Les fondamentaux de la veille

flibmann@fla-consultants.com

Spécialiste de l'information électronique (recherche à la demande, formation), François Libmann est ingénieur de l'École centrale de Lyon. Il a complété sa formation au Dartmouth College, aux États-Unis. Après avoir débuté dans l'innovation, il crée FLA Consultants (www.fla-consultants.com) en 1977. En 1995 il crée Bases Publications (www.bases-publications.com) qui édite les lettres *Bases* et *Netsources* consacrées à la recherche dans les banques de données et sur Internet avec un point de vue d'utilisateur exigeant.

Les grands serveurs toujours au service des veilleurs

[analyse] **Plus convivial, plus démocratique, souvent gratuit, le web et son accès aisé à l'information présente une alternative séduisante à la recherche sophistiquée, encore parfois ardue, et payante dans les bases de données professionnelles. Pourtant, pour un veilleur avisé, l'exigence de sources qualifiées et la pertinence des résultats finals passe nécessairement par l'utilisation conjointe des grands serveurs. Ceux-ci peaufinent leur approche depuis des décennies, en adaptant leurs interfaces de recherche au profil de leur utilisateur.**

Au cours des vingt dernières années, la démocratisation de l'internet et notamment l'avènement du web ont révolutionné l'accès à l'information et donc les pratiques de veille : il y a bien un « avant web » et un « après web ». Inventé en 1989 au CERN de Genève, le World Wide Web a, en effet, induit durablement de nouvelles règles du jeu informationnel qui se sont appliquées dans le nouvel espace ainsi créé. Les jeunes générations de veilleurs en oublieraient presque que Google n'a que dix ans et que les sources d'informations électroniques sont bien antérieures au web, et même à l'Internet. Les années 1970 ont ainsi vu la création de plusieurs bases de données qui perdurent encore aujourd'hui. Si l'on met de côté Lexis et sa clientèle juridique, l'essentiel du développement des serveurs se fit sur l'aspect scientifique et technique, en grande partie par l'informatisation des bulletins d'*abstracts* qui existaient dans la quasi-totalité des domaines scientifiques.

De la DSI aux interfaces conviviales du web

Certes, les « serveurs de bases de données[1] » n'ont pas été créés spécifiquement pour la veille (veille historiquement « technologique »)… Mais ils se sont rapidement imposés comme des outils incontournables grâce aux informations qualifiées fournies par des producteurs reconnus et à des outils de recherche puissants. Outre l'utilisation pour des recherches rétrospectives, la formule DSI (diffusion sélective d'information aussi appelée « profil ») s'est très largement répandue et a constitué pendant de très nombreuses années l'outil de veille fondamental pour un très grand nombre de chercheurs.

Néanmoins, l'accès direct à ces richesses est resté longtemps limité aux seuls professionnels de l'information, souvent documentalistes, formés au langage d'interrogation du serveur, mais c'était le prix de la performance… et le seul moyen de l'utiliser. Notons que, bien avant la démocratisation de l'Internet, les grands serveurs avaient fait des tentatives pour élargir leurs clientèles vers l'utilisateur final, mais toutes furent vouées à l'échec. Il a fallu attendre le développement d'Internet et avec lui d'interfaces conviviales pour que les utilisateurs finals accèdent directement tant aux serveurs classiques qu'aux nouveaux services qui se sont développés (des serveurs tels CSA Illumina ou Scopus ou des accès directs aux sites de certaines banques de données).

Aujourd'hui, tous les serveurs utilisent Internet comme « tuyau » entre eux et leurs utilisateurs après avoir abandonné la norme X25 (Transpac). Certains comme Dialog ou STN ont développé des interfaces spécifiques « *end-users* » en maintenant et développant par ailleurs celles destinées aux professionnels de l'information. D'autres, comme SCOPUS dans le domaine scientifique, Factiva et L'Européenne de données, tous deux plutôt dans l'actualité, ont développé des interfaces qui peuvent être utilisées soit de façon élémentaire comme on utilise en général Google, soit de façon sophistiquée avec de nombreuses possibilités disponibles.

1 On appelle serveur de bases de données un (gros) ordinateur dont les mémoires contiennent un grand nombre de banques de données structurées ainsi qu'un langage d'interrogation spécifique. Les exemples les plus connus sont Dialog, Questel ou STN. Plus récemment sont apparus des agrégateurs, proches dans le concept, qui regroupent sur leurs machines le texte intégral (le plus souvent) de milliers ou de dizaines de milliers de sources, en général de presse, sur lesquelles une recherche simultanée est possible. On peut citer Factiva, LexisNexis ou EDD comme exemples.

QUAND SPOUTNIK « DÉCLENCHE » DIALOG

Tout démarre lorsque les Américains découvrent avec stupeur, le 4 octobre 1957, que les Russes, en lançant le premier satellite, les ont devancés dans la course à l'espace. Très vexés, ils cherchent comment rattraper leur retard. Parmi les solutions évoquées figure en bonne place la possibilité d'accéder plus facilement et plus rapidement aux archives des publications scientifiques et différentes mesures sont alors prises. Roger Summit entre comme stagiaire à la Lockheed Missiles and Space Company avec le directeur du traitement de l'information dès 1960. Ses travaux débouchent sur la création en 1964, dans cette société, d'un laboratoire destiné à explorer les possibilités de recherche automatique d'information.
Le premier prototype de Dialog, serveur qui existe toujours, est lancé début 1967 avec les 200 000 références de la banque de données STAR de la Nasa.
En 1969 est ajoutée la banque de données Eric sur les sciences de l'éducation et la décision est prise en 1970 de transformer Dialog en service commercial.
Dans le même temps ou presque, fin 1972 ou début 1973, selon les sources, est lancé un système/serveur de même nature baptisé Orbit au sein de SDC (System Development Corporation, un *spin-off* de la RAND Corporation) à Santa Monica en Californie. Notons que SDC Orbit fut le premier à offrir l'information brevet en chargeant World Patents Index de Derwent en 1976. Après avoir eu de nombreux propriétaires successifs, Orbit est racheté par Questel en janvier 1994 et la fusion des deux serveurs est réalisée quelques années après.
Toujours au début des années 1970, en 1973 précisément mais procédant d'une démarche différente dans un milieu différent, celui des avocats, Mead Data lance le service Lexis proposant des textes juridiques en texte intégral, saisis à Taïwan, bien avant l'OCR, parce que les opérateurs de saisie ne comprenant pas l'anglais faisaient moins d'erreurs. Cette activité sera revendue à Reed Elsevier fin 1994. Mead Data Central lance par ailleurs Nexis en 1979, selon le même principe que Lexis mais pour la presse en démarrant avec... six sources (plus de 10 000 aujourd'hui).
On voit donc que la décennie 1970 fut décisive pour les banques de données avec des initiatives américaines, les Européens ne s'y intéressant que dans un deuxième temps, d'abord avec le serveur IRS de l'Agence spatiale européenne puis, au niveau français, avec la création de Télésystèmes Questel en 1979.

L'ère des sources gratuites

Dès avant les années 2000, du fait de l'abondance toujours plus grande de l'information gratuite sur le web et de l'hypothèse sous-jacente qu'il n'y avait plus de questions complexes, de nombreux « gourous » avaient prédit la mort des serveurs professionnels et la disparition des professionnels de l'information tels les cochers de fiacres. C'est ce qui a présidé au lancement en 1997 de Reuters Business Briefing[2] service payant et aux possibilités de recherche très limitées mais qui permettait de visualiser un très grand nombre de documents (la facturation était basée sur un faible coût horaire de connexion).

Par ailleurs, les sources gratuites se sont multipliées. On peut les classer en quatre catégories :
- toutes les informations mises en ligne par les internautes eux-mêmes ; ceci est illustré aujourd'hui par le monde du web 2.0 ;
- les informations mises à disposition par des entreprises (données commerciales en particulier) ou des associations de certains domaines, comme celui de l'environnement ;
- tout ce qui relève de l'*open acces*, en général des publications scientifiques, grâce à un nouveau modèle économique : l'auteur (en fait sa structure) supporte les frais d'édition, l'accès à l'information étant gratuit en particulier sous sa forme électronique ;
- les informations autrefois payantes et mises en ligne gratuitement pour des raisons politiques. C'est alors le contribuable qui paie, même s'il s'agit souvent d'information professionnelle. C'est le cas pour TED (appels d'offres européens), Celex (droit européen), Legifrance (droit français), etc.

Par ailleurs, des moteurs de recherche gratuits tels que Google Scholar ou SCIRUS d'Elsevier[3] se sont développés pour donner accès à l'information scientifique. Si la qualité de Google Scholar est assez souvent critiquée[4], il n'empêche que cet outil est énormément utilisé.

Ces moteurs ne donnent cependant pas nécessairement de lien vers l'information totalement gratuite (le document primaire en texte intégral), en fournissant parfois comme résultats des références bibliographiques issues d'un serveur payant.

Des serveurs toujours incontournables

Malgré ce développement du gratuit ou du pseudo gratuit, les serveurs existent toujours et se portent plutôt bien. On observe à la fois des concentrations et des rachats d'entreprises, signe de vitalité du secteur[5] Les serveurs ne sont évidemment pas restés sans évoluer et ont pour la plupart développé des outils permettant de visualiser les résultats de façon plus performante ou d'obtenir des représentations graphiques sophistiquées de ces résultats.

Ils ont aussi, surtout dans le domaine des brevets, cherché à élargir leur clientèle aux ingénieurs des services de R&D.

Mais si les serveurs agrégateurs se portent bien parce qu'ils ont suffisamment d'utilisateurs, on observe dans le même ///////

temps, dans nombre de centres de recherche, un mouvement que l'on peut qualifier d'inquiétant. En effet, certains centres de documentation ayant vu leurs effectifs fortement réduits ou ayant tout simplement disparu, les chercheurs effectuent leur veille eux-mêmes.

Ils utilisent des stratégies très simples, voire simplistes, avec des outils qui ne sont en général ni les plus complets ni les plus performants, considérant que la connaissance qu'ils ont de leur domaine de recherche leur permettra de séparer le bon grain de l'ivraie.

Or on peut faire l'hypothèse que ces recherches sont moins performantes que celles mises au point en collaboration avec des professionnels de l'information. On peut donc légitimement s'inquiéter des conséquences à plus ou moins long terme sur la qualité des résultats des recherches.

Un autre facteur limitant souvent l'utilisation des grands serveurs (mais hors de l'industrie pharmaceutique ou pétrolière, par exemple) est la dramatique absence de culture des dirigeants sur cette question. Leur connaissance se limite la plupart du temps à Google, outil simple et gratuit dont l'existence leur semble justifier de ne pas recourir à des professionnels et des outils professionnels payants.

La profession devrait donc d'urgence lancer des campagnes d'image en direction des dirigeants.

La facilité en général à l'encontre de la qualité

Aujourd'hui, les serveurs tels EDD, Factiva ou Lexis-Nexis proposent des interfaces qui s'adressent tant aux professionnels qu'aux utilisateurs finals ; ces derniers toutefois, dans la majorité des cas, appliquent à ces outils les mêmes recettes « simplistes » qu'ils emploient avec Google, et sont donc très loin d'exploiter toutes la puissance des moteurs. Or, sauf pour les quelques questions réellement simples ou les démarches de première approche (tout comme d'ailleurs sur les moteurs du web), la réflexion sur les stratégies de recherches et sur la formulation des requêtes permet d'obtenir des résultats plus pertinents, et la formation des professionnels comme des utilisateurs reste un point essentiel pour des veilles électroniques de qualité. Il faudrait enfin considérer ou plutôt reconsidérer que la première valeur ajoutée est dans la recherche elle-même avant d'être dans le traitement des résultats. C'est ce que vise à montrer les concours des stratégies de recherches organisé par *Bases* et i-expo en association avec l'ADBS.

Il ne faut enfin pas confondre la recherche sur un sujet précis pour lequel de bonnes stratégies sont nécessaires et les résultats en nombre suffisamment réduit pour être traitées à la main, avec les stratégies de *text mining* utilisant des outils qui « ramassent » un grand nombre de documents, en général sur des sites gratuits, et qui en font une analyse statistique et/ou sémantique.

Vers une approche multisources

Aujourd'hui, la veille sur les sources électroniques passe obligatoirement par le web visible comme invisible. Mais, face à des informations toujours plus nombreuses et éclatées, à des flux de moins en moins maîtrisés, on constate chez les professionnels de la veille une attention renforcée à la qualification des sources d'information : le travail sur ces sources (re)devient capital, et on accepte à nouveau (quoique avec réticence) que l'information à valeur ajoutée puisse être payante.

Les serveurs d'information professionnels ou les services de médiation sont souvent utilisés avec plus d'intelligence, après un travail de recherche sur le web, et les deux chaînes d'acquisition, au lieu de s'ignorer ou de se combattre, se complètent ainsi harmonieusement.

2 Voir *Bases*, n° 131, septembre 1997 : « Reuters : Cap vers le simplisme ».
3 Voir *Bases*, n° 244, décembre 2007 : « Le positionnement délicat et subtil de SCIRUS ».
4 Voir par exemple le « Pan » de Peter Jasco dans le numéro de mars-avril 2008 du magazine *américain Online*.
5 Parmi les plus récents on notera : la reprise par Dow Jones des 50 % de Reuters dans Factiva dont il est maintenant le seul propriétaire ; le rachat de Reuters par Thomson ; le rachat de Dialog (les serveurs Dialog et DataStar) par Proquest, nouveau nom de CSA (Cambridge Scientific Abstracts) après qu'il eut racheté... Proquest

EXEMPLE DE RECHERCHE ELABOREE EFFECTUEE SUR LE SERVEUR PRESSEDD