



## **Comparaison Entre la technologie « TAG » et l'analyse de fichiers logs**

**Laurent Patureau**  
Co-fondateur d'**IDfr**  
Editeur de **Wysistat**  
16, Boulevard Winston CHURCHILL  
25 000 BESANCON

Tel : 03 81 48 03 05  
Fax : 03 81 48 04 83  
Mail : [contact@wysistat.com](mailto:contact@wysistat.com)  
URL : <http://www.wysistat.com>





## I. Les technologies

### **I.1. L'analyse de fichiers log**

L'analyse de fichiers log consiste à analyser les fichiers que le serveur Internet (IIS, Apache... ) remplit à chaque demande d'un élément : une page html, une image, une animation flash...

Techniquement, l'internaute demande une page au serveur qui héberge le site. Le serveur Web au cours du processus qui lui permet de délivrer l'élément demandé va « logger » dans un fichier au format texte les éléments de la demande comme :

- L'IP de l'internaute qui fait la requête
- L'heure de la requête
- Le protocole de la requête (http, https... )
- Si la requête est un get (demande d'un élément) et un post (envoi d'information)
- La requête elle-même : /dossier/fichier.html
- La conclusion du traitement, par un code : 200 pour élément fourni, 404 pour non disponible...
- Le poids de l'élément envoyé
- La signature du navigateur (cette information permet d'identifier le navigateur, le système d'exploitation).
- La provenance de l'internaute : la requête n-1

Exemple de ligne :

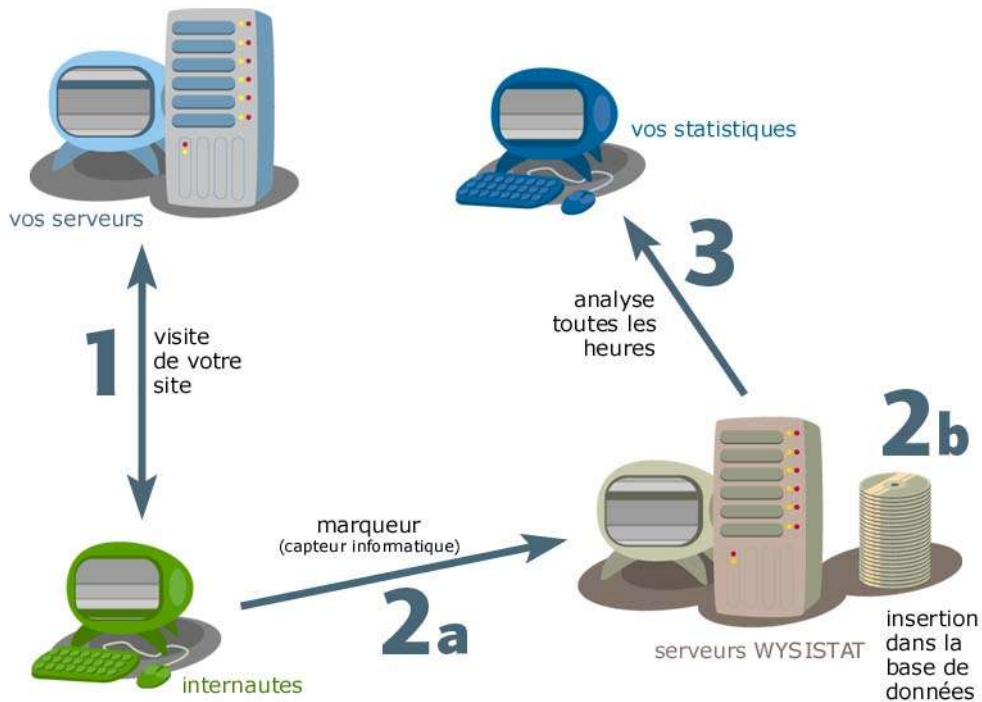
```
194.57.11.165 - - [20/Oct/2003:09:59:02 +0200] "GET /marche_public/styles.css
HTTP/1.0" 200 3372
"http://www.idfr.net/marche_public/index.php?keyword=ticket+cinema&restriction="
"Mozilla/4.0 (compatible; MSIE 4.01; Windows 98; Hotbar 4.0)"
```

L'analyse de ces fichiers permet, grâce aux informations que le serveur HTTP indique, de suivre l'ensemble des requêtes émises sur le serveur HTTP. En se basant sur l'adresse IP, on peut faire un regroupement par session de consultation. Cette notion permet de se rapprocher de la notion de visite utilisée par les outils de mesure d'audience.

Toutes les requêtes sont renseignées dans ces fichiers. On donne souvent le nom de Hits (demande unitaire) à ces requêtes. On dit ainsi que l'analyse de fichiers log est basée sur l'analyse des hits.

## **I.1. La technologies des marqueurs TAG**

La technologies des marqueurs TAG se base sur l'utilisation de script positionnés au sein des pages à auditer. Ces marqueurs vont déclencher le chargement d'une image qui permettra de comptabiliser les visites, les pages vues et les autres données.



*Principe de fonctionnement de la technologies des marqueurs TAG*

Cette technologie est dite tiers car elle utilise, en plus du couple serveur d'hébergement / Internaute, un troisième serveur dit tiers qui assure la comptabilisation des données.

Grâce à l'emploi du javascript, les données disponibles au niveau du serveur tiers sont plus complètes que celles renseignées par le serveur HTTP de l'hébergement. Le serveur de comptage insert ces informations dans une base de données et analyse soit en direct, soit à fréquence régulière les données (toutes les heures, tous les jours... ).

## **II. Comparatif**

### **II.1. Comparatif technique :**

#### **Installation, mise en œuvre :**

La mise en œuvre d'une solution d'analyse de fichiers log nécessite l'installation sur le serveur d'hébergement d'un programme qui va faire le traitement. Cela augmente les fonctions que le serveur doit remplir. L'analyse des fichiers log peut être assez longue et nécessite beaucoup de ressources processeur et de RAM pour des sites à audience moyenne (à partir de 100 000 pages vues par mois).

La mise en œuvre d'une solution de marqueur TAG ne nécessite pas d'installation sur le serveur, à l'exception des marqueurs à positionner sur les pages. Certain outil comme Wysistat utilise un même marqueur pour toutes les pages d'un site ce qui facilite le marquage des sites.

#### **Hébergement sur plusieurs serveurs :**

Lorsque le site est hébergé sur plusieurs serveurs, l'analyse de fichiers log n'est pas possible à moins de regrouper les fichiers de log en un seul.

Les solutions de marqueur TAG sont très efficaces dans ce cas-là car l'audience est directement regroupée sur le serveur tiers donc ce type d'architecture système ne vient pas perturber le comptage.

Un bon exemple est la technologie *Akamai* qui démultiplie les centres d'hébergement et qui rend ainsi caduque l'analyse de fichiers log du serveur source alors que les solutions de marqueurs TAG sont compatibles avec cette technologie.

De plus un changement d'hébergement ne changera rien si vous utilisez une solution de marqueurs TAG alors qu'il entraînera généralement la perte de l'historique avec une analyse de fichiers log.

#### **Evolution des solutions :**

Lors d'une mise à jour d'une solution d'analyse de fichiers log, il est nécessaire de mettre à jour le programme installé sur le serveur : opération parfois délicate.

La technologie des marqueurs TAG centralise le système et assure des mises régulières par le prestataire de service sans intervention de la part du client.

#### **Evolution des moteurs de recherche :**

Régulièrement, les moteurs de recherche changent leur méthode d'interrogation.

Avec un programme d'analyse de fichiers log, il est nécessaire de mettre un Update régulier ce qui est rarement le cas.

La technologie des marqueurs TAG permet de s'affranchir de cette étape car le prestataire maintient à jour ces données et assure une veille à la place de ses clients.

#### **Problème de cache :**

Lorsqu'un internaute fait une requête depuis son poste de travail, la requête n'arrive pas forcément au serveur qui héberge le site. Deux éléments, que sont le cache navigateur et les caches proxys, peuvent répondre à la demande de l'internaute sans avoir besoin de faire une demande au serveur d'hébergement. Ainsi la requête n'est pas transmise au serveur qui ne va donc pas avoir connaissance de la demande. Cette requête ne sera donc pas prise en compte dans l'analyse.

La technologie des marqueurs TAG force la connexion au serveur tiers, même à travers le cache d'un proxy ou le cache du navigateur assurant ainsi la prise en compte de l'ensemble des visites et des pages vues.

### **Topologie des sites Internet :**

L'analyse de fichiers log correspond mal aux sites dynamiques dont un fichier peut en réalité produire plusieurs pages différentes selon les paramètres passés lors de son appel. A l'extrême, certains sites n'ont qu'un seul fichier, l'analyse de la navigation est alors limitée avec l'analyse de fichiers log car ils ne voient qu'un seul fichier.

La technologie des marqueurs TAG permet de différencier les pages réelles, en tenant compte des paramètres passés au fichier. Cela permet de s'adapter à toutes les technologies dynamiques.

Par ailleurs, un cas mettant en faute l'analyse de fichiers log est le cas d'un site constitué d'un seul fichier Flash qui comporte nombre de séquences. Seule la technologie des marqueurs TAG permet de voir où vont les internautes au sein même du Flash. En effet, des outils comme Wysistat ont des solutions pour voir où l'internaute clique et qu'elles sont les parties du Flash qui ont été vues.

## **II.2. Comparatif au niveau des données :**

### **Désactivation d'adresse IP :**

La technologie des marqueurs TAG permet d'éliminer des adresses IP du comptage afin de ne pas fausser les résultats, l'IP de la société et du prestataire qui développe le site par exemple. Cela permet de voir l'audience réelle sans tenir compte de l'audience dite « interne » qui peut, dans certain cas, représenter plus de 50% de l'audience globale du site.

### **Spiders des moteurs de recherche :**

Les « spiders » des moteurs de recherche (robots qui scrutent les sites Internet) sont au même titre qu'un internaute pris en compte au niveau des fichiers log. La technologie des marqueurs TAG permet d'éliminer ces robots automatiques qui peuvent générer à eux seuls une grande part de l'audience.

### **Nationalités :**

Pour l'analyse des nationalités des internautes, l'analyse de fichiers log n'a à sa disposition que l'adresse IP de l'internaute et un resolve de cette dernière. Cette donnée ne permet d'établir la nationalité que sur 50 % des internautes, (le reste donnant une extension en .com, .net, ...). La technologie des marqueurs TAG est plus puissante à ce niveau-là car les services ont leur propre base d'adresse IP qui leur assure une haute qualité dans la résolution de la nationalité.

### **Notion de visites vs notion de sessions**

L'analyse de fichiers log va calculer un nombre de sessions (appelé par abus de langage visite) à partir d'un regroupement basé sur l'adresse IP. Pour les entreprises utilisant un proxy en sortie, l'ensemble des requêtes semblent alors provenir d'un même ordinateur (le proxy) car la seule adresse IP vu par le serveur est celle du proxy. Ainsi, l'analyseur de fichiers log va regrouper en une seule visite un ensemble de visites distinctes.

La technologie des marqueurs TAG, utilisant les cookies, permet d'identifier chaque poste, même à travers un proxy. Dans le cas défavorable où le cookie n'est pas accepté (moins de 2%), l'heuristique IP/Browser est alors employée ce qui produit un résultat similaire à l'analyse de fichiers log, mais seulement pour 2% des visites.

### **Notion de pages vues vs notion de hits**

L'analyse de fichiers log a comme données à sa disposition les hits demandés à la machine (toute requête faite au serveur HTTP, que ce soit demander une page, une image, un fichier css, js, une animation Flash, ...). Le programme doit alors extraire de ce flot de hits les « pages » au sens de pages de contenu. Cette étape qui peut sembler simple est en réalité assez délicate car de nombreuses problématiques se présentent.

En effet, le cas des frames va automatiquement multiplier le nombre de pages (un écran avec une frame est constitué de 3 fichiers donnant donc 3 pages vues pour l'analyse de fichiers log). Les éléments Flash sont souvent comptés comme des pages or il s'agit bien souvent d'images animées constituant une page.

La technologie des marqueurs TAG permet de définir simplement et précisément la notion de page en n'insérant un marqueur que dans les pages de contenu. Cette technologie est basée sur la notion de page vue.