

Cour 02 : Internet et recherche biologique

Le réseau internet révolutionne peu à peu les pratiques des chercheurs tant du point de vue de leurs recherches documentaires que du point de vue de la diffusion de leurs travaux. Si l'opportunité d'accéder à une masse d'informations colossale « d'un simple clic » est une perspective particulièrement séduisante pour tout chercheur en biologie, la découverte d'informations pertinentes dans cet « océan » s'avère en réalité relativement difficile, et cela malgré l'existence d'un nombre croissant d'outils de recherche mis à la disposition des internautes.

Le nombre de pages web accessibles à partir des différents outils de recherche d'informations est en constante progression. Cette profusion d'informations disponibles sur le web est bien entendu bénéfique aux utilisateurs, car l'augmentation de la taille de la base de données interrogée accroît la probabilité de trouver des informations pertinentes sur un sujet donné.

Le mode de fonctionnement des outils de recherche sur internet peut être classés en trois catégories:

- les moteurs de recherche (ou robots) dont l'objectif est d'indexer en « texte intégral » les pages web, sans intervention humaine et sans critères de qualité associés, dans une base de données (on parle également d'index) dans laquelle les usagers peuvent effectuer des recherches par mots clés, *via* une interface spécifique.
- les méta-moteurs, dont le principe est d'interroger simultanément, *via* une interface unique (site web ou logiciel installé sur un poste client), un nombre plus ou moins élevé de moteurs de recherche différents.
- les annuaires (ou répertoires), dont les principales caractéristiques sont leur mode d'organisation hiérarchique et la possibilité de recherche en « furetant » dans différentes catégories. Contrairement aux pages web présentes dans l'index des moteurs de recherche, les pages répertoriées par les annuaires sont sélectionnées sur des critères de qualité et ne sont pas indexées en texte intégral ; ce sont des notices descriptives créées par des éditeurs qui sont indexées.

En outre, les domaines de connaissances couverts par les moteurs et annuaires de recherche sont variables. On distingue ainsi les outils dits généralistes, qui concernent tous les domaines

de connaissance, des outils sélectifs ou spécifiques qui couvrent un champ scientifique ou disciplinaire donné.

2. Outils de recherche étudiés

Le nombre d'outils de recherche d'informations disponibles sur internet étant très élevé, pour des raisons de faisabilité, seul un nombre limité de 8 outils a été inclus dans cette étude, 5 moteurs, 1 méta-moteur et 2 annuaires.

2.1. Moteurs de recherche

- *Altavista* (<http://www.altavista.com>) est, probablement en raison de son ancienneté (1995), un des moteurs généralistes les plus populaires et les plus fréquemment testés
- *Google* (<http://www.google.com>), apparu assez récemment (1998), est le moteur généraliste qui possède à l'heure actuelle l'index le plus important en terme de nombre de pages indexées (plus de 1,5 milliards).
- *Bioview* (<http://www.bioview.com>) est un moteur spécifiquement dédié aux biologistes avec un index composé de pages web uniquement sélectionnées dans le domaine de la biologie.
- *Scirus* (<http://www.scirus.com>) permet d'interroger l'index généraliste de *Fast* (<http://www.fast.com>) limité aux seules pages web ayant un contenu scientifique. Il permet également d'interroger 4 bases de données d'Elsevier dont l'accès est subordonné à la souscription d'un abonnement. Cette dernière possibilité n'a pas été exploitée pour cette étude car l'information proposée n'est pas en libre accès.
- *Search4science* (<http://www.search4science.com>) est un moteur spécifiquement dédié à la recherche d'informations en sciences. Il offre la possibilité de rechercher dans deux index différents: (1) celui de *Northern Light* lorsque l'utilisateur opte pour l'option de recherche *Dynamic search*. Cette interface propose pour chaque terme de recherche saisi par l'utilisateur une série de synonymes, permettant d'élargir ou de restreindre la recherche en cours. Il présente la spécificité de regrouper les résultats dans des répertoires pour faciliter leur exploitation; (2) celui de *Google* lorsque l'utilisateur opte pour l'option *Direct search*. Cette dernière option étant très proche d'une interrogation directe de *Google* via sa propre interface, seule l'option *Dynamic search* a été évaluée.

2.2. Méta-moteur

- *Copernic* (<http://www.copernic.com>) est un métamoteur « client » au sens où son utilisation nécessite l'installation préalable d'un logiciel sur l'ordinateur de l'utilisateur. Deux versions sont disponibles: une version payante qui offre des possibilités avancées de recherche par domaine ou type d'informations recherchés et est plus spécialement dédiée au public des professionnels de l'information, et une version gratuite permettant d'interroger 10 moteurs de recherche généralistes simultanément. Une grande majorité d'utilisateurs en biologie n'ayant pas accès à la version payante, seule la version gratuite a été testée dans cette étude.

2.3. Annuaires

- *Infomine* (<http://infomine.ucr.edu/Main.html>) est un annuaire « spécifique », développé par un réseau de bibliothèques universitaires californiennes. Il répertorie environ 20000 sites qui font autorité en sciences pour un public d'universitaires et de chercheurs.
- *Open Directory Project (ODP)* (<http://www.dmoz.org>) est un annuaire généraliste, dont la sélection et l'indexation des pages web sont réalisées par des éditeurs volontaires, spécialistes chacun dans leur domaine d'activité. Il s'agit d'un des annuaires généralistes dont la taille d'index est la plus importante.