

# Mesurer le temps de chargement d'un site Web, en ignorant la pub ?

Stéphane Bortzmeyer  
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 8 octobre 2012

<https://www.bortzmeyer.org/mesurer-web-sans-pub.html>

---

Je suis sûr qu'une partie de mes lecteurs vont apprécier le défi. Le problème a été discuté dans un groupe de travail ARCEP, où il s'agit de mesurer la qualité de l'accès à l'Internet via un FAI. Outre les mesures de bas niveau (les seules sérieuses, à mon avis <<https://www.bortzmeyer.org/qos-ou-qoe.html>>), on voudrait mesurer un truc plus représentatif de l'utilisation habituelle de l'Internet, l'accès à une page Web. Mais les pages Web « importantes » contiennent souvent des tas d'éléments « extérieurs » comme Google Analytics, les publicités, etc. Comment les exclure ?

Je me focalise sur l'aspect technique. Sur la question de fond, il faut se demander si c'est une bonne idée d'exclure quelque chose qui fait partie intégrante du vécu de l'utilisateur...

Le premier algorithme qui vient à l'esprit est d'exclure les contenus externes qui sont désignés par un nom de domaine différent. Prenons la page d'accueil de TF1, <<http://www.tf1.fr/>>. Elle référence 31 noms de domaines différents, dont 19 sont en dehors de `tf1.fr` (par exemple `w.estat.com`, utilisé pour des statistiques). En examinant ces noms à la main, les choses semblent simples : les noms externes à `tf1.fr` sont bien pour du contenu « externe », dont l'affichage n'est pas indispensable.

Maintenant, testons un autre média, Libération. La page <<http://www.liberation.fr/>> contient 62 noms de domaine (qui peuvent être des liens à suivre, pas forcément du contenu chargé automatiquement). Mais, cette fois, le problème est que du contenu de la page est chargé à partir de noms en dehors de `liberation.fr` comme `s0.libe.com`. Oui, `libe.com` a le même titulaire que `liberation.fr` mais allez expliquer cela à un logiciel.

À part des problèmes comme celui-ci, l'examen rapide de quelques sites Web français populaires (bien placés dans Alexa) semble quand même indiquer un gros effort des webmestres pour placer le contenu sous **leur** nom de domaine. On trouve nettement moins de noms de domaine appartenant à des CDN, par exemple, alors que c'était très courant à une époque. Ce premier algorithme ne semble donc pas catastrophique mais des différences comme `libe.com` vs. `liberation.fr` le prennent en défaut.

Deuxième algorithme possible, utiliser les listes noires de logiciels comme Ghostery ou Adblock Plus pour éliminer le contenu « externe ».

Et y a-t-il d'autres possibilités ? Vous pouvez indiquer vos suggestions sur SeenThis <<http://seenthis.net/messages/90963>>.