

Comment bien utiliser les flux RSS et les agents de surveillance

Un texte rédigé en 2008 pour l'ouvrage Netrecherche de Véronique Mesguich et Armelle Thomas

Il y a quelques années de cela la veille Internet était presque entièrement dévolue à des solutions (logiciels et services en ligne) permettant de surveiller des pages web et vous alertant, par un email ou un signal visuel sur votre bureau, lorsque l'une d'elles venait à changer. A ces solutions vient maintenant s'ajouter un autre moyen, plus "élégant", de surveiller le web : les flux RSS. Plus élégant parce les métadonnées qu'ils embarquent formatent ces flux afin de leur assurer la meilleure diffusion possible : cela veut dire pas de surlignage artisanal de mots nouveaux ou de problème avec la datation automatique d'une page, la date de publication fait en effet partie de ces métadonnées intégrées par défaut au flux.

Impressionnant dans les possibilités de mixage et de filtrage de l'information qu'il offre, le RSS est toutefois insuffisant lorsqu'il s'agit de mener une veille à 360 degrés. Le problème vient du fait qu'il ne permet de surveiller que ce qui est "autorisé" par l'auteur du flux d'informations diffusé. Or lorsque l'on mène une veille avancée sur un concurrent on aura besoin de faire feu de tout bois. Ainsi tout ce qui change sur son site web aura de l'intérêt et sera susceptible d'être interprété : charte graphique, plan du site, ajout de vidéos, de photos,... Autant de signaux faibles qui, agrégés à d'autres pourront aider à mieux comprendre la stratégie d'un concurrent.

Quelles solutions pour quels usages?

Il faut donc se résoudre à utiliser en parallèle deux solutions, les agrégateurs de flux RSS que l'on réservera au traitement et au filtrage de gros volumes de données de type "news" et billets de blogs, et les agents de surveillance qui permettront de se focaliser sur tel ou tel site web de concurrents, fournisseurs,...

Une fois encore le choix des bons outils dépendra des objectifs que vous vous êtes fixés. Le schéma ci-dessous a pour but de vous aider à faire ces choix.

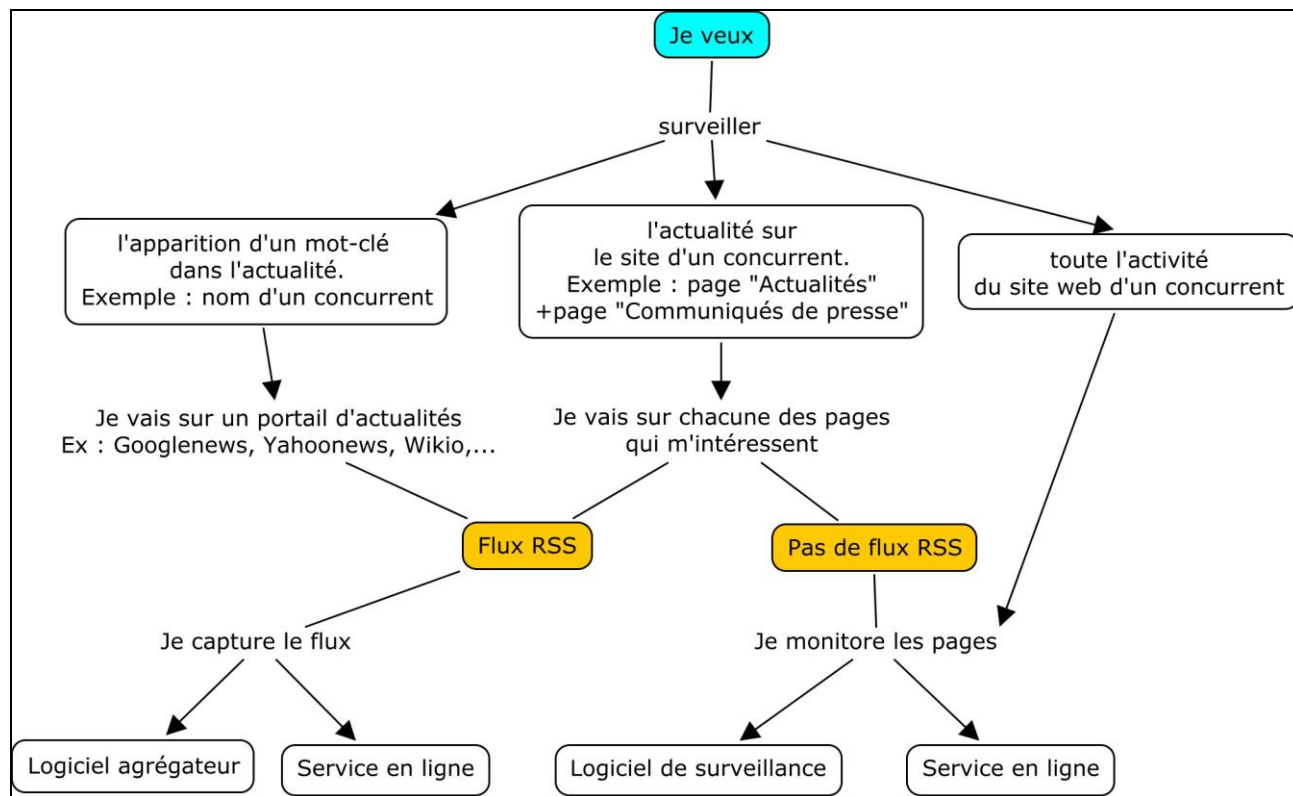


Schéma 1 : Choisir un type d'outil en fonction de son objectif

Que l'on ait besoin d'un agrégateur de flux RSS ou d'un agent de surveillance on voit sur ce schéma qu'il y a toujours un second choix à effectuer : logiciel à installer sur son poste ou sur un serveur ou service en ligne accessible après identification. Chacun dispose de ses propres avantages et inconvénients :

Logiciels à installer	Services en ligne
Avantages : • Possibilité de stocker les pages modifiées afin d'analyser leurs changements sur la longue durée ou de pratiquer des opérations de traitement automatique du langage (TAL). • Visualisation des changements par surlignage • Création de rapports de veille.	Avantages : • Pas de problèmes d'espace-disque. • Accès à partir de n'importe quelle connexion Internet • Surveillance automatisée ne dépendant plus de votre connexion ou de l'extinction de la machine. • Protègent mieux votre anonymat.
Inconvénients : • Peuvent rapidement générer beaucoup de volume. • Nécessitent une connexion permanente et un ordinateur toujours allumé pour le lancement automatique des surveillances.	Inconvénients : • Peu de fonctionnalités avancées dans ces services (sauf Trackengine). • Envoi d'une notification de modification de page mais pas d'une copie de page surlignée (sauf Trackengine).

Tableau 1 : avantages comparés des logiciels et services en ligne

Il faut insister sur l'importance du stockage des pages modifiées sur un serveur local car c'est à partir d'elles que l'on sera en mesure de discerner d'éventuels changements de stratégie. Par ailleurs cela rendra possible la réalisation d'opérations de traitement automatique du langage sur le corpus constitué (catégorisation, émergence de thématiques, détection de tendances,...).

Mettre en place la surveillance concrètement

Une fois que nous savons quels types de surveillance nous avons à mener et avec quelles familles d'outils, comment s'y prendre concrètement ? Les tableaux étape/action qui suivent vous donnent des modes d'emploi génériques. Pour les agents de surveillance la procédure est la suivante :

Etape	Action
1.	Copier l'url d'une page à surveiller dans votre presse-papier (CTRL+C)
2.	Collez-la dans la fenêtre de votre logiciel ou service de surveillance en ligne réservée à cet effet (CTRL+C)
3.	Réglez les options de surveillance : a. Périodicité b. Présence conditionnelle de certains mots-clés c. Gestion des logins et mots de passe d. Gestion des images
4.	Choisissez les options d'envoi des alertes a. Notification par email b. Périodicité des envois d'alertes

Tableau 2 : Mode d'emploi générique pour la mise sous surveillance d'une page web avec un agent de surveillance.

Pour les agrégateurs la procédure est très proche :

Etape	Action
1.	Recherchez le logo (XML ou RSS) du flux RSS de la page qui vous intéresse.
2.	Copiez l'adresse du lien correspondant (commande accessible d'un clic droit sur le logo).

3.	Intégrez-la dans votre agrégateur (Collez).
4.	Réglez les options de surveillance des flux qui vous sont proposées par l'outil que vous utilisez : a. Périodicité b. Gestion des logins et mots de passe,... c. Présence conditionnelle de certains mots-clés
5.	Gérez la présentation de vos fils RSS. L'information étant déjà structurée dans des balises, vous pouvez jouer très facilement sur sa mise en page dans votre agrégateur afin d'en faciliter la lecture.
6.	Vous serez désormais informé de chaque nouvel item (article, billet) entrant dans le flux RSS

Tableau 3 : Mode d'emploi générique pour la mise sous surveillance d'un flux RSS avec un agrégateur.

NB : si vous mettez sous surveillance le flux RSS d'une requête dans un portail d'actualités comme GoogleNews il est important d'effectuer auparavant un tri par dates des résultats. Cela vous permettra de toujours voir apparaître les articles les plus récents en premier dans votre agrégateur ; essentiel pour un veilleur.

Les outils

Famille	Nom	Prix	Adresse	Commentaire
Agents de surveillance	Website Watcher	A partir de 29,95 €	www.aignes.com	Une référence depuis des années. Pour des besoins de veille avancés.
	Diffbot	Gratuit	www.diffbot.com	Extension gratuite pour Firefox. Fonctionnalités limitées mais outil pratique.
Agents de surveillance en ligne	Trackengine	De 20 à 60\$/an	www.trackengine.com	Très nombreuses fonctionnalités. Plus puissant que de nombreux logiciels. Envoie les pages modifiées avec surlignage des changements.
	ChangeDetect	De 2\$ à 40\$/mois	www.changedetect.com	Fonctionnalités intéressantes : surveillance de fichiers .doc ou .pdf., accès aux sites nécessitant une identification, filtrage par expressions ou mots-clés.
Logiciels d'agrégation de flux RSS	RSSbandit	Gratuit (Open source)	www.rssbandit.org	Permet notamment un filtrage des flux entrants par mots-clés.

	Feedreader	Gratuit (Open source)	www.feedreader.com	Nombreuses possibilités de filtrage et de recherche dans les flux entrants.
	FeedException	Gratuit (freeware)	www.newsgator.com	Aussi puissant que le précédent. Se synchronise avec l'agrégateur en ligne Newsgator.
Services d'agrégation de flux RSS	Netvibes	Gratuit	www.netvibes.com	Service très facile à utiliser qui permet de se créer un portail d'actualité (et bien plus).
	Google Reader	Gratuit	www.google.com/reader	Très nombreuses fonctionnalités dont partage et consultation en mode déconnecté