Opinions Libres

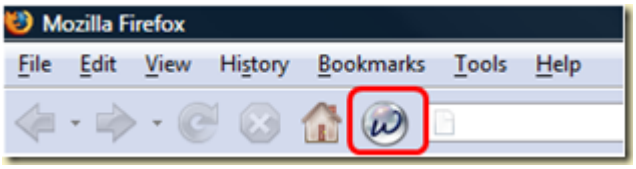
le blog d'Olivier Ezratty

Tutorial Outwit – récupération de listes

Après avoir décrit l'objet du logiciel **Outwit Hub**, passons à un petit tutorial qui vous permettra de l'expérimenter pour une application pratique: la récupération de listes de sites web. Il vous faudra d'abord installer ou disposer de Firefox 3.x de Firefox. Ensuite, vous installerez la **bêta d'Outwit Hub**. Elle fonctionne sur Windows, MacOS comme sur Linux.

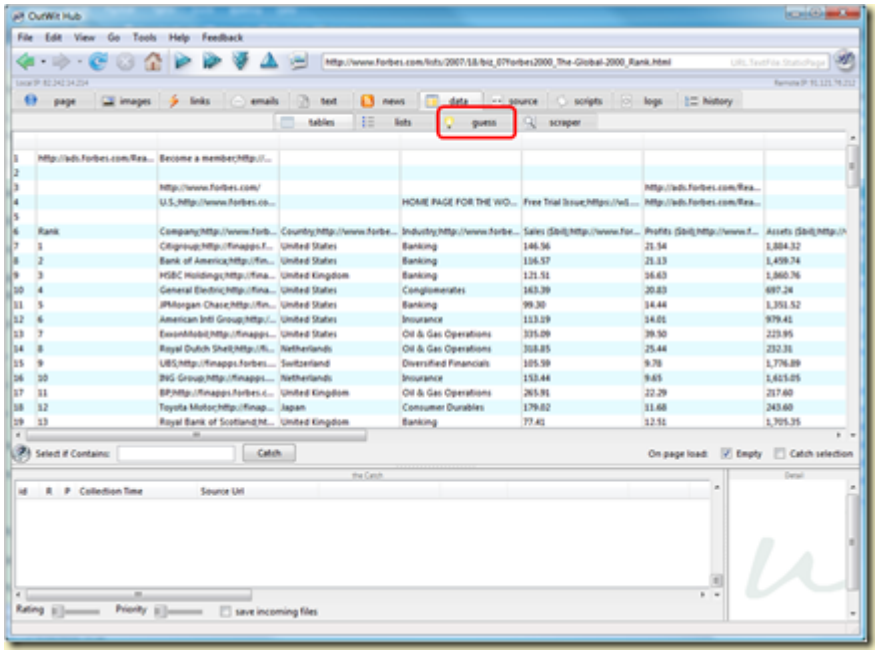
Je vais utiliser ici un exemple de récupération de données structurées déjà exploité pour la préparation de mes supports de cours sur l'économie de l'innovation : la liste du **Forbes 2000** qui regroupe les 2000 plus grandes entreprises mondiales. Elle est porteuse d'une très grande richesse d'informations. Cette liste apparait sous forme de tableaux que l'on peut copier coller à la main dans Excel. Mais il faut charger 20 pages ce qui est bien fastidieux.

Alors, lançons Firefox puis Outwit Hub à partir de l'icone installée dans la toolbar de Firefox par ce dernier comme indiqué ci-dessous.

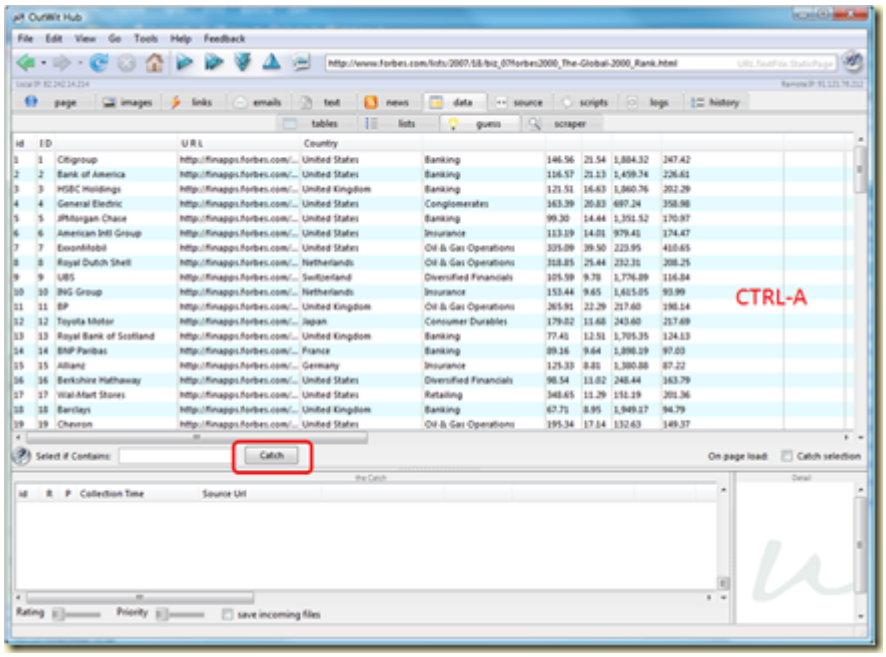


Dans la barre d'URL d'Outwit Hub, collez **l'URL de Forbes**. Faites "Enter" et la page s'affiche comme dans tout navigateur. Ensuite, cliquez sur le tab "**Data**" comme indiqué ci-dessous.

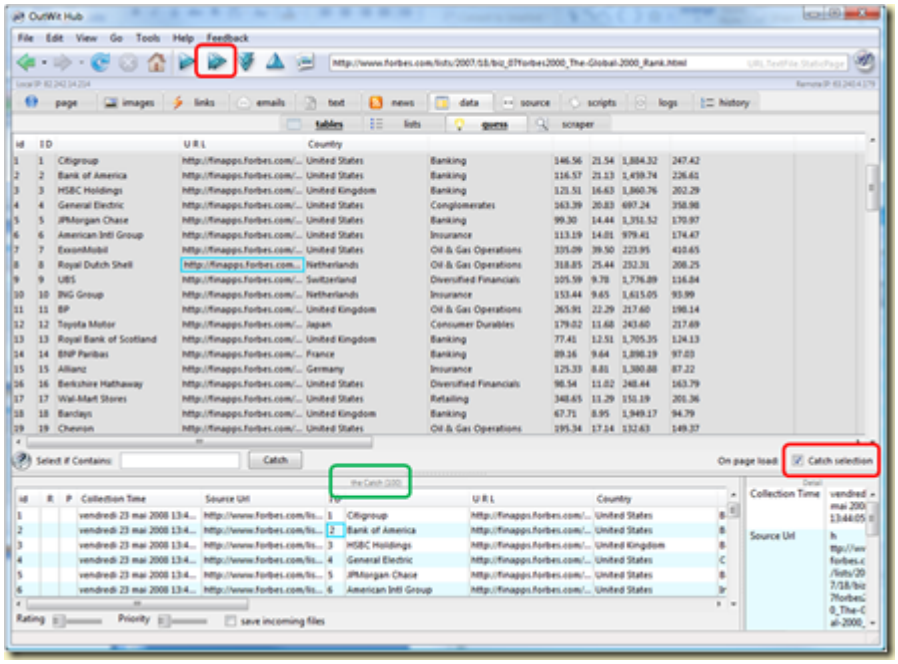
Cela affiche les données tabulées de la page HTML. Mais on n'a pas encore un beau tableau. Pour ce faire, cliquez sur le bouton "**Guess**" comme indiqué ci-dessous. Cette fonction déclenche la détection automatique d'un véritable tableau de données structurées par Outwit.



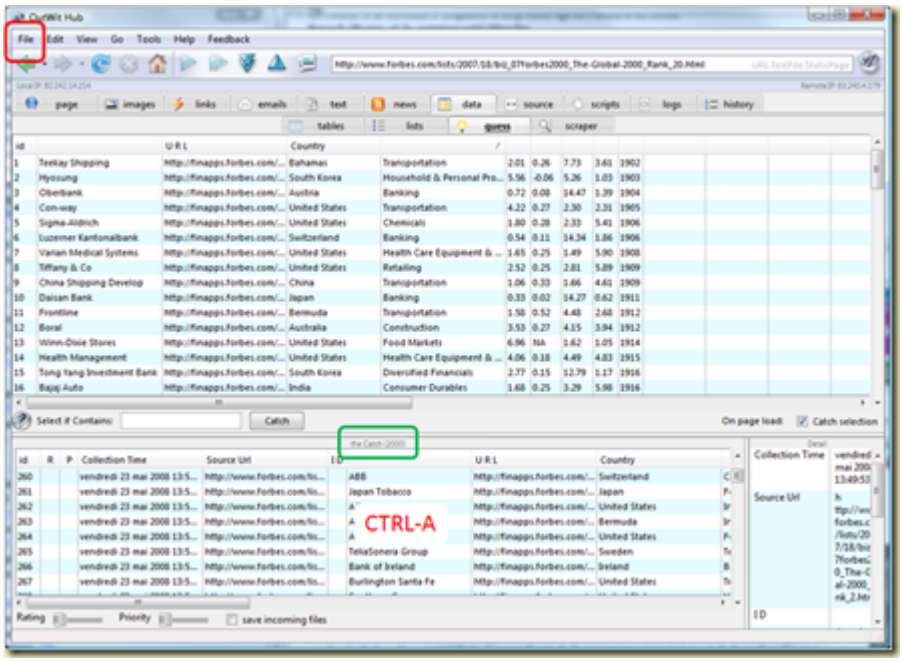
Le résultat apparait ci-dessous. A partir de là, on va sélectionner toutes les lignes dans la liste, par exemple en cliquant sur une ligne et en faisant “CTRL-A” avec le clavier. Et puis, on va la copier dans la zone en bas de la fenêtre qui est le “**Catch**” en cliquant sur le bouton du même nom. Le catch, c’est ce que l’on a attrapé dans les pages et qui s’accumule au fur et à mesure. Ensuite, on peut récupérer ce qu’il y a dans le catch de différentes manières selon qu’il s’agit d’images ou de listes.



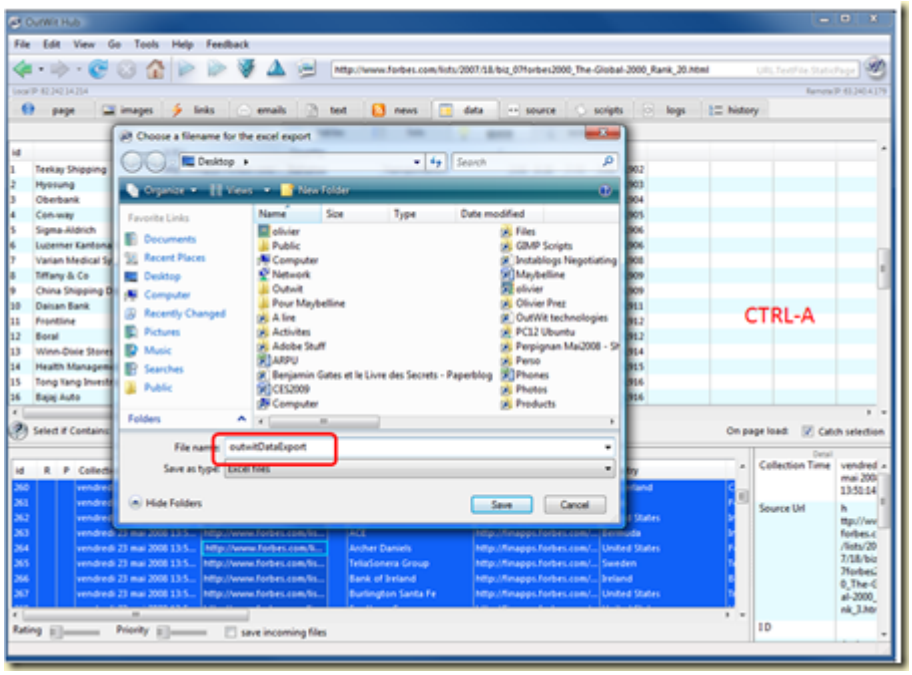
Maintenant, nous allons lancer le processus qui va automatiquement récupérer la suite du Forbes 2000 dans les 19 pages web suivantes. Il faut d’abord sélectionner la checkbox “**Catch selection**” qui indique que dans toute ouverture de page ou sélection de page suivante, le contenu identifié sera automatiquement basculé dans le catch. Ensuite, on cliquera sur la “**double flèche droite**” (fast forward) qui est dans la barre d’outils en haut de Outwit. Elle déclenchera l’analyse automatique des pages suivantes jusqu’à la vingtième. Si on veut le faire à la main, on peut utiliser le bouton “flèche droite” (“Play”) page par page. On peut aussi arrêter le scan automatique des pages avec **ESC** ou en cliquant à nouveau sur la double flèche.



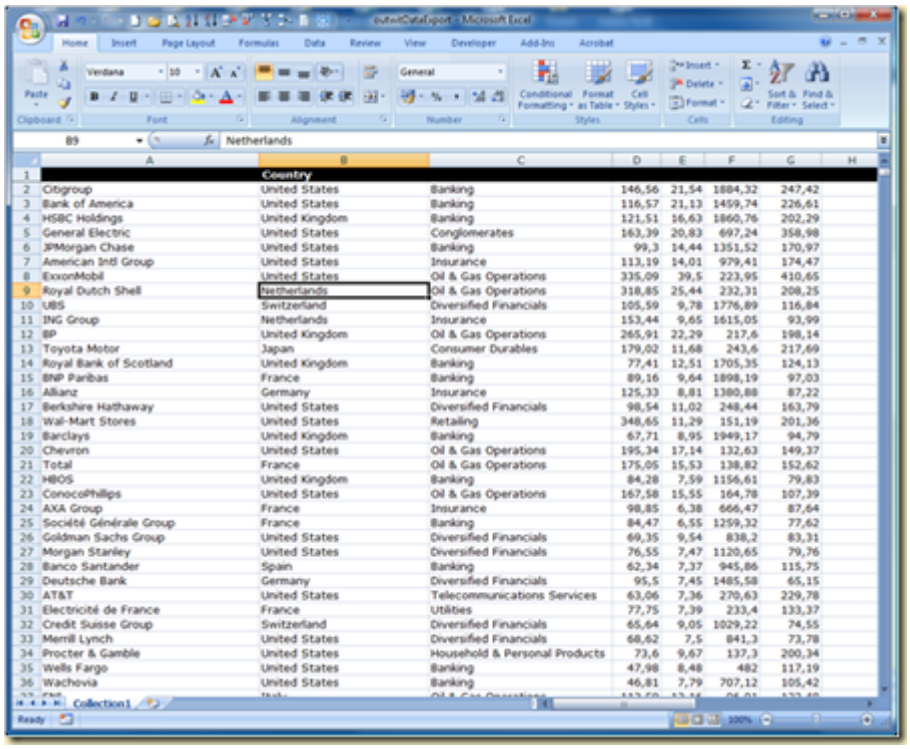
Une fois les 20 pages scannées ce qui prend quelques minutes à peine, le catch comprend bien 2000 entrées. On va maintenant le sauvegarder pour l'exploiter.



Pour ce faire, on va dans le menu "**File**" et on lance "**Export selection as…**". Le seul format supporté est pour l'instant Excel. C'est en fait un schéma XML supporté par Excel, mais je ne sais pas trop lequel. Et il n'est pas lisible dans OpenOffice 2.4. Un export CSV est en tout cas prévu à terme. Et on peut tout de même copier la sélection du haut ou celle du catch dans le presse papier pour la récupérer ailleurs (avec le bouton droit de la souris et "**Copy**").



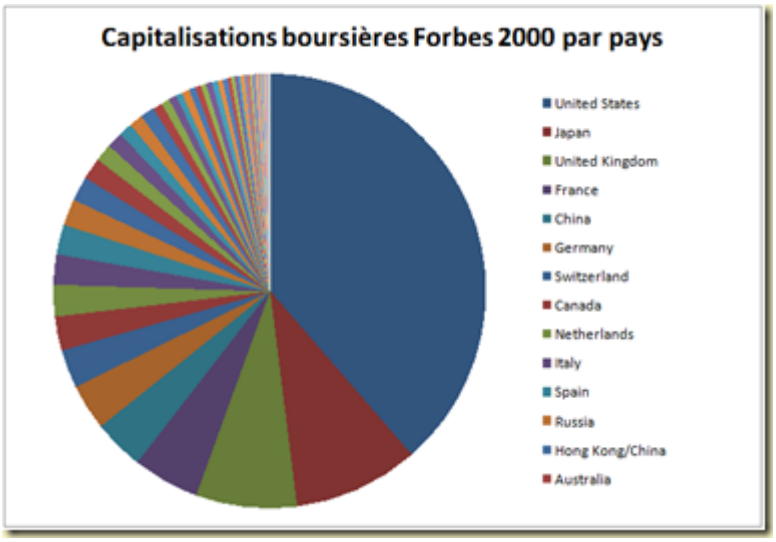
Et on ouvre le fichier sous **Excel**, version 2007 US dans l'exemple ci-dessous (ça fonctionne peut-être sous OpenOffice). On élimine ensuite les colonnes inutiles. On peut constater que la colonne "Rank" n'a pas été bien récupérée mais ce n'est pas grave dans ce cas là. C'est un défaut qui sera traité dans les évolutions d'Outwit pour faire en sorte que la détection automatique de la structure des tables d'une page ne soit pas refaire à chaque page.



	Country					
Citigroup	United States	Banking	146,56	21,54	1884,32	247,42
Bank of America	United States	Banking	116,57	21,13	1459,74	226,61
HSBC Holdings	United Kingdom	Banking	121,51	16,63	1860,76	202,29
General Electric	United States	Conglomerates	163,39	20,83	697,24	358,98
JPMorgan Chase	United States	Banking	99,3	14,44	1351,52	170,97
American Intl Group	United States	Insurance	113,19	14,01	979,41	174,47
ExxonMobil	United States	Oil & Gas Operations	335,09	39,5	223,95	410,65
Royal Dutch Shell	Netherlands	Oil & Gas Operations	318,85	25,44	232,31	208,25
UBS	Switzerland	Diversified Financials	105,59	9,78	1776,89	116,84
ING Group	Netherlands	Insurance	153,44	9,65	1615,05	93,99
BP	United Kingdom	Oil & Gas Operations	265,91	22,29	217,6	198,14
Toyota Motor	Japan	Consumer Durables	179,02	11,68	243,6	217,69
Royal Bank of Scotland	United Kingdom	Banking	77,41	12,51	1705,35	124,13
BNP Paribas	France	Banking	89,16	9,64	1898,19	97,03
Allianz	Germany	Insurance	125,33	8,81	1380,88	87,22
Berkshire Hathaway	United States	Diversified Financials	98,54	11,02	248,44	163,79
Wal-Mart Stores	United States	Retailing	348,65	11,29	151,19	201,36
Barclays	United Kingdom	Banking	67,71	8,95	1949,17	94,79
Chevron	United States	Oil & Gas Operations	195,34	17,14	132,63	149,37
Total	France	Oil & Gas Operations	175,05	15,53	138,82	152,62
HBOS	United Kingdom	Banking	84,28	7,59	1156,61	79,83
ConocoPhillips	United States	Oil & Gas Operations	167,58	15,55	164,78	107,39
AXA Group	France	Insurance	98,85	6,38	666,47	87,64
Société Générale Group	France	Banking	84,47	6,55	1259,32	77,62
Goldman Sachs Group	United States	Diversified Financials	69,35	9,54	838,2	83,31
Morgan Stanley	United States	Diversified Financials	76,55	7,47	1120,65	79,76
Banco Santander	Spain	Banking	62,34	7,37	945,86	115,75
Deutsche Bank	Germany	Diversified Financials	95,5	7,45	1485,58	65,15
AT&T	United States	Telecommunications Services	63,06	7,36	270,63	229,78
Electricité de France	France	Utilities	77,75	7,39	233,4	133,37
Credit Suisse Group	Switzerland	Diversified Financials	65,64	9,05	1029,22	74,55
Merrill Lynch	United States	Diversified Financials	68,62	7,5	841,3	73,78
Procter & Gamble	United States	Household & Personal Products	73,6	9,67	137,3	200,34
Wells Fargo	United States	Banking	47,98	8,48	482	117,19
Wachovia	United States	Banking	46,81	7,79	707,12	105,42

Pour exploiter les données, il faut les normaliser au format français : on sélectionne les colonnes de chiffres, on utilise la fonction de Recherche/Remplacement d'Excel pour remplacer les "," (virgules) par rien du tout, et ensuite les "." par "," et le tour est joué. Là encore, cette transformation devrait être effectuée automatiquement par Outwit Hub après la fin de la bêta en fonction de vos "Regional Settings".

Et hop, on créé un petit tableau dynamique croisé et un camember avec par exemple la répartition par pays de la capitalisation boursière des 2000 plus grandes entreprises mondiales.



Pour l'instant, ce processus automatique (Guess) ne fonctionne pas encore parfaitement sur tous les sites. J'ai pu le tester avec succès sur Kelkoo.fr et quelques autres sites comme des sites de recherche d'emploi. Mais entre les quatre fonctionnalités d'extraction de données sous l'onglet "Data" (trois sont automatiques: tables, listes et guess, et une est manuelle : scrapers), il y a en général toujours un moyen de s'en sortir. D'autres tutoriaux suivront qui seront publiés sur le site d'Outwit.

A vous de jouer maintenant…

Le tutoriel suivant traitera de la récupération d'images.

Article mis à jour le 1ier août 2008 pour tenir compte de la compatibilité d'Outwit Hub avec Firefox. Mais le tutoriel n'a pas été encore retesté dans cette version.

Cet article a été publié le 24 mai 2008 et édité en PDF le 15 mars 2024.