OUTIL D'EXTRACTION AUTOMATIQUE ET DE TRAITEMENT EN AMONT DE L'INFORMATION :

Frédérique de Ruiter, Master 2 Intelligence économique et territoriale,
frederique.deruiter@neuf.fr

Cheikh Omar Baro, Master 2 Intelligence économique et territoriale,
chobaro@yahoo.fr

Luc Quoniam, Professeur des Universités,
quoniam@univ-tln.fr

Arie de Ruiter, Architecte de systèmes d'information,
arie.ruiter@neuf.fr

Université du Sud Toulon Var.

Résumé :

Nous présentons dans les pages qui suivent un outil d'extraction automatisée de références bibliographiques qui fonctionne à partir de certains sites nationaux d'Amazon. Cet outil permet de paramétrer en amont l'extraction de l'information, le tri ainsi que son classement sur une base de données exploitable suivant les besoins de l'utilisateur.

Summary:

In this article we present a tool for automated extraction of bibliographic references based on products on sale on the national sites of Amazon. The tool permits to parameterise the extraction of information, the sorting as well as the classification within a database depending on user requirements.

Mots clés : Outil de recherche automatique, Macro Search, Amazon, Bibliographie, Traitement de l'information.

Key words: Automatic Search Tool, Macro Search, Amazon, Bibliography, Data processing.

OUTIL D'EXTRACTION AUTOMATIQUE ET DE TRAITEMENT EN AMONT DE L'INFORMATION :

Dans le contexte actuel de **l'économie informationnelle**, marquée par la turbulence, les entreprises qui se maintiennent et accroissent leurs performances sont celles qui savent être à l'écoute systématique de leur environnement. Mais, l'information étant devenue **très accessible**, voire même surabondante, le plus important est l'usage que l'on en fait, grâce au traitement qui lui donne du **sens** et permet son **actualisation**.

Comme l'a souligné Herbert Simon[1]"Les systèmes de traitement de l'information de notre monde contemporain baignent dans une abondance excessive d'informations et de symboles. Dans un tel monde, la ressource rare n'est pas l'information, mais la capacité de traitement pour s'occuper de cette information". On mesure ainsi l'importance de la capacité de traitement, particulièrement pour une structure de veille dont l'objectif principal est l'aide à la décision.

Dans le cadre de ce travail, nous considérerons la Veille comme "un processus dont l'objectif est d'aboutir à une prise de décision"[Antonio Da Silva, 2002]

Pour ce faire, il faut disposer d'une bonne méthode de sélection des information, sans laquelle "il n'est pas de veille stratégique viable"[LESCA, 1998].

Il convient ainsi de disposer de la méthode et de l'outil pertinents. Comme toute collecte d'informations elle requiert pour être efficace, une stratégie de recherche. En effet pour éviter les interminables itérations de la recherche traditionnelle de bibliographie, il importe de réfléchir à une méthode de recherche.

L'objectif de notre travail consistera ainsi à définir la bonne stratégie et l'outil le plus pertinent pour l'acquisition d'une base bibliographique actualisée et susceptible d'être intégrée dans une base de donnée aux fins de son exploitation adéquate.

L'outil "Macro Recherche" sur ce point précis permet de lancer une extraction automatique de références d'ouvrages pertinents qui donne en résultat une conséquente base bibliographique après annulation des doublons, ainsi que la capacité de filtrer, par exemple par l'année de publication. Grâce aux requêtes il a été possible d'extraire à la fois les documents primaires (les références sur l'ouvrage) et les documents secondaires notamment les résumés des ouvrages.

Nous exposerons dans les pages qui suivent la restitution de l'expérience et une application de cet outil à la recherche bibliographique sur le thème de l'Intelligence Economique.

1 – CADRAGE THEORIQUE

1.1 – Analyse des besoins

Diverses études sur l'analyse des comportements des internautes lorsqu'ils effectuent une recherche sur les moteurs ont montré une tendance à s'arrêter sur les premiers résultats ou sur les premières pages[2].

Ce qui pose un réel problème vu le nombre relativement élevé des pages renvoyées par les moteurs de recherche. A fortiori quand certaines études révèlent l'absence de pertinence fréquente des premiers résultats du fait de la présence de liens sponsorisés[3].

Concevoir un outil permettant de lever cet écueil afin d'exploiter au mieux les requêtes faites à travers les moteurs de recherche demeure important.

[1] Cité par Pateyron E., (1997), Veille stratégique, in Encyclopédie de gestion, pp. 183-194.

[2] Première étude co-réalisée par Jupiter Research et Iprospect : analyse des tendances observées sur trois ans (2002, 2004, 2006) ; deuxième étude de Harvest Digital et Métro Research : janvier 2006.

Source : httt://veillepme.blogspirit.com/etudes/06/06/2006

[3] Jean Véronis. (2006) "Etude comparative de six moteurs de recherche" [disponible sur htt://www.up.univ-mrs.fr/veronis/pdf/2006-etude-comparative.pdf]

Etant entendu que l'objectif est de fournir la bonne information, celle qui répond aux besoins du décideur, une surveillance systématique des publications sur un domaine déterminé semble d'un intérêt stratégique pour une structure de veille.

En effet, une revue de bibliographie est nécessaire pour à la fois l'information spécialisée fournie par les monographies mais aussi pour l'approfondissement de certaines questions qu'elle rend aisé. Elle permet également de repérer les auteurs importants, les éditeurs, les titres d'ouvrages et éventuellement les sites web à connaître afin de suivre un domaine.

La notion de bibliographie peut recouvrir diverses acceptions mais nous retiendrons la suivante « une liste de références ou de notices bibliographiques classées selon certains critères pour permettre le repérage des documents référencés » [Beaudiquez, 1989].

La méthode traditionnelle que pratique souvent étudiants et chercheurs rompus à la recherche d'auteurs de référence est l'interrogation de bases de données de bibliothèques, la consultation éventuelle de renvois bibliographiques en fin d'articles ou d'ouvrages.

Cette recherche traditionnelle requiert de connaître les caractéristiques des langages utilisés par les spécialistes de la documentation pour décrire le contenu des documents scientifiques, qu'il s'agisse de livres, de chapitres d'ouvrages, d'articles de périodiques, etc. [Piolat, Annie, 2002].

Le travail a consisté à définir la manière de lever l'écueil de la recherche manuelle de documents dans un contexte de surcharge de l'information et de l'absence de normalisation de l'indexation par les principaux moteurs de recherche? La connaissance de la description des produits au niveau de chaque page d'Amazon, a permis d'établir les requêtes de la macro. L'organisation des résultats obtenus au sein d'une feuille permettra leur intégration dans une base de données Winisis[4].

[4] Logiciel « freeware » développé et distribué par l'Unesco.

1.2 – Le choix de la source d'informations :

Deux raisons principales justifient le choix porté sur Amazon :

Amazon offre la possibilité de lancer la recherche sur ses sites en France, au Canada, en Grande Bretagne, en Allemagne ainsi que le site commercial lui-même.

En outre, La présentation des pages est similaire sur les différents sites nationaux. La structure des documents notamment en ligne (Internet) peut être très variable, ce qui complique d'autant les méthodes de recherche automatique et simultanée sur plusieurs sources. La similarité de présentation des pages sur Amazon permet de pallier ce problème.

2 – METHODES ET RESULTATS

2.1 – Définition des mots clés

Afin d'alimenter la feuille "Searchstring", nous avons recueilli un ensemble de vingt sept mots-clés, principalement à travers le Référentiel de formation en Intelligence Economique en France.

L'ensemble de mots-clés est un mix de mots français et anglais sur le thème de l'Intelligence Economique.

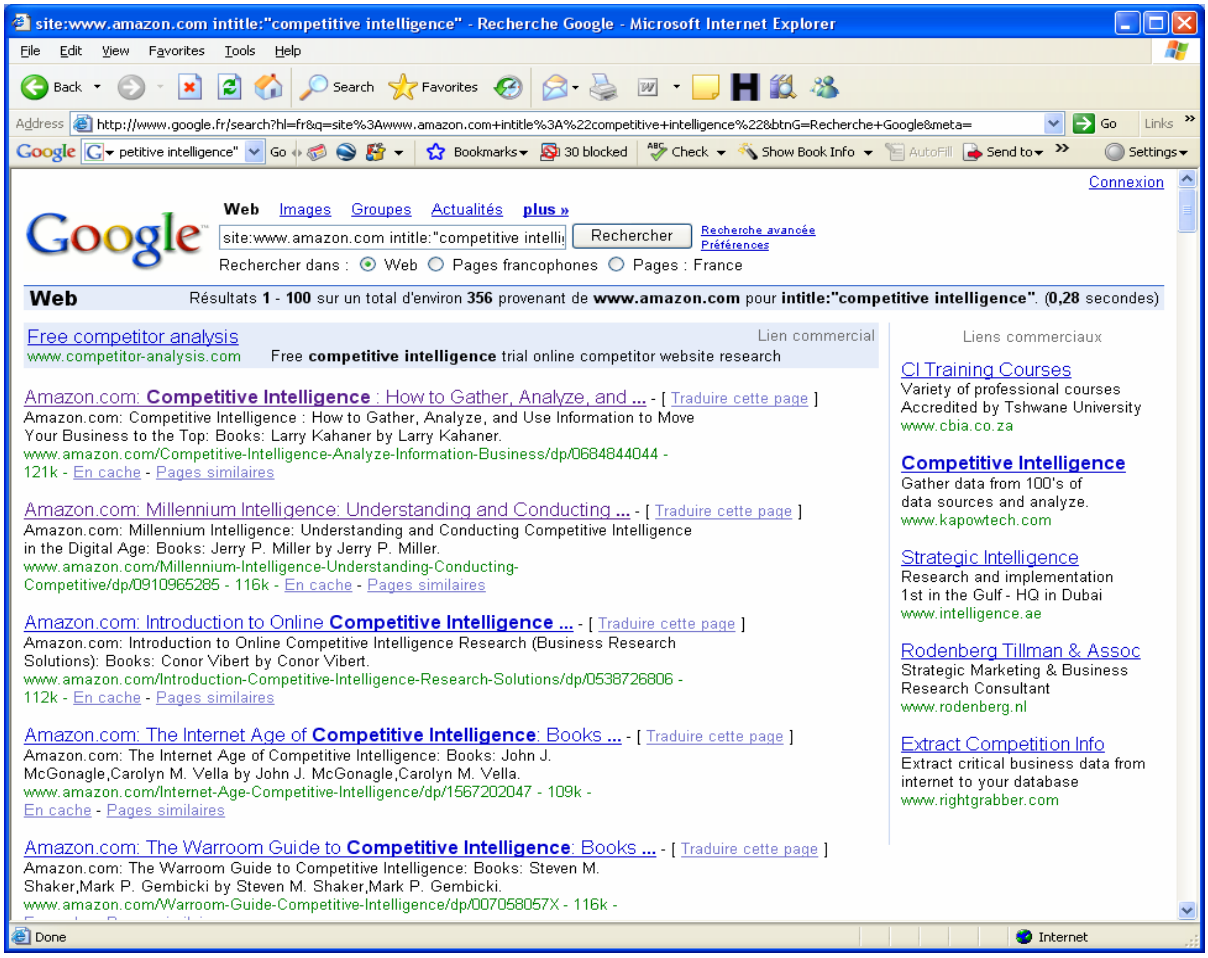
Toutefois, notons qu'un chercheur sélectionnera ses propres mots-clés « au fur et à mesure » et lancera rarement une requête sur la base immédiate de 27 mots-clés simultanément.

2.2 – La stratégie de recherche : la requête sur Google.

Afin de trouver les résultats de toutes les pages sur **Amazon.com** avec un titre contenant, par exemple, **'competitive intelligence'** on lance la requête suivant sur Google :

⇒ **site:www.amazon.com intitle:"competitive intelligence"**

Celle-ci donne comme résultat :



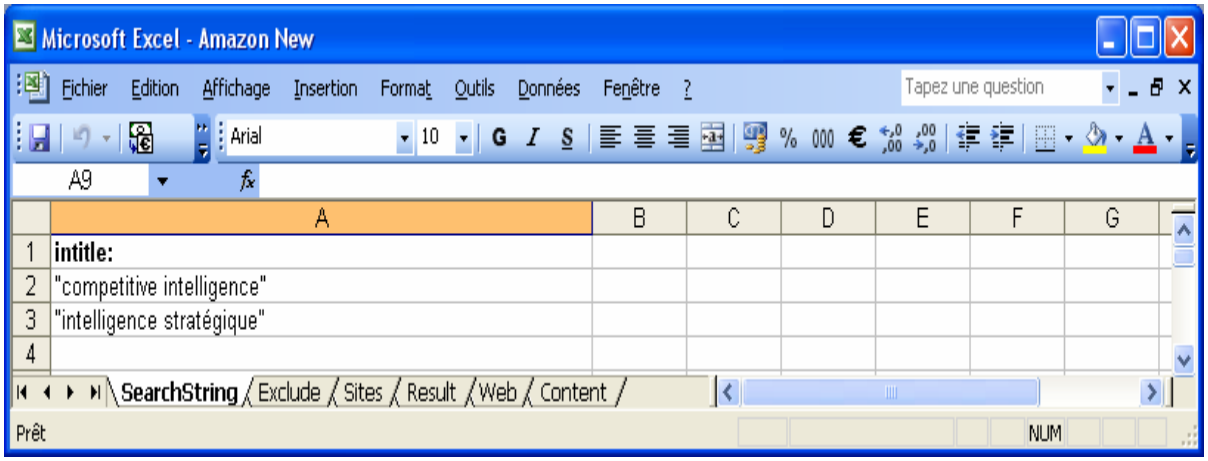
Ce qui est intéressant dans ce résultat est la traduction par Google de cette requête en l'URL suivant (voir 'Address' ci-dessus) :

⇒ http://www.google.fr/search?hl=fr&q=site%3Awww.amazon.com+intitle%3A%22competitive+intelligence%22&btnG=Recherche+Google&meta=

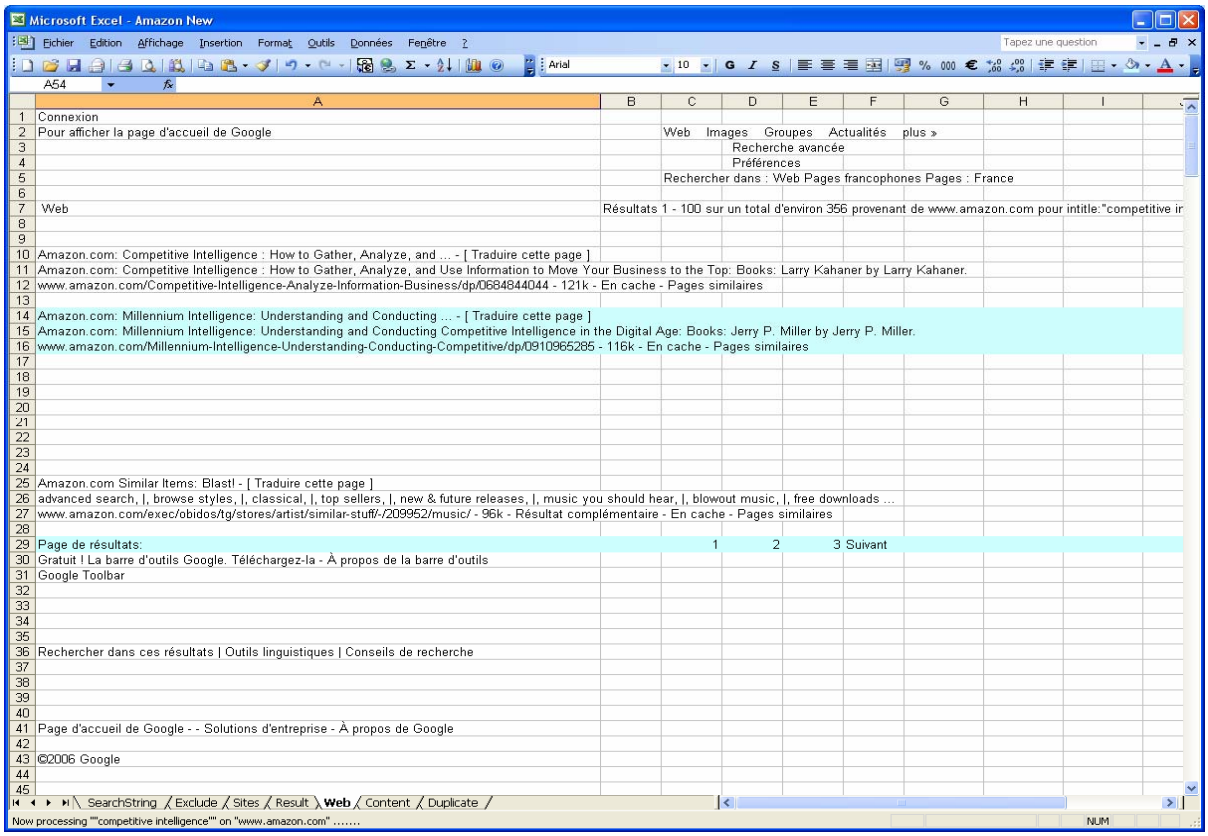
Dans cet URL on retrouve bien notre requête lancée sur Google (avec une traduction de certains caractères en représentation hexadécimale).

2.3 – L'utilisation de Visual Basic dans Excel

Nous avons développé une macro en Visual Basic dans un fichier Excel, effectuant des requêtes automatisées sur Google à partir d'une feuille **'SearchString'**, contenant notre sélection de mot-clés pertinents pour la recherche.



Afin de récupérer les résultats d'une requête dans une feuille Excel, la macro établit une connexion, utilisant un URL composé de la façon précédemment indiquée, qui donne le résultat suivant dans une feuille **'Web'** du fichier Excel[5] :



[5] Cette copie d'écran est abrégée afin de montrer la façon dont la fin de la page de résultats de Google apparaît dans la feuille Excel

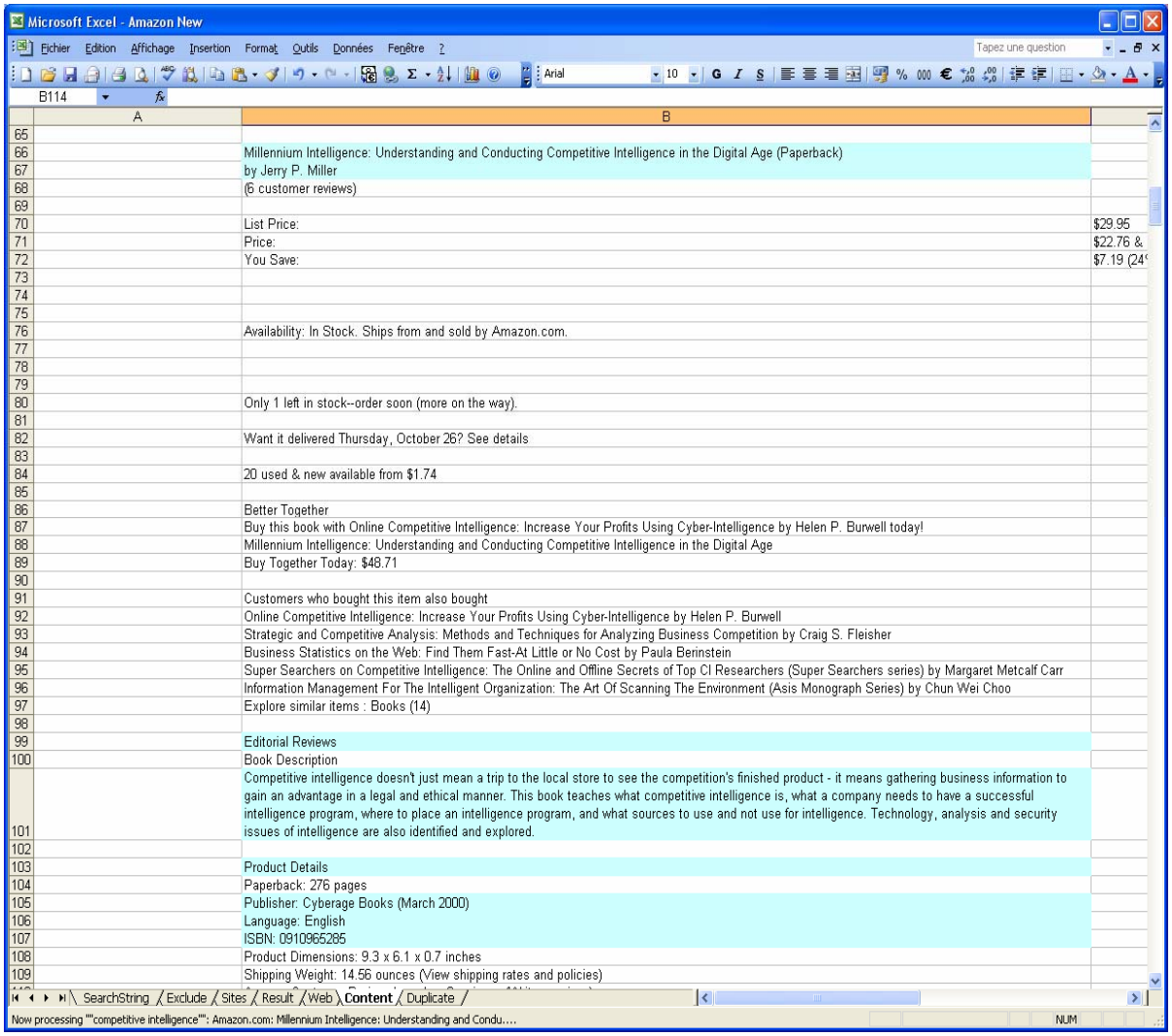
De cette feuille on peut déduire que chaque résultat de Google consiste en 3 lignes et que la troisième ligne contient l'URL de la page détaillée du produit sur Amazon.com.

De plus, si une ligne commençant avec le texte **'Page de résultats :'** contient le mot **'Suivant'** dans une des colonnes sur la même ligne, une prochaine page de résultats de Google est à traiter.

Cette prochaine page peut être récupérée en ajoutant le texte **'&start=100&sa=N'** à l'URL utilisé lors de la connexion établie pour la première page.

On en déduit ainsi que dans les préférences de Google le nombre de résultats à afficher est à 100. Afin de contourner cette préférence, l'expérimentation avec les URL sur Google a montré qu'en ajoutant le texte **'&num=100'**, l'affichage se fait par défaut en nombre de résultats égale à 100.

Afin de récupérer la page détaillée du produit sur Amazon.com, la macro établit une connexion, utilisant l'URL trouvée dans la troisième ligne de chaque résultat de Google, donnant la représentation suivante dans une feuille **'Content'** du fichier Excel :



Chaque page de présentation d'un ouvrage sur Amazon, contient différentes rubriques dont il a fallu tenir compte pour mieux extraire les informations nous intéressant dans le cadre de notre travail.

Pour ce faire l'extraction suit la logique suivante :

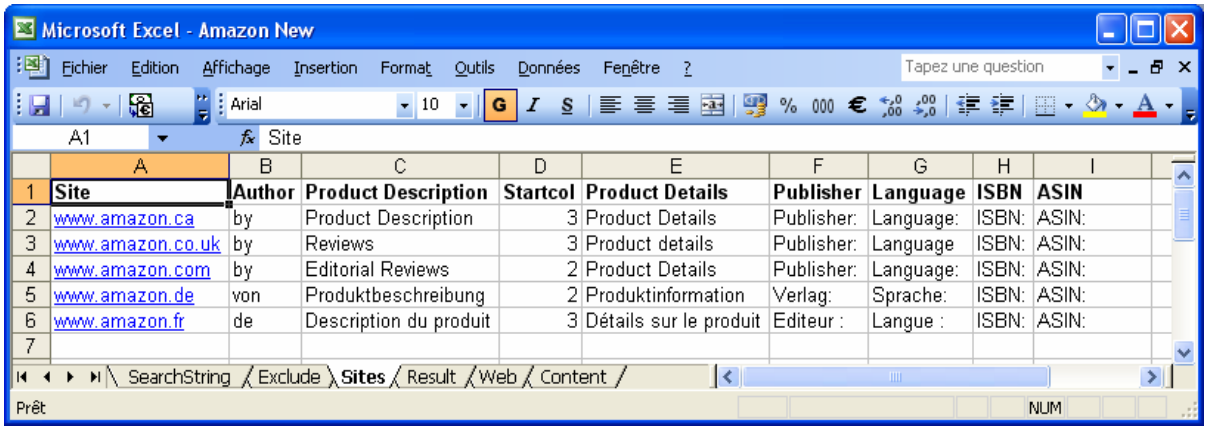
Afin de trouver la colonne contenant les détails sur le produit, la macro cherche le texte **'Product Details'**. Elle récupère ainsi l'éditeur (cellule commençant avec **'Publisher:'**), l'année de publication (les 4 chiffres avant la fin **')'** de l'éditeur), la langue (**'Language:'**), l'ISBN, ou l'ASIN, (respectivement **'ISBN:'** ou **'ASIN:'**), l'auteur (**'by'**) et le titre (la ligne précédente de celle-ci de l'auteur).

Pour trouver une description du produit, la macro cherche pour le texte **'Editorial Reviews'** dans les cellules de la même colonne que les autres données.

Afin de trouver une description 'utile' la macro prend le plus long texte dans les 6 cellules suivant la cellule du texte **'Editorial Reviews'**.

2.4 – La recherche Multi - sites Amazon :

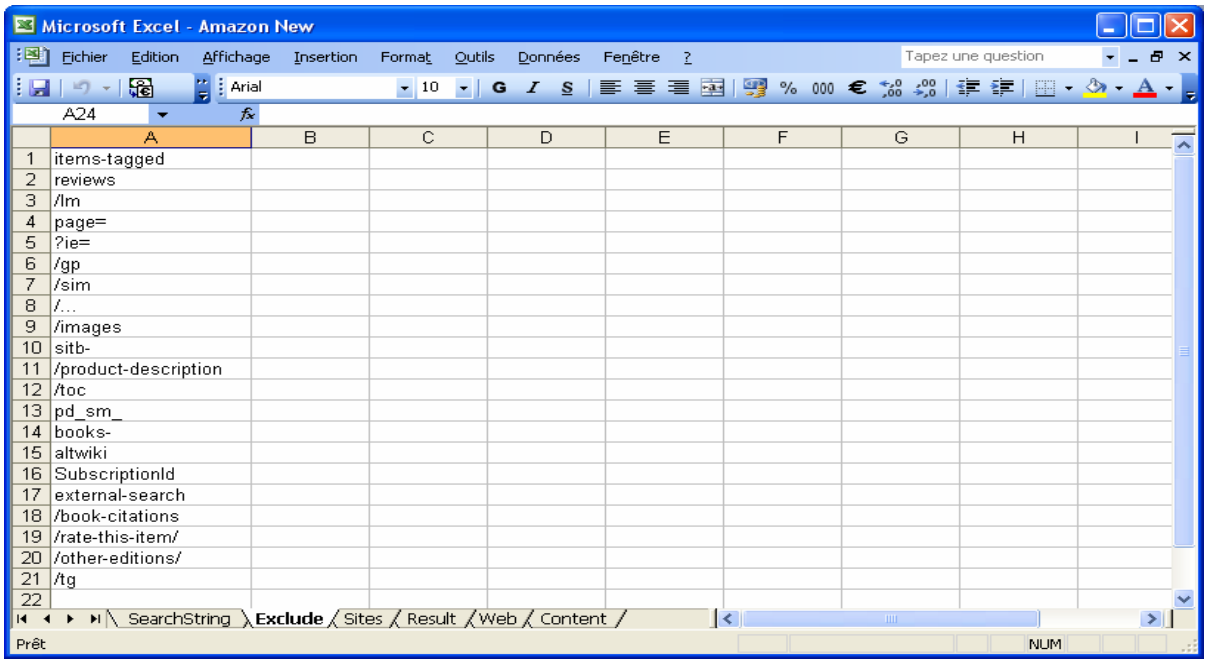
Afin de chercher des produits sur les différents sites nationaux d'Amazon, le fichier Excel contient une feuille nommée **'Sites'** (voir ci-dessous) dans laquelle on retrouve ces différents sites et la traduction de mots clés utilisés pour la recherche de détails du produit, comme décrit précédemment.



	A	B	C	D	E	F	G	H	I
1	**Site**	**Author**	**Product Description**	**Startcol**	**Product Details**	**Publisher**	**Language**	**ISBN**	**ASIN**
2	www.amazon.ca	by	Product Description	3	Product Details	Publisher:	Language:	ISBN:	ASIN:
3	www.amazon.co.uk	by	Reviews	3	Product details	Publisher:	Language	ISBN:	ASIN:
4	www.amazon.com	by	Editorial Reviews	2	Product Details	Publisher:	Language:	ISBN:	ASIN:
5	www.amazon.de	von	Produktbeschreibung	2	Produktinformation	Verlag:	Sprache:	ISBN:	ASIN:
6	www.amazon.fr	de	Description du produit	3	Détails sur le produit	Editeur :	Langue :	ISBN:	ASIN:
7									

2.5 – L'exclusion de certains URL d'Amazon

L'expérimentation sur les résultats renvoyés par Google a montré, qu'il existe des URL d'Amazon qui ne contiennent pas de détails sur un produit spécifique. Afin d'exclure ces URL de la recherche, le fichier Excel contient une feuille nommée **'Exclude'** (voir exemple ci-dessous), contenant toutes les parties de texte trouvés dans un URL à exclure.



2.6 – La signalisation de doublons

Le doublon est la double occurrence d'un ouvrage identifiable par son ISBN ou ASIN. Afin d'indiquer si un produit a déjà été trouvé avant, la macro stocke les **ISBN/ASIN**'s dans une feuille temporaire **'Duplicate'**.

Les résultats de la feuille **'Result'** contiennent une colonne **'Duplicate'** qui aura une valeur **'No'** pour la première occurrence de l'ISBN/ASIN et **'Yes'** pour chaque récurrence.

Par exemple, le même ouvrage a souvent été commercialisé simultanément sur Amazon.uk, Amazon.com et Amazon. ca (…)

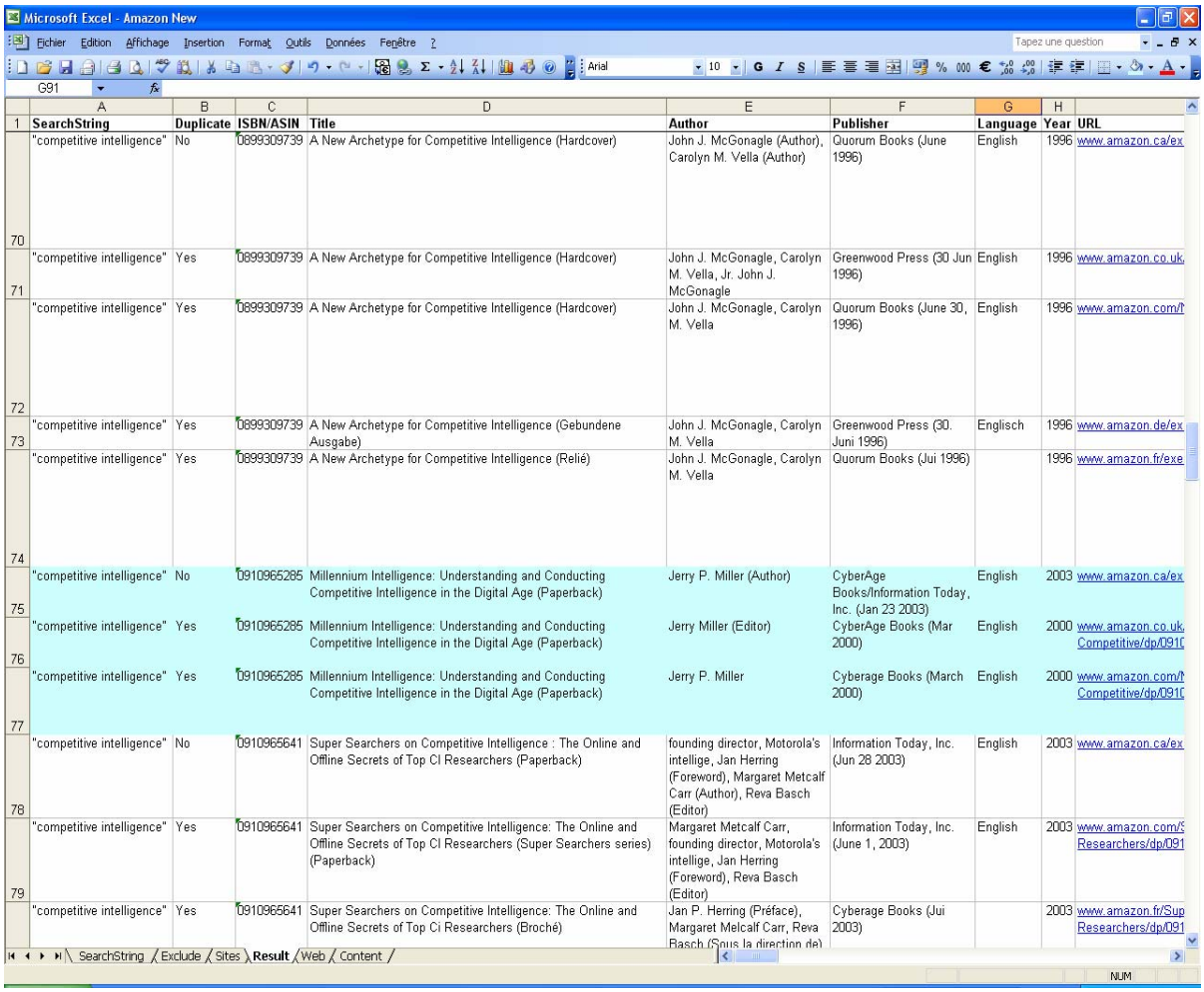
2.7 – Résultats

L'utilisation de la macro avec une base de 27 mots – clés portant sur l'Intelligence économique a généré 2291 résultats.

Après traitement de ces résultats, notamment l'élimination des doublons, nous nous retrouvons avec 1090 ouvrages. L'extraction automatique a duré 3h40mn.

La base de résultats a été classée sur une feuille Excel, ce qui permet un traitement, un tri suivant l'année, l'auteur…

Un extrait de ces résultats, dans lesquels nous retrouvons 3 occurrences du livre qui nous a servi comme exemple précédemment, est présenté ci-dessous :



Microsoft Excel - Amazon New

	A	B	C	D	E	F	G	H	
1	SearchString	Duplicate	ISBN/ASIN	Title	Author	Publisher	Language	Year	URL
70	"competitive intelligence"	No	0899309739	A New Archetype for Competitive Intelligence (Hardcover)	John J. McGonagle (Author), Carolyn M. Vella (Author)	Quorum Books (June 1996)	English	1996	www.amazon.ca/ex
71	"competitive intelligence"	Yes	0899309739	A New Archetype for Competitive Intelligence (Hardcover)	John J. McGonagle, Carolyn M. Vella, Jr. John J. McGonagle	Greenwood Press (30 Jun 1996)	English	1996	www.amazon.co.uk/
72	"competitive intelligence"	Yes	0899309739	A New Archetype for Competitive Intelligence (Hardcover)	John J. McGonagle, Carolyn M. Vella	Quorum Books (June 30, 1996)	English	1996	www.amazon.com/
73	"competitive intelligence"	Yes	0899309739	A New Archetype for Competitive Intelligence (Gebundene Ausgabe)	John J. McGonagle, Carolyn M. Vella	Greenwood Press (30. Juni 1996)	Englisch	1996	www.amazon.de/ex
74	"competitive intelligence"	Yes	0899309739	A New Archetype for Competitive Intelligence (Relié)	John J. McGonagle, Carolyn M. Vella	Quorum Books (Jui 1996)		1996	www.amazon.fr/exe
75	"competitive intelligence"	No	0910965285	Millennium Intelligence: Understanding and Conducting Competitive Intelligence in the Digital Age (Paperback)	Jerry P. Miller (Author)	CyberAge Books/Information Today, Inc. (Jan 23 2003)	English	2003	www.amazon.ca/ex
76	"competitive intelligence"	Yes	0910965285	Millennium Intelligence: Understanding and Conducting Competitive Intelligence in the Digital Age (Paperback)	Jerry Miller (Editor)	CyberAge Books (Mar 2000)	English	2000	www.amazon.co.uk/ Competitive/dp/091
77	"competitive intelligence"	Yes	0910965285	Millennium Intelligence: Understanding and Conducting Competitive Intelligence in the Digital Age (Paperback)	Jerry P. Miller	Cyberage Books (March 2000)	English	2000	www.amazon.com/ Competitive/dp/091
78	"competitive intelligence"	No	0910965641	Super Searchers on Competitive Intelligence : The Online and Offline Secrets of Top CI Researchers (Paperback)	founding director, Motorola's intellige, Jan Herring (Foreword), Margaret Metcalf Carr (Author), Reva Basch (Editor)	Information Today, Inc. (Jun 28 2003)	English	2003	www.amazon.ca/ex
79	"competitive intelligence"	Yes	0910965641	Super Searchers on Competitive Intelligence: The Online and Offline Secrets of Top CI Researchers (Super Searchers series) (Paperback)	Margaret Metcalf Carr, founding director, Motorola's intellige, Jan Herring (Foreword), Reva Basch (Editor)	Information Today, Inc. (June 1, 2003)	English	2003	www.amazon.com/ Researchers/dp/091
	"competitive intelligence"	Yes	0910965641	Super Searchers on Competitive Intelligence: The Online and Offline Secrets of Top Ci Researchers (Broché)	Jan P. Herring (Préface), Margaret Melcalf Carr, Reva Basch (Sous la direction de)	Cyberage Books (Jui 2003)		2003	www.amazon.fr/Sup Researchers/dp/091

SearchString | Exclude | Sites | Result | Web | Content

NUM

3 - CONCLUSION

Le travail effectué avec la macro Excel nous a permis d'extraire un nombre considérable d'ouvrages. La macro balayant les différents sites Amazon, pourrait être assimilée à un méta moteur. Cette caractéristique est à notre sens son principal intérêt. En effet, cet outil permet, pour une même requête, d'interroger plusieurs moteurs de façon simultanée, de rapatrier les résultats, le synthétiser et même proposer un récapitulatif des réponses données. En ce sens, c'est un outil qui, vue sa relative facilité de fabrication est d'une utilité essentielle pour l'usage d'une petite structure de recherche. Mais le premier biais qui apparaît est le nombre considérable de doublons notamment autant de doublons que d'ouvrages recensés. Ceci s'explique par le fait que la recherche a été réalisée effectivement sur différents sites nationaux. En ce sens, la sélection par l'ISBN a permis de faire le tri nécessaire. L'utilisation de cet outil à l'aide de termes généraux risque de générer beaucoup d'informations non pertinentes. Il importe de souligner qu'une recherche plus ciblée sur un nombre limité de mots-clés pourrait donner des résultats plus pertinents. La macro est donc spécialement adaptée à une recherche ponctuelle avec une stratégie bien définie au préalable, et avec un objectif de recherche spécifié. Il importe de remarquer que l'utilisation de cet outil est efficace dans le cadre de recherches sur des sujets très pointus où l'information est plus rarissime.

BIBLIOGRAPHIE

Beaudiquez M., (1989), Guide Bibliographie générale. Méthodologie et Pratique. Paris : Saur.

Blanco S., Sélection de l'information à caractère anticipatif : un processus d'intelligence collective

Dou H., Hassanaly P., Quoniam L., Latela A., 51990) : Veille technologique et information documentaire : de l'usage de la bibliométrie dans les services documentaires. Documentaliste-science de l'information, vol.27, n°3, mai juin-juin, p ;132-141 ; cité par Lopés Da Silva, opus cité.

Lefèvre Ph., (2000), La Recherche d'informations. Paris : Hermès.

Léon A., (2006), Savoir Chercher et Interroger : les repères méthodologiques, in " Ressources électroniques pour les étudiants, la recherche et l'enseignement, Formist, pp 59-61.

Lesca H., (1996) : veille stratégique : comment sélectionner les informations pertinentes ? Concepts, méthodologie, expérimentation, résultats. Conférence internationale de management stratégique. Lille, 13615 mai, pp1612.

Lesca, H., Schuler, M. (1998). Veille stratégique : comment ne pas être noyé sous les informations. In Economies et sociétés, sciences de gestion, série S.G., n°2/1998, PP159-177.

Lopes Da Silva, A. (2002). L'information et l'entreprise, des savoirs à capitaliser, méthodes, outils et applications à la veille. Université de Droit, Eco. & Sciences d'Aix-Marseille III.

Pateyron E.l, (1997), Veille stratégique, in "Encyclopédie de gestion", tome 1, pp. 183-194.

Piolat A., (2003), La Recherche documentaire. Manuel à l'usage des étudiants, doctorants et jeunes chercheurs.

Rostaing H., (1993), Veille technologique et Bibliométrie : concepts, outils applications, Thèse