DES PHARES DANS LA NUIT

LA RECHERCHE DOCUMENTAIRE SUR INTERNET

PIERRE-MARIE BELBENOIT-AVICH
Service commun de la documentation, Lyon I

Le développement des réseaux Internet a déjà fait couler beaucoup d'encre : tellement peut-être que certains fantasment sur les avantages qu'ils pourront en tirer. Il est assez intéressant, amusant dirais-je, de voir se répéter, à chaque étape du transfert de l'information, les mêmes fantasmagories, au sens du Petit Larousse : *« Procédé qui consiste à faire apparaître des formes irréelles dans une salle obscure... ; spectacle enchanteur, féerique ».*

La recherche en ligne devait résoudre tous les problèmes, à leur tour les CD-ROM apparurent comme la panacée, et maintenant voici qu'Internet est vécu par beaucoup comme la solution idéale, le remède à tous les maux de l'information documentaire. Il n'est pas lieu ici de préciser les limites des CD-ROM au sens strict, mais on pourrait citer un article anglo-saxon récent qui traite ce support, avec un brin de provocation, d'antédiluvien. *Sic transit gloria mundi...*

La multiplicité des sites

Quant à Internet, le premier malentendu n'a-t-il pas porté sur son nom ? *Super Highways :* les Autoroutes de l'information ? Comme si Internet n'était pas plutôt seulement fait de voies à un ou deux sens, enchevêtrement se prolongeant à l'infini, bref, intrinsèquement, un WEB, ou en bon français une toile d'araignée. Il n'est que de voir comment on y saute littéralement d'un site à l'autre, nonobstant la distance.

Plus encore, le problème fondamental d'Internet réside sans doute dans la multiplicité du nombre de sites qui y fourmillent. Il ne s'agit pas, comme c'était le cas pour les serveurs en ligne traditionnels, de points d'accès bien connus, définis quant à leur accès, leur contenu et leur langage d'interrogation. Cette façon d'obtenir l'information était relativement simple, à condition de savoir manier tel ou tel logiciel. Travailler sur Dialog, Data-Star, ou STN équivalait à pouvoir utiliser des bases bien connues où l'on était en quelque sorte entre « gens de bonne compagnie », autrement dit, entre experts se distinguant des néophytes.

Internet offre, bien au contraire, une information disséminée à l'extrême. En septembre 1995, on y estimait le nombre de sites commerciaux supérieur à 6 200 000 unités. Il en existe largement plus aujourd'hui. Comment tous les connaître ? Comment tous les identifier ? Et comment les juger ? Abîme sans fond d'information ! Masse mouvante et évolutive, puisque, à l'été 1995, la durée moyenne mondiale d'un site était évaluée – car que faire d'autre ? – à six mois. Cela a fait la fortune de livres intitulés *Pages jaunes d'Internet,* comme si la forme papier, avec ses inévitables lenteurs, pouvait suivre le rythme du monde virtuel. On ne doit pas s'étonner de voir déjà bradés des livres publiés en 1993 sur ce sujet.

La complexité d'Internet ne réside pas seulement en cette multiplicité des sites. Elle consiste aussi dans le fait qu'ils sont accessibles sur diffé-

rentes voies d'accès : les News Groups ou groupes de discussion, ou encore forums, qui autorisent des discussions constantes, impromptues et informelles avec tous ceux qui, sur la planète, s'intéressent à tel ou tel sujet ; en février 1996, on estimait leur nombre à près de 12 000. S'y ajoute un type de site qui a eu son heure de gloire en 1993 – ce qui paraît déjà vieux – le WAIS ou Wide Area Information Server, dont on s'attendait à ce qu'il permette, grâce à une interface, d'interroger des bases sans avoir à se soucier de leurs logiciels d'interrogation. Mentionnons aussi Gopher, créé à la même époque, par l'Université de Minnesota, qui voulait *« creuser des galeries souterraines »* d'un site à l'autre, à l'image du rongeur canadien dont il porte le nom. Son principal inconvénient se situe dans la lourdeur de son arborescence qui le rend malcommode à utiliser.

La mode est maintenant au WEB (World Wide Web), qui contient des sites identifiés par leur adresse électronique. Il est assez facile de s'y créer une page, comme l'on dit, et nombre de sociétés de tous poils, ou, en ce qui nous concerne, de bases documentaires, s'y sont engouffrées, faisant tout d'abord miroiter un accès gratuit, vite devenu payant. Mais, là encore, impossible de s'y retrouver. On a vite réalisé qu'ajuster ensemble des sites n'offrait aucun intérêt. Pour les rendre attractifs, intéressants et utilisables, il fallait créer des bases et des logiciels qui liaient ces bases à des interfaces ou des outils de développement.

L'évaluation des sites

Le problème sera ensuite l'analyse de ces sites, dont le nombre double plus ou moins tous les trois mois, d'autant qu'ils sont dynamiques et que leur adresse et leurs objectifs changent.

La grosse difficulté pour indexer les sites Internet est en outre qu'ils sont continuellement évolutifs, alors que l'indexation d'un article est faite une fois pour toutes. Il faut donc veiller à ce que l'analyse qui en est faite reste continuellement valable et à jour. Il faut estimer, juger, évaluer ces sites. On l'a écrit : sur Internet, il y a beaucoup d'informations, mais toutes n'ont pas la même valeur. Et d'autres auteurs d'ajouter : *« Ce dont on a besoin aujourd'hui, c'est d'une intelligence sélective, parce que l'information intelligente est souvent le résultat de tri, de classement ; les moteurs générateurs d'intelligence auront donc à examiner de vastes volumes de références à partir d'archives électroniques »*.

Les utilisateurs ont en effet tendance à consulter les résultats sans se poser de questions, en particulier sur la nature ou l'exhaustivité de la base qui est proposée. Ce sera la responsabilité des bibliothécaires et documentalistes que de se tenir informés de la nature de ces bases. Ce ne sera plus leur rôle de faire eux-mêmes la recherche, mais plutôt de guider les utilisateurs au travers d'un labyrinthe sans cesse grandissant, d'expliquer où trouver l'information pertinente.

Les moteurs de recherche

S'est donc créé et imposé en 1995 le concept de moteur de recherche sur Internet *(search engine)*. Les grands groupes les ont lancés pour faire face à la recherche d'informations. Avant de parler de certains d'entre eux, nous pourrions en quelques mots expliquer leur mode de fonctionnement.

Ces machines peuvent utiliser deux techniques d'indexation, deux manières d'analyser l'information brute et dispersée : d'une part – et c'est le cas du moteur de recherches bien connu Yahoo – l'indexation se fait sur les résumés descriptifs des sites. Elle se veut exhaustive et sans notion de pertinence. Dans d'autres cas, et on mentionnera ici Webcraler, Lycos, Infoseek, Opentext, Excita, Altavista, Mackindey Directory, la recherche se fait par mots-clés.

Elle peut aussi se faire de manière informatique ou – et c'est alors beaucoup plus pertinent – grâce à une équipe d'indexeurs qui analysent non seulement les nouveaux sites mais aussi les anciens pour voir si leur contenu et leurs objectifs n'ont pas changé. C'est le cas par exemple d'Electronic Information Engineering, qui emploie une demi-douzaine de personnes dont le travail est de vérifier continuellement chaque site Web. Par ailleurs, ce moteur fait un contrôle de qualité, si bien que l'indication d'un site est en soi un critère. Ce contrôle évalue la rapidité de l'information, l'exhaustivité et le lien

LA GROSSE DIFFICULTÉ POUR INDEXER LES SITES INTERNET EST QU'ILS SONT CONTINUELLEMENT ÉVOLUTIFS, ALORS QUE L'INDEXATION D'UN ARTICLE EST FAITE UNE FOIS POUR TOUTES

avec un service de fourniture de l'information et offre ainsi une valeur ajoutée qui estime la qualité et fournit un certain nombre de commentaires.

Ce service, comme beaucoup d'autres, offre une « salle de lecture virtuelle », qui est seule à être gratuite, et met à disposition des usagers dictionnaires, ouvrages de références, thésaurus, listes d'acronymes, glossaires et un forum de formation. Electronic Information Engineering analyse aussi bien des bases gratuites que payantes. Lycos, de son côté, employait, en décembre 1995, une trentaine d'indexeurs. Au contraire, Infoseek a, pour recenser les sites Internet et les fournir au travers d'un écran frontal d'accueil, adopté une approche automatique.

Certains de ces moteurs de recherche sont encore gratuits, d'autres sont

payants, tel par exemple, Engineering Information Village, qui veut offrir un ensemble de données formelles et informelles. Les usagers individuels paient 50 $ par mois et les institutions 8 000 $, pour un maximum de huit utilisateurs. Environmental Route Net propose en combinaison les données propres de l'éditeur Cambridge Scientific Abstracts et des sites Web. La vraie valeur de ce service réside dans le nombre des sites analysés : près de 6 000 *home pages*, dont 600 seulement ont été retenues. Le coût est de 1 999 $ par an, sans charge supplémentaire par document retenu.

D'autres moteurs ont préféré des financements publicitaires tel Webcrawler, maintenant possédé par America Online.

Les meilleurs

Passons en revue quelques-uns de ces moteurs d'indexation :

– ***Infoseek*** est un site qui peut traiter 100 000 sites différents. Pendant les périodes de pointe, 500 connexions sont analysées par minute ; la plupart d'entre elles prennent moins d'une seconde. En septembre 1995, Infoseek prévoyait de doubler l'information dont il disposait, en dépouillant en particulier des News, des sites Internet en FTP ou Gopher, et la plupart des listes d'adresses Internet connues. Une façon intéressante de mener une recherche correctement est d'abord de travailler par thème ou sujet, et ensuite d'introduire un mot spécifique.

Infoseek offre aussi un service commercial d'accès aux bases de données dans un éventail très large, à prix réduit. Enfin, les utilisateurs peuvent se créer un profil personnalisé, avec les références classées selon un ordre de pertinence. Lancé en février 1995, Infoseek revendiquait 60 000 utilisateurs sept mois plus tard.

– ***Folio Retriever*** autorise l'utilisateur à organiser sa documentation sur Internet par centres d'intérêt. C'est une approche originale. Ici, l'utilisateur ne se voit pas offrir une indexation, mais indique l'adresse du site URL qu'il veut joindre, ainsi que le nombre maximum de liens hypertexte qu'il désire employer. Folio Retriever fait également une analyse et une sélection des meilleurs sites.

– ***StarWeb*** a été créé par Cuadra à Los Angeles. Il utilise une information déjà existante : description des références et texte intégral. Cette machine offre une interface CGI-Bin qui connecte le logiciel Star à n'importe quel site Web. Elle donne une information qui a été contrôlée et l'accès à des sites différents, à partir d'un seul logiciel.

– ***World Trade Centre State*** fournit une information non seulement commerciale, mais aussi concernant toutes les publications gouvernementales, les News et services, et même un centre d'information médicale. Il ressemble à Lycos, Webcrawler et Yahoo.

– ***Electronic Library*** fournit quant à lui dix millions de pages en texte intégral pour 9$95 par mois. Il dépouille 150 titres de périodiques, des milliers de livres, plus de 20 000 photos et en sélectionne la partie la plus pertinente.

– ***Altavista***, créé plus récemment en janvier 1996, utilise la logique booléenne et peut être restreint à un champ. Les utilisateurs peuvent chercher par mot, phrase, titre, adresse URL et serveur.

– ***Yahoo.*** Aucun site n'est aussi populaire que Yahoo… Créé en 1994 par deux étudiants de la Stanford University, il revendiquait à la fin de 1995 trois millions de recherches par jour, soit un équivalent de 300 000 utilisateurs. Il offre également des possibilités de recherche en texte intégral.

– enfin, ***OCLC***, bien connu des bibliothèques universitaires, intervient également dans ce créneau et propose une approche un peu différente. Il a créé Netfirst, qui détaille les sources Internet et indexe les pages Web, les Usenet News Groups, les sites FTP, les journaux électroniques, les catalogues de bibliothèques. C'est une base de données avec format structuré, descripteurs et utilisation de la classification Dewey. Les sites y sont recensés s'ils sont assez importants et si leur durée de vie prévisible semble assez longue. Ce site est également offert sur CD-Rom.

Liste de quelques sites

http://altavista.digital.com	(Altavista)
http://directory.net	(Directory de 9 000 sites)
http://www.endinfosys.com	(Voyager)
http://www.edoc.com/ejournal/	(Virtual library electronic journals)
http://www.folio.com	(Folio Retriever)
http://www.hum.gu.se/w3vl/w3vl.html	(Virtual library on humanities)
http://www.internetwfc.com	(World trade center)
http://www.infomkt.ibm.com	(Infomarket)
http://www.infoseek.com	(Infoseek)
http://www.infotrieve.com	(Infotrieve medline)
http://krscience.dialog.com	(Basescience)
http://www.lycos.cs.cmu.edu/	(Lycos)
http://www.medscape.com	(Medscape)
http://webcrawler.com	(Webcrawler)

Moteurs et CD

Une autre sorte de moteurs d'indexation peut être utilisée sous forme CD avant de passer sur Internet. On mentionnera Verity. Il comprend trois outils d'indexation puissants :

– Topic remote Web, qui permet d'indexer les pages de texte HTML ;

– Topic file indexer, qui construit et offre un index avec thésaurus hiérarchisé ;

– un système d'indexation avancé, qui inclut l'optimisation des CD-Rom et un support d'indexation Acrobat, PDF et SGML.

Verity était, à la fin de 1995, utilisé par plus de 650 sociétés dans le monde. Les données Internet sont insérées sur CD, puis vérifiées et augmentées à chaque connexion. Le support CD est très intéressant pour travailler sur des machines d'indexation, car il permet une approche rapide, appréciable pour tous les services payants, comme l'est précisément Verity.

Mais la recherche documentaire sur Internet ne se fait pas seulement par le biais des machines à indexer. Il est fascinant de voir que les serveurs ou bibliographies traditionnelles ont déjà plus ou moins occupé cette niche commerciale. Ainsi Infotrieve a ouvert un accès Medline sur Internet, proposé directement à l'utilisateur final ; tout comme Sciencebase, cette fois créé très récemment par Dialog, et qui permet de travailler simultanément sur plusieurs bases bibliographiques, quelle que soit la discipline, sans connaissance préalable du thésaurus ou du logiciel d'interrogation.

NOTRE RÔLE N'EST PLUS CELUI D'UN INTERMÉDIAIRE OBLIGÉ, INDISPENSABLE, ENTRE L'UTILISATEUR ET LA CONNAISSANCE

Les phares et les bibliothèques

On peut se demander finalement en quoi ces phares de la documentation Internet concernent les bibliothèques.

Les machines d'indexation touchent tous les secteurs de la connaissance : lettres, droit, sciences économiques, sciences pures et appliquées, médecine, pharmacie, etc., tous domaines dans lesquels la recherche d'informations, exhaustive, rapide, complète et fiable, est essentielle. Les bibliothèques sont concernées par l'offre documentaire d'Internet, tout comme elles l'ont été par les bases en ligne. La différence majeure est que l'utilisateur final, quel qu'il soit, bénéficiera d'un accès direct aux autoroutes de l'information.

Notre rôle n'est plus celui d'un intermédiaire obligé, indispensable, entre l'utilisateur et la connaissance. Nous pouvons avoir une fonction de conseil, d'orientation, de veille documentaire enfin, qui permette aux utilisateurs des bibliothèques de mieux appréhender l'ensemble de ces nouvelles ressources.

Avril 1996

BIBLIOGRAPHIE

Andrieu, Olivier. – « Le langage d'interrogation d'Internet est purement un langage documentaire », *Archimag*, mars 1996, n° 92, p. 20-21.

Belbenoit-Avich, Pierre-Marie. – « La bibliothèque électronique, bibliothèque d'aujourd'hui ou de demain ? », *Bulletin des bibliothèques de France*, 1993, n° 6, p. 60-65.

Blake, Paul. – « Database traditional get caught up in the Web », *Information Today*, September 1995, p. 17-18.

« Databases specialists add value to the Web », *Information World Review*, July/August 1995, p. 1- 3.

Gartner, Richard. – « The art of tracking humanities », *Information World Review*, March 1996, p. 24–25.

Hekins, Lelia. – « Seeking out the search engine for you », *Information World Review*, January 1996, p. 12.

Maignien, Yannick. – « La Bibliothèque virtuelle ou de l'*Ars Memoria* à *Xanadu* », *Bulletin des bibliothèques de France*, n° 2, 1995, p. 8-17.

Maignien, Yannick. – « *Lector ex machina* », *Le Débat*, n° 86, sept.-oct. 1995, p. 155.

Miller, Tom. – « Three steps to the Net Heaven », *Information World Review*, October 1995, p. 31.

Ojala, M. – « The pitfalls of simpler searching... and the potential for professional expertise », *Information World Review*, July/August 1995, p. 8.

Schildermans, Jozef. – « De l'importance d'une bonne requête », *Data News*, 15 mars 1996, p. 29–31.