Le métier de veilleur repose principalement sur l'analyse des variations d'information et sur l'identification des signaux faibles. Ce sont là deux des principales missions qui lui incombent au quotidien. On pourrait alors asséner des banalités sur la croissance de l'information accessible, sur la production de « connaissances » et sur la difficulté à s'y retrouver mais quelque soit cet ordre de grandeur, au-delà d'un certain seuil, la masse d'information implique inévitablement une démarche structurée, l'acquisition de « bonnes pratiques » et le recours à des outils.

Comment identifier les sources d'information sur Internet ?

Par Frédéric Martinet, KB Crawl SAS

Frédéric Martinet, consultant au sein de KB Crawl SAS et webmaster éditorialiste du site web dédié à l'intelligence économique, à la veille et à la recherche d'information www.actulligence.com

Le travail du veilleur porte finalement assez rarement sur la totalité de l'information. En général, ce dernier travaille sur un secteur économique spécifique, sur certains types d'informations (brevets, lois, articles de presse,…) et donc sur un périmètre réduit.

Comment définir correctement le périmètre de veille utile au client final ?

Le problème est de définir correctement ce périmètre utile non pas au veilleur mais à ses « clients » internes ou externes.

1. La première étape de toute démarche de veille repose donc sur l'appréciation et la définition de ce périmètre de veille. Pour une problématique donnée il s'agira de définir la typologie des sources d'informations à même d'alimenter le processus de veille en données brutes mais ciblées.

2. Une fois cette typologie de sources plus ou moins précisément établie, il s'agira pour le veilleur de se lancer dans une phase de sourcing visant à identifier précisément les sources qu'il lui faudra surveiller.

Le sourcing ne doit cependant pas aller sans une réflexion préalable sur le périmètre de veille et sur le type de sources à privilégier afin d'éviter toute dispersion superflue.

Comment établir une typologie de sources ?

Tout d'abord il faut comprendre que si l'on parle de définition initiale des typologies des sources utiles sur une problématique de veille, c'est bien parce que cette dernière sera probablement amenée à évoluer. Tout processus de veille est un processus itératif. Il l'est car la problématique initiale de veille évoluera sensiblement ou radicalement en fonction des résultats que la veille produira ou en fonction de variations de l'environnement politique, sociétal, juridique.

Afin de choisir les types de source à mettre en surveillance il faudra donc se poser les questions suivantes :

- Quelle est la criticité de la décision à prendre ?
- Qui peuvent être les détenteurs de l'information dont j'ai besoin ?
- Quel est le niveau de fiabilité de la source ?
- Quel est le délai dont je dispose avant d'identifier l'information ?

Comment partir à la recherche de sources ?

Il est parfois bon de rappeler l'évidence…mais il semble logique et efficace de commencer par effectuer une recherche interne. On balaiera les banques de liens du service documentation, ou bien diffusés sur l'Intranet, on interrogera les services a priori les plus concernés par le sujet. La collecte de favoris Internet référencés chez

Concrètement comment cela se traduit-il ?

Imaginons qu'une société de produits alimentaires soit soumise à un problème de contamination et qu'elle ait diffusé un avertissement pour un rappel de ces produits. Cette dernière, une fois l'urgence sanitaire contenue va devoir gérer le devenir de son image particulièrement important dans des entreprises parfois à faible rentabilité et relevant de besoins primaires (se nourrir). Si le consommateur a peur ou doute, il ne consommera plus.
La criticité de la décision à prendre est donc forte : communiquer ou ne pas communiquer vers le client, et de quelle façon, avec quel message.
Les détenteurs de l'information sont les consommateurs eux-mêmes, et de fait leurs opinions sont diffusées sur les médias qui leurs sont accessibles : magazines d'associations de consommateurs, forums, blogs. L'information peut, par ailleurs, être collectée par des intermédiaires tels que les instituts de sondage (enquêtes d'opinion, baromètres,...)
Le niveau de fiabilité n'est pas alors essentiel. L'important, dans ce cas précis, sera principalement la volumétrie des opinions négatives.
Enfin le délai est dans ce cas court. Pour enrayer l'hémorragie liée à l'érosion d'une image de marque il faut aller vite. On privilégiera donc les sources d'informations récentes (actualités, forums, blogs,...) ou influentes.
Ce premier travail fait, il permettra au veilleur de partir à la recherche de sources en y voyant plus clair sur les points de départ de sa recherche de sources et sur les options à privilégier lors de sa recherche.

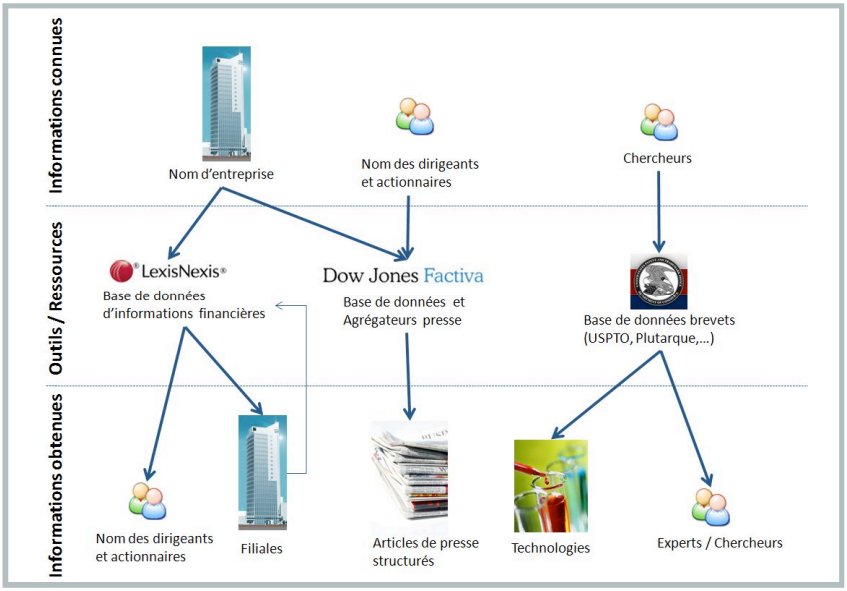
certaines personnes clés est aussi un des points à prendre en considération. Cette première phase permet souvent de collecter des sources utiles car préalablement qualifiées par un individu concerné par le sujet voire expert. D'ailleurs en dehors des sources d'information formelle, elle permettra aussi d'identifier des experts sur un sujet, des référents, des relais d'information qui doivent être considérés comme une source importante d'information.

La deuxième partie du sourcing s'effectuera à partir d'Internet qui permettra d'identifier des sources d'information sur Internet mais aussi des sources d'information hors-lignes : revues spécialisées, personnes ressources.

On pourrait distinguer l'information sur Internet en trois types : l'information structurée émanant de bases de données professionnelles d'Information Scientifique et Technique (1) et économique par ailleurs faisant souvent partie du Web profond (2) , l'information non-structurée (3) et les informations personnelles qui foisonnent désormais sur les réseaux sociaux.

Chacune de ces sources d'information comporte des outils de recherche avancés qui à partir de certains types d'entrée permettent d'obtenir des sorties.

La phase initiale de recherche interne aura permis de collecter des bribes d'informations essentielles qui permettront d'alimenter ces outils d'interrogation et de générer des sources d'information en sorties qui viendront alimenter le processus de veille.



Concernant l'information structurée on pourra par exemple :

- Injecter le nom d'une entreprise dans une base de données d'informations sur les entreprises afin d'en extraire le nom des dirigeants qui feront des mots clés efficaces pour surveiller les déclarations d'une entreprise et les activités tierces des dirigeants afin d'identifier leur stratégie de communication.
- L'utilisation du nom de certaines entreprises pourra permettre d'identifier d'autres filiales en utilisant là encore des bases de données financières et économiques.
- Les bases d'information structurée peuvent aussi permettre d'avoir accès à un fond documentaire important d'articles de presse à partir du simple nom d'une entreprise. Bien que difficiles à analyser, ces corpus souvent massifs peuvent être traités avec des outils de text-mining afin d'en extraire du sens. La structuration de ces informations rend alors l'utilisation beaucoup plus simple d'outils de traitements statistiques que dans le cas d'information déstructurée.
- Si l'on connaît le nom d'experts scientifiques ou de chercheurs, on pourra utiliser ces derniers dans une base de données de brevets (4) afin de surveiller les technologies sur lesquels ces derniers travaillent et par là même souvent celles sur lesquelles travaillent leur entreprise. Il sera par ailleurs possible via ce type de recherche d'identifier d'autres experts par l'analyse des co-citations.

Exemples d'utilisation d'outils de recherche d'information structurée pour l'identification de sources d'information et d'alertes complémentaires

notes ... notes ... notes ...

• (1) Dès les années 1970, le champ de l'IST fut étendu à l'information économique et financière. « Histoire de l'Information Scientifique et Technique » de Martine Comberousse – p.92 – Armand Colin.

• (2) Terme introduit en 2001 par Brightplanet, le Web profond correspond aux documents sur Internet non accessibles en texte intégral via des moteurs de recherche : espaces réservés aux inscrits, espaces payants, bases de données professionnelles payantes mais aussi sites non référencés, pages solitaires. « The Deep Web: Surfacing Hidden Value" – Michael K. Bergman – The Journal of Electronic Publishing – August, 2001, Volume 7, Issue 1.

• (3)L'information dite non-structurée ou déstructurée est une information dont les méta données (auteur, date, références, mots clés, abstact,...) ne sont pas clairement spécifiées et accessibles en dehors du corps du document rendant leur exploitation informatique complexe contrairement à l'information structurée contenue dans des bases de données professionnelles.
Cf « Valorisation de l'information non-structurée » par l'Apil, le Cigref et l'Aproged – Octobre 2007. L'information Web peut être considérée comme déstructurée, les méta balises étant souvent incomplètes voire erronées.

(4) Le site de l'Inpi par exemple en France – http://www.plutarque.com – ou bien encore le site http://www.uspto.gov/ du Bureau Américain des marques déposées et des brevets

L'information non-structurée est la plus délicate à exploiter.

D'abord elle n'est pas qualifiée et émane de sources dont il est difficile souvent d'estimer la fiabilité. N'importe quelle entreprise ou individu pouvant désormais publier de ´information sur Internet, cette dernière peut être sujette à manipulation, à de la désinformation et son caractère peu fiable renforce le risque d'intoxication informationnelle.

Les moteurs de recherche grands publics permettent toutefois rapidement d'élargir son champ d'exploration en phase de sourcing. Certains moteurs ont par exemple développé des algorithmes permettant de déterminer des sites « proches » thématiquement ou partageant un nombre de liens identiques ou adjacents conséquents . Par ailleurs, il est aussi possible de connaître les liens pointant vers un site Web . La notion de sérendipité s'applique alors : par exploration des liens hypertextes sortants ou entrants sur un site on pourra identifier des thématiques proches de celle(s) clairement exprimée(s) en phase initiale.

Avec quelques mots clés il est possible de rapidement établir les thématiques et les-sous thématiques d'un périmètre de veille en s'appuyant sur des outils de recherche en ligne proposant une clusterisation des résultats , c'est-à-dire une structuration à la volée en ensembles et sous-ensembles thématiques d'un corpus de sources répondant à une demande utilisateur.

Enfin les outils issus du web 2.0 sont aussi particulièrement riches et amènent surtout une qualification humaine non monétisée.

Concernant l'information sur les personnes, un simple nom peut aujourd'hui vous permettre d'établir un portrait précis d'un individu.

Exploiter un nom de personne en l'injectant dans un réseau social professionnel vous permettra de connaître ses anciens employeurs, ses amis, ses préférences politiques, les projets sur lesquels il a travaillé et d'avoir une approche qualitative de la recherche d'information. Pour un béotien sur une thématique à aborder, c'est là tout autant de pistes et de mots clés qui pourront permettre de démarrer efficacement une veille.

L'identification de sites ressources peut quant à elle s'effectuer par le biais d'outils de « social bookmarking ». A travers une simple URL vous pourrez identifier les sites que les utilisateurs de ces réseaux ont saisis dans le même répertoire ou définis avec les mêmes mots.

On le voit donc, le travail de sourcing nécessite une démarche structurée abordable y compris par des hommes n'étant pas familiers avec la thématique de veille. Une démarche initiale de recherche et d'identification de ressources internes permettra rapidement d'établir des points de départ qui une fois exploités correctement à travers différents outils ciblés permettront de définir à la fois mots clés de mise en surveillance et sources à surveiller. Loin des concepts et théories, la méthodologie proposée ici repose aussi sur une bonne connaissance des outils qui vont constituer autant de leviers multiplicateurs des bribes d'information collectées initialement.

FRÉDÉRIC MARTINET

LE SOURCING SELON KB CRAWL

Le livre blanc sur la veille réalisé par KB Crawl (www.kbcrawl.net/fr/telechargement/livre-blanc.html) y consacre tout un chapitre. Pour Frédéric Martinet et Antoine Montoux, "un sourcing correctement effectué permet à partir d'un petit nombre de sources initial, grâce à la nature même de la structure hypertextuelle de l'information sur Internet, de collecter un nombre important de sources complémentaires. Toutefois, sur certaines thématiques très pointues ou sur certains marchés de niche, le nombre de sources identifiables peut être parfois limité.
Le sourcing doit s'attacher à aller à partir de sources connues, déjà exploitées, réputées ou renommées vers des sources de moins en moins notoires.

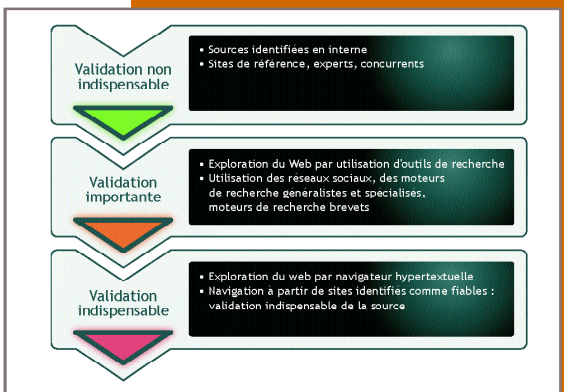


Fig 6 : Nécessité de validation des sources en fonction de leur qualification

Au fur et à mesure de cette exploration thématique du Web, la démarche de validation des sources deviendra de plus en plus essentielle.
L'ensemble des sources mises sous surveillance et leur qualification est un des maillons clés du dispositif de veille sur Internet. Il doit régulièrement être mis à jour.
Au-delà des sources qui disparaissent et qu'il suffi t de supprimer du dispositif de veille, de nouvelles sources peuvent apparaître. Afi n d'identifi er ces nouvelles sources, il est toujours nécessaire d'intégrer dans son plan de surveillance des annuaires thématiques ou des sources généralistes qui feront souvent référence à d'autres ressources. Une fois ces sources identifi ées, il suffi ra de les intégrer dans la surveillance automatique. On pourra par exemple mettre sous surveillance les espaces publicitaires des moteurs de recherche afi n de voir quel concurrent se positionne sur un produit ou les rubriques des annuaires correspondant à sa propre activité.".