

Stratégie et choix technologiques des éditeurs de logiciels de veille

Par Frédéric Martinet – Actulligence Consulting



Conférences Expert & Ateliers Solutions
Entreprise & Web applications
19 février 2010, Hotel Scribe Opéra **** Paris

Solutions for Enterprise Advanced Research

www.search2010.org



Présentation F. Martinet – Actulligence Consulting

- ▶ 9 ans d'activité professionnelle en intelligence économique
- ▶ Lancement d'Actulligence Consulting en Octobre 2009
- ▶ Blogueur spécialisé en veille, intelligence économique et recherche d'information depuis 2001 sur www.actulligence.com
- ▶ Formateur et consultant pour de grandes entreprises françaises
- ▶ Intervenant en troisièmes cycles spécialisés veille et intelligence économique
- ▶ Maître de conférence associé IUT de Montluçon – Université Blaise Pascal



Champ d'étude

- ▶ Les éditeurs de logiciels de veille francophones c'est-à-dire (liste non exhaustive)

1. Digimind
2. Ami Software
3. KB Crawl
4. Spotter
5. iScope
6. Synthesio
7. ...

- ▶ Il ne s'agit pas d'une étude « exhaustive »
- ▶ Les opinions émises n'engagent que moi



Constats préalables sur le marché

- ▶ Des éditeurs tous différentes les uns les autres par :
 1. Leur taille (bien que tous soient relativement petits)
 2. Les langages de programmation utilisés
 3. Les technologies embarquées ou développées
 4. Le champ fonctionnel
 5. L'ergonomie
 6. L'approche sectorielle ou pas
 7. La prestation de services & l'accompagnement

- ▶ Des clients de plus en plus matures aussi grâce à des formations initiales en veille / intelligence économique ne dédaignant plus les outils

- ▶ Un bouleversement assez profond du paysage de l'information numérique qui a eu un impact non négligeable



Les choix que nous allons aborder et détailler

- ▶ SaaS ou « in house »
- ▶ Spécialisations fonctionnelles
- ▶ Sourcing or not sourcing
- ▶ Scraping or not scraping
- ▶ Sémantique ou pas



SaaS ou in-house

► In-House

1. Meilleure confidentialité des sujets traités vis-à-vis de l'éditeur (secteur de la Défense et de la veille techno de pointe sensibles à cet argument)
2. Possession de l'application et des données
3. Modèle économique souvent différent (récurrent moins important souvent en Tierce Maintenance Applicative)
4. Ne pas partager les mêmes infrastructures que ses concurrents

► SaaS

1. Gestion plus simple du projet et délais de mise en place plus rapide. Idéal pour les projets de courte durée mais pas seulement
2. Maintenance et montée en version souvent plus simple quand externalisée
3. Anonymat mieux préservée (proxy du prestataire, requêtes sur les sites noyées dans la masse)
4. Contournement des politiques de filtrage internes



Spécialisations fonctionnelles

- ▶ Traiter tout le cycle de la veille ou seulement une partie?
- ▶ Certains outils se spécialisent sur la partie collecte
- ▶ D'autres sont plus fort sur la partie diffusion
 1. Newsletters automatisées
 2. Génération de rapport
 3. Minisites
 4. RSS
 5. ...
- ▶ La maîtrise technologique de tout le cycle de la veille reste complexe
 1. La partie collecte est souvent bien traitée mais avec des degrés d'automatisation ou une simplicité de mise en œuvre variable
 2. En général, carence des outils sur le traitement de l'information
 3. La partie collaborative est principalement orientée Workflow
- ▶ Le volet diffusion reste conditionnée fortement par la qualité de l'information et sa mise en forme



Sourcing or not sourcing

► Trois grands courants

1. Outils clés en main y compris sources surveillées
2. Do it Yourself
3. Un chemin intermédiaire : livraison de packages métiers, ou par type de sources (presse en ligne francophones, blogs techno, ...)

► Outils clés en main:

1. Facilité de mise en place d'un projet avec y compris l'intégration de sources spécifiques
2. Peu performants sur les domaines avec de nombreuses sources protégées
3. Qualité de l'information extraite souvent meilleure
4. Nécessite de bons outils de filtres (filtres booléens + « intervention humaine »)

► Do It Yourself

1. Peu pertinent pour la veille image « toutes citations »
2. Intéressant pour toutes les thématiques de veille avec un faible nombre de sources à forte valeur ajoutée
3. Pénible à paramétrer

► Packages de sources

1. Bien pour démarrer rapidement un projet tout en maîtrisant son périmètre de veille
2. Démarche de « démarrage »

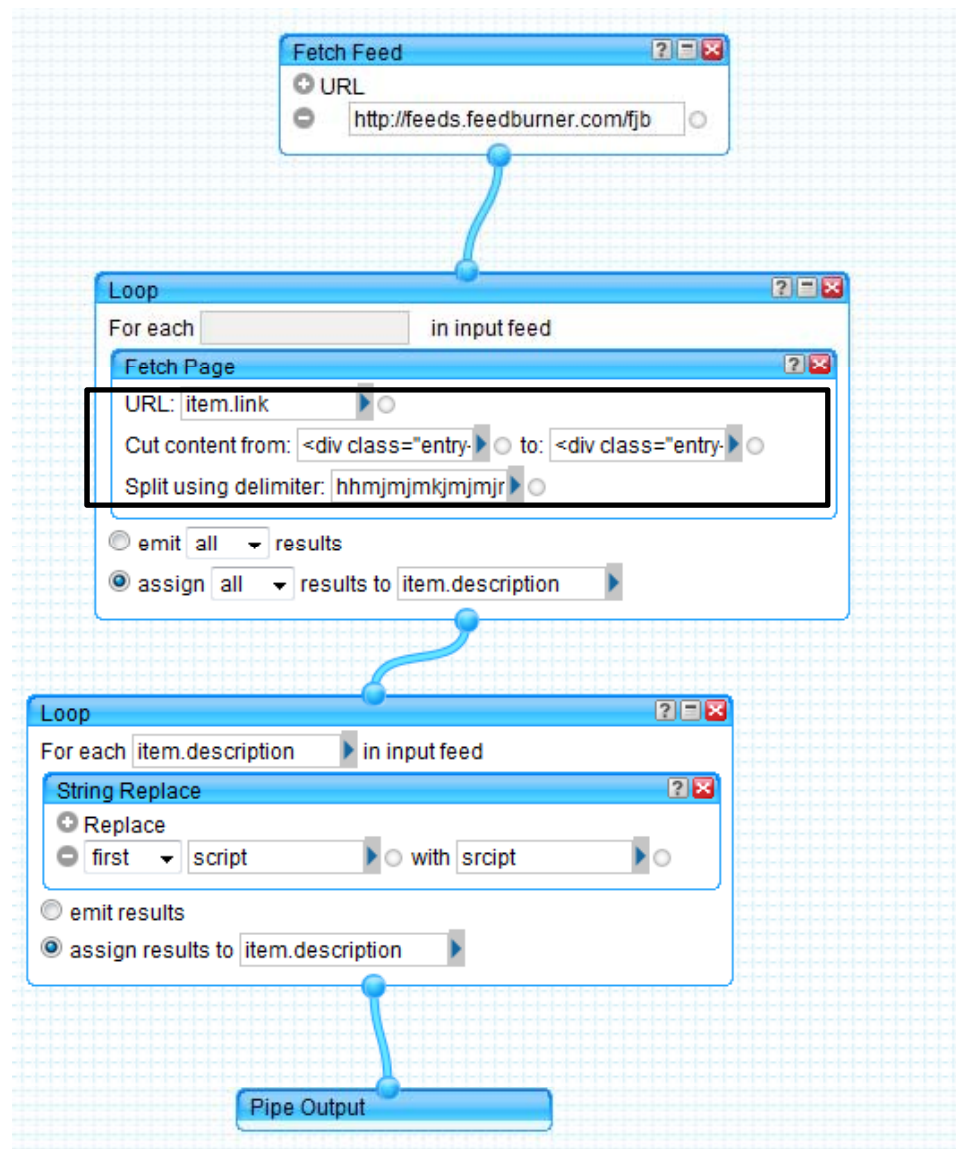


Scraping or not scraping (1)

- ▶ Fonctionnalité visant à rendre possible le passage d'un web structuré à une information structurée
- ▶ Niveau 0 : Surveillance d'une page, de ses modifications et / ou de ses nouveaux liens
- ▶ Niveau 1 : Sélection de certaines parties par :
 1. Filtres automatiques (exclusion de certaines formes type dates, chaînes numériques, ...)
 2. Couplage éventuelle avec une base de connaissance (formats publicitaires par exemple)
 3. Chaînes de caractères à l'intérieur des liens
- ▶ Niveau 2 : ciblage automatisé d'une zone intéressante (corps d'article)
 1. Définition des marqueurs à l'intérieur du code source : long, taux de réussite élevé, souvent outsourcé
 2. Algorithmes maison : facile, taux de réussite variable
 3. Utilisation du modèle DOM pour cibler toute une zone



Exemple d'utilisation de balises à l'intérieur du code source



Scraping or not scraping : Web Harvesting (2)

- ▶ Permet réellement de passer d'un contenu web déstructuré, monobloc à un contenu Web structuré, segmenté
 1. Soit découpage à la main
 - Très long
 - Se justifie pour les sites comportant de nombreuses pages
 - Ex : veille tarifaire
 2. Utilisation du model DOM
 - Plus simple à mettre en place
 - Interfaces graphique par sélection de zones ou par exploration du modèle DOM
 - Particulièrement utile pour la veille presse en ligne : extraction du titre, de l'auteur, de la date, de la rubrique, ... EX : <http://web-harvest.sourceforge.net/samples.php?num=4>
- ▶ Cette démarche reste longue à mettre en place et alourdit déjà une phase initiale de paramétrage difficile pour les utilisateurs
- ▶ L'annonce de l'adoption par Google des microformats et du RDF pourrait permettre d'atteindre plus facilement cet objectif



Sémantique or not sémantique

► Le sémantique peut-il facilement s'appliquer à la veille ?

1. Complexe car hétérogénéité des sources et des langages y compris parfois sur des veilles sectorielles et thématiques
2. Le sémantique sur des terrains inexplorés ou mouvants demeure lourd à maintenir
 - Base de connaissances
 - Arbres des connaissances
 - Taxonomies
3. Le sémantique est délicat dans des contextes internationaux, ses algorithmes reposant lourdement sur la structure de chacune des langues

► Quelles alternatives ?

1. Le contenu brut peut toutefois être enrichi par l'extraction des entités nommées, l'identification des prises de paroles, des verbes d'actions et de relation
2. L'utilisation d'algorithmes statistiques poussés peut améliorer l'expérience de veille. Le couplage avec un bon outil de search reste un moyen de concilier collecte et capitalisation en facilitant l'accessibilité des données



Et maintenant ?!

A vos questions !



Frédéric Martinet

Consultant Intelligence Economique, Veille stratégique et Veille image

Actulligence Consulting

+33 (0) 6 19 05 41 37

frederic.martinet@actulligence.com

www.actulligence.com

