxx Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+FunWebProducts) - - 200
- Etc.

19 hits pour la même adresse IP, générés par Google Desktop et séparés de 5 min à chaque fois ont également pu être observés :

Exemple d'heure de Hit :

- 16:50:08
- 16:55:14
- 17:00:21
- 17:05:42
- 17:10:49
- 17:16:09
- 17:21:30

Plusieurs doublons ou triplons pour la page d'accueil

Dans le troisième cas, nous avons retenu toutes les pages d'accueil rechargées par la même adresse IP après moins de 4 secondes.

Cela donne un total de 33 hits à ignorer car le premier chargement de la page d'accueil ne peut avoir le temps d'aboutir. C'est donc bien une demande serveur (notifiée par le fichier LOG) mais qui ne se transformera a priori jamais en affichage de page chez le client puisque celui-ci redemande immédiatement la même page, avant même que la première demande ait abouti. Ces hits sont en fait générés par le visiteur, souvent de manière involontaire.

Voilà les cas les plus fréquents :

- - Double-clic sur un lien (c'est un réflexe pour beaucoup de personnes).
- - Nouveau clic sur un lien une ou deux secondes après avoir cliqué lorsque rien ne se passe (lenteur du réseau) et qu'on doute alors avoir vraiment cliqué la première fois.
- - Rechargement de la page (F5) avant le chargement complet de celle-ci.

A ce stade, nous arrivons à 888 hits conservés (contre 892 chargements pour XiTi) qui contiennent sans doute encore des aspirateurs ou des robots non identifiables facilement.

Le système par marqueurs tel que celui utilisé par XiTi mesure donc plus de visiteurs que le fichier LOG, contrairement aux apparences et suppositions initiales.

Pour mettre en évidence cela, nous avons pris toutes les adresses IP du fichier LOG fourni (robots compris). Nous avons ensuite pris toutes les adresses IP mesurées par le TAG XiTi sur la page d'accueil du site. En croisant les deux données, on trouve 60 adresses IP mesurées par XiTi et non présentes dans le fichier LOG.

Effectuons maintenant un calcul approximatif afin d'apprécier la différence de mesure des deux systèmes :

- En regardant dans XiTi on compte 556 visiteurs (dont 556 IP différentes) sur la page d'accueil (pour les 892 chargements). 60 adresses IP représentent donc plus de 10% d'écart dans la mesure des visiteurs en faveur de XiTi. On a donc vraisemblablement ce même ratio pour les chargements de pages.
- En continuant l'estimation, l'analyseur de fichiers LOG raterait environ 10% (892) = 90 hits. Sur 1220 hits on aurait alors :
 - 1220 (total) - 892 (si comme XiTi) + 90 = 418 hits de robots (soit un total de 33% de hits ne représentant pas des visiteurs réels). Et seulement : 1220 - 418 = 802 hits réels sur la page d'accueil.

En conclusion :

Un fichier LOG semblant montrer que plus de hits sont enregistrés que le système TAG (XiTi) peut en fait mesurer moins d'informations sur les visiteurs réels du site. En effet, le système TAG basé sur du JavaScript parvient à déjouer les caches et autres robots plus facilement que les LOG. Paradoxalement, plus le site est gros, plus il y a d'effet cache et par conséquent les LOG doivent manquer de nombreuses pages. Et dans le cas des petits sites, le volume des robots (indexation des moteurs par exemple) peut rapidement représenter une part importante du volume total du site, et donc fausser de manière notable l'audience mesurée par l'analyseur de LOG.



www.xiti.com/Contact.aspx - Service Client : 0 825 06 94 84

XiTi.com est un service de la société AT Internet – Siège social : 85 avenue JF Kennedy – 33700 Mérignac – France – RCS Bordeaux B 403 261 258. Les marques et logos figurant dans ce document sont des marques enregistrées ou non appartenant à la société AT Internet ou à des tiers. Toute utilisation, non autorisée explicitement par les titulaires des marques précitées, est strictement interdite. Toute reproduction partielle ou totale de ce document, sans autorisation expresse d'AT Internet est interdite. AT Internet se réserve le droit de mettre à jour le présent document à tout moment et sans préavis. Document et informations non contractuels. © AT Internet – 2008

DE.S.4-00000369 – v.1.0 (mise à jour 05/05/2008)