Opinions Libres
le blog d'Olivier Ezratty

Moissonner le web avec Outwit

Le web, ses sites et les moteurs de recherche ont une particularité : ils présentent des tonnes d'information, mais en général, faiblement structurées.

La couche de présentation "web/HTML" détruit la structure originelle des informations, très souvent stockées dans des bases de données. Il en résulte des silos de données disparates difficiles à exploiter. Pourtant, des trésors d'informations sont disponibles qui pourraient être mieux exploitées : données financières, listes diverses, comparaisons de prix, listes d'objets à vendre, etc.

Cela fait des années que les chercheurs et spécialistes des standards du web cherchent une réponse. Elle s'appelle le plus souvent "web sémantique", voire "web services". Le "**web sémantique**" consiste à publier les données sur le web avec des informations sur leur structure, leur sens, permettant une exploitation intelligente par les logiciels. Il s'appuie notamment sur le standard W3C "RDF" qui permet de spécifier la structure des données et leur sémantique (tel champ est un nom, tel autre est une adresse, etc). Le seul hic, c'est que ce standard n'est pas adopté et que le web reste un bazar toujours faiblement structuré au niveau des données publiées.

Les "services web" permettant quant à eux d'interroger les sites à partir de logiciels, comme si on interrogeait une base de données (pour faire simple), ne se sont pas plus généralisés. Les moteurs de recherche comme Google Search fonctionnent de leur côté en utilisant des batteries de statistiques, et pas du tout par sémantique.

Résultat, le web sémantique est pour l'instant dans les limbes. Il est à peine utilisé au sein des entreprises et très rarement sur Internet. Les logiciels en sont réduits à deviner la structure des informations publiées pour les exploiter. On pourrait appeler cela le "web sémantique implicite".

C'est pour cela qu'il existe plein d'outils souvent en Shareware pour aller grappiller des informations structurées sur les sites web. On appelle cela des "scraper", ou bien des outils de téléchargement pour ce qui est des images. Mais il est difficile de trouver des outils de ce genre qui soient à la fois faciles d'emploi, flexibles et extensibles.

Répondre à ce besoin est la mission d'Outwit. Avec l'idée de faciliter la récupération d'informations structurées de sites web qui… la déstructurent !

L'équipe

Je connais **Jean-Christophe Combaz** depuis deux ans. Diplômé de la Harvard Business School, c'est un entrepreneur en herbe au long parcours. Il a créé **Calliscope** en 1987 et l'a dirigé en France et en Californie. Il a créé ensuite la filiale européenne de l'éditeur de logiciels de conception 3D Ray Dream, puis a démarré le portail **AdForum** dédié au marché de la publicité, dont il a conçu le site.

C'est un véritable visionnaire et un passionné du web sémantique. Il a plein d'idées sur les services qu'il pourrait rendre, et de manière très pragmatique. Il avait développé dès 2006 un prototype d'outil permettant de récupérer des données sur les sites web, et qui ne tournait que sur Macintosh.

Après une grosse année de préparation, il a lancé son entreprise, Outwit Technologies, et trouvé des business angels pour financer l'aventure et divers conseils tels qu'Etienne Krieger (CEO de **Navidis**) et votre serviteur. De la est sorti une première bêta d'Outwit, créée avec trois développeurs, et qui est maintenant téléchargeable.

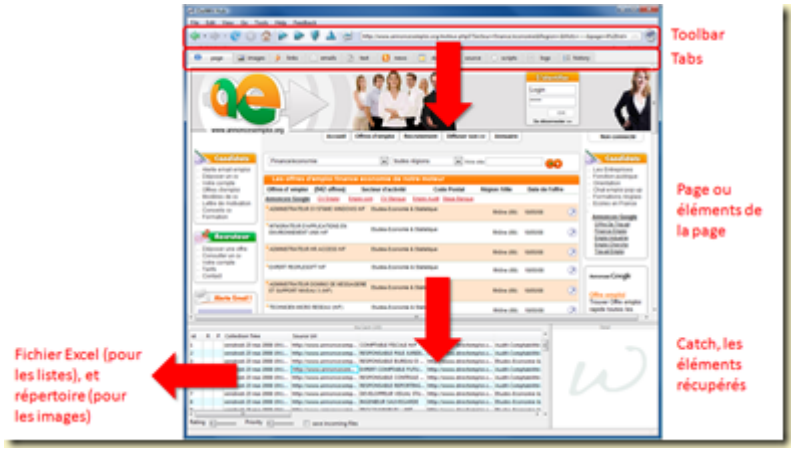
Le produit

Le "Outwit Hub" est une extension Firefox compatible avec la version 3.X du navigateur (depuis fin juillet 2008). C'est en fait une application à part entière qui intègre le navigateur de Firefox et se lance à partir de Firefox.

Outwit Hub est téléchargeable en bêta 1 **ici**. Son mode d'emploi rapide est **ici**. Le produit étant encore en développement, son aide en ligne est succincte.

Le principe d'utilisation d'Outwit Hub s'articule en trois phases :

- On ouvre une page web et on sélectionne les éléments à récupérer dans le haut de la fenêtre (images, listes, emails, tableaux, RSS).
- On envoie la sélection dans le "catch", en mode manuel ou automatique, c'est une liste en bas de la fenêtre d'Outwit Hub. Le catch va accumuler les éléments récupérés jusqu'à ce qu'on efface ce qu'il contient. Le mode manuel consiste à appuyer sur le bouton "**Catch**" et le mode automatique, à sélectionner la case "**Catch selection**" en bas à droite. Il s'appliquera aux pages web suivantes obtenues soit avec une nouvelle URL, soit avec les boutons de déplacement (flèches droites simples ou doubles).
- La sauvegarde des éléments récupérés sur le disque, en mode manuel ou automatique (optionnelle pour les images). Pour des images, c'est leur récupération dans un répertoire. Pour des listes et tableaux, c'est la sauvegarde dans un fichier Excel ou la copie dans le presse-papier après leur sélection.



Le logiciel est plein de ressources mais il est encore un peu tôt pour les décrire en détail car il faut déjà s'approprier les fonctions de base.

En gros, Outwit Hub sera scriptable et on pourra créer des "scrapers" personnalisés de sites divers. Et l'ensemble fonctionnera dans un mode communautaire. Le site web d'Outwit jouera ainsi un rôle fédérateur des utilisateurs et contributeurs d'Outwit.

De plus, on pourra développer des "Outfits", des applications spécifiques créées avec le moteur d'Outwit qui est au coeur du Hub. Le modèle économique reposera sur leur développement et leur distribution, ainsi que sur un financement du site communautaire d'Outwit par la publicité. Tout est à construire.

Tout ceci constitue le plan à exécuter pour les mois à venir !

Les utilisateurs

Qui pourra utiliser Outwit ?

Pour ce qui est de la récupération d'images, c'est un peu tout le monde.

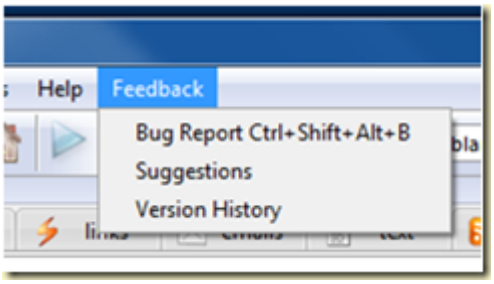
Pour ce qui est de listes, cela concernera les "chercheurs d'informations structurées" avec notamment : les enseignants et les étudiants, les métiers du marketing, de la communication, de la veille technologie, les économistes et les investisseurs financiers.

Et ce, à l'échelle mondiale. Ce qui explique pourquoi le logiciel est actuellement développé avec une interface utilisateur en anglais.

L'objectif de Jean-Christophe Combaz est de créer ou permettre la création par des tiers d'applications à partir du noyau de l'Outwit Hub. A terme, Outwit Hub sera plutôt destinés à des utilisateurs avancés du web.

Comment le découvrir ?

Je vous propose de découvrir le Outwit Hub grâce à deux petits tutoriaux "faits main" illustrés dans les deux posts suivants. L'un pour récupérer des listes, comme celle des 2000 plus grandes entreprises mondiales du classement Forbes 2000. Et l'autre, pour récupérer des images sur Google Image. Il s'agit de tutoriaux sur des fonctions basiques du logiciel car celui-ci permet de réaliser des choses bien plus sophistiquées… qu'il reste à documenter !



Ensuite, vous pourrez faites quelques tests vous-mêmes et utiliser le menu “**Feedback**” d’Outwit pour faire suggestion et reporter les bugs et dysfonctionnements que vous identifierez. Et en étant indulgent : ce n’est qu’une bêta non finalisée.

Article modifié le 1ier août 2008 pour indiquer qu’Outwit Hub est maintenant compatible avec Firefox 3.X.

Cet article a été publié le 24 mai 2008 et édité en PDF le 18 mars 2024.