Bibenligne.org

La Recherche documentaire sur l'Internet en sciences humaines et sociales

Véronique Ginouvès
Phonothèque MMSH

Faire une recherche sur Internet : une interaction entre ces trois principes

1 Formuler son sujet de recherche

2 Connaître l'offre documentaire sur Internet

3 S'adapter aux outils et à leurs usages

Une interaction entre trois principes :
1. Formuler son sujet de recherche

- Formuler/cerner le sujet ;
- Créer une liste des idées et des concepts associés ;
- Les traduire en mots-clés ;
- Trouver ses synonymes dans le champs sémantique ;
- Les traduire en différentes langues ;
- Connaître les opérateurs booléens pour exclure, ajouter ou retrancher.

Une interaction entre trois principes :
2. Connaître l'offre documentaire sur Internet

- Repérer les grandes sources d'information sur le domaine de recherche, repérer les auteurs ;
- Dresser une typologie des types de documents ;
- Déterminer à quelle « échelle » se situe son questionnement : faut-il affiner ou élargir le(les) champ(s) disciplinaires ?
- Quel portail sur le sujet ? Quelle base de données bibliographique ?
- Où trouver le texte intégral ?

Une interaction entre trois principes :
3. S'adapter aux outils et à leurs usages

- Différencier les moteurs, des banques de données ou des répertoires ;
- Connaître la syntaxe des équations de recherche : elle est différente selon les moteurs mais aussi les catalogues de bibliothèque.
- Comprendre l'architecture d'une base de données pour une exploitation optimum (ex : index, thésaurus, RAMEAU...)
- Connaître les outils de veille (alertes, syndication de contenu)

Des repères pour s'adapter

Distinguer au moment de la recherche

- les univers **fermés** (les catalogues de bibliothèque, les sites spécialisés…)
- les univers **infinis** (web, courriel, forums…)

Et surtout savoir utiliser les sources documentaires classiques car tout ne se trouve pas sur la toile…

Les moteurs

Un moteur visite périodiquement une partie des fichiers statiques accessibles sur Internet et met à jour un index comprenant tout ou partie des mots des fichiers visités. Le résultat est une liste de pages web. Un moteur donne une image du Web déjà obsolète car il peut y avoir parfois des décalages importants entre ce que le moteur a « aspiré » et la forme actuelle du site.

La pertinence des résultats

- Le nombre de liens qui pointent vers la page
- La qualité du code en particulier la balise html <title>Titre de la page</title>
- Les mots clés contenu dans l'URL ex : www.mediterranee.france3.fr
- La fréquence de mise à jour du site
- Le nombre de requêtes faites à travers le moteur

Le mariage de tous ces critères pour le calcul de pertinence est un secret industriel…

Les moteurs spécialisés

Scirus (Elsevier) depuis 2001 : http://www.scirus.com

Google Scholar depuis novembre 2004 : http://scholar.google.com

Base - Bielefeld Academic Search Engine depuis 2004 : http://www.base-search.net/

In-Extenso créé en 2002, arrêté en 2009

OAIster créé 2002, devenu le moteur d'OCLC en octobre 2009

Google Scholar

Moteur de recherche spécialisé dans la littérature universitaire lancé fin 2004, encore en version *beta* en novembre 2009.

Base de données pluridisciplinaire

Adresse : http://scholar.google.com/

Qu'interroger sur Google Scholar

La couverture de Google Scholar ne peut pas être définie avec précision. L'objectif est de retrouver les documents du web invisible du monde scientifique. La base de données est multidisciplinaire avec une prépondérance des ressources en sciences exactes et médicales.

Dans Google Scholar vous y trouvez …

- Des fichiers en texte intégral à accès libre et payants en format HTML et PDF : éditeurs scientifiques, sociétés savantes, répertoires de pre-prints, serveurs universitaires...
- Des articles avec évaluation par les pairs (peer reviewed), des thèses, des livres, des articles en pré-publication, des rapports…

Google Scholar ne fournit aucune liste des éditeurs commerciaux ni de liste des serveurs d'archives qu'il indexe.

Vous ne savez rien sur…

- La période couverte par ces documents
- Le volume des documents présents sur Google Scholar
- Les langues représentées même si l'anglais est prédominant.
- La mise à jour des informations.

Recherche simple

La recherche se fait dans le **texte intégral en langage naturel**.

Les synonymes sont essentiels

Pour aller plus loin :

- la recherche avancée
- les préfixes

La recherche avancée

- Les champs se remplissent les fonctions des booléens AND, NOT, OR et de la recherche stricte avec guillemets
- Certaines champs proposent de rechercher par auteur, nom de publication et intervalle de dates de la recherche
- Vous pouvez rechercher le terme dans tout le document ou seulement dans le titre
- Vous pouvez limiter la recherche à un ou plusieurs domaine(s) parmi 7

Scirus

- Scirus est un moteur de recherche scientifique développé par l'éditeur scientifique Elsevier
- Son « Advisory board » est composé de chercheurs et de professionnels de l'informations
- Scirus annonce l'indexation de plus de 350 millions de pages issues du web scientifiques (novembre 2009)
- Adresse : http://www.scirus.com

Scirus vous explique son fonctionnement

Scirus indexe le texte intégral des pages Web qu'il visite.

- Les ressources obtenues suites à vos interrogations sont toutes accessibles depuis le Web, mais un certain nombre seront d'accès payant (Lexis Nexis, Science Direct d'Elsevier...)
- Vous interrogez dans Scirus aussi bien le texte intégral d'articles de recherches que le texte de sites d'informations institutionnels (plaquettes de laboratoire en ligne, sites d'universités)

Scirus vous explique son fonctionnement, mais :

- La politique d'indexation et sa fréquence sont clairement annoncées
- La nature commerciale du moteur tend à mettre en avant les ressources issues des portails payants
- Malgré la volonté annoncée de créer un moteur scientifique vous y trouvez un certain nombre de pages non scientifiques (pages de sites personnels ou commerciaux).

La recherche sur Scirus

- Un formulaire de recherche simplifié vous permet de faire vos recherches sur les mots langage naturel et par expression exacte, de sélectionner vos sources (Journal, Ressources privilégiées, Autres ressources).
- Un formulaire de recherche avancée vous permet d'utiliser les opérateurs booléens ainsi que de sélectionner les résultats selon leur date de publication, type de données, types de formats, origines des ressources, grands domaines de la science

Des “recettes” pour les moteurs

- Soyez naturel: plutôt que de chercher une liste de synonymes, tapez ce que vous savez sur le sujet qui vous intéresse, les sites sont généralement écrits dans un langage simple.
- Privilégiez les mots les plus précis possibles: une recherche sur "opinel" rapporte 3 fois moins de résultats qu'une recherche sur "couteau".
- Utilisez les expressions complètes entre guillemets ou séparées par des tirets
- Les mots importants d'abord : Google fournit des résultats plus précis si les mots-clés importants figurent en début de liste.
- Excluez les mots inutiles : le signe "-" Ex. : Corse -hotel
- Contournez les pages inaccessibles: tronquez les adresses des pages inaccessibles pour remonter à la racine. Exemple: si la page http://www.fao.org/sd/EGdirect/EGre0033.htm ne fonctionne pas, essayez http://www.fao.org/

Des outils spécialisés : blogs et annuaires

Bibenligne
http:/www.bibenligne.org

Internet per gli umanisti
http://biblio.lett.unitn.it/

La boîte à outil des historiens
http://laboiteaoutils.blogspot.com

Les bases de données - 1

Chaque base donne des informations en fonction de son principe de fonctionnement

Le catalogue SUDOC fournit :

- Des notices bibliographiques (9 millions)
- La localisation des ouvrages
- Les PEB

Mais uniquement des BU ou des bibliothèques adhérentes au réseau

Les bases de données - 2

Le catalogue Worldcat, est l'une des bases de données de l'OCLC (Online Computer Library Center) :

vous accédez aux catalogues de plus de 10 000 bibliothèques à travers le monde

proposant 1,4 milliards de notices de documents

en 31 langues.

Les bases de données - 2

Les points forts de Worldcat :

- La création de bibliographies, publiques ou non, exportables vers des logiciels (EndNote, Refworks, Zotero) ;
- L'intégration du portail OAISTER en octobre 2009 pour accéder à des documents numériques ;
- L'accès à un index auteur original : Worldcat indentities http://orlabs.oclc.org/identities

Les bases de données - 3

Deux bases à accès réservé :

Université de Provence : ENT

environnement numérique de travail

http://entu1.phocean.fr/uPortal/render.userLayoutRootNode.uP

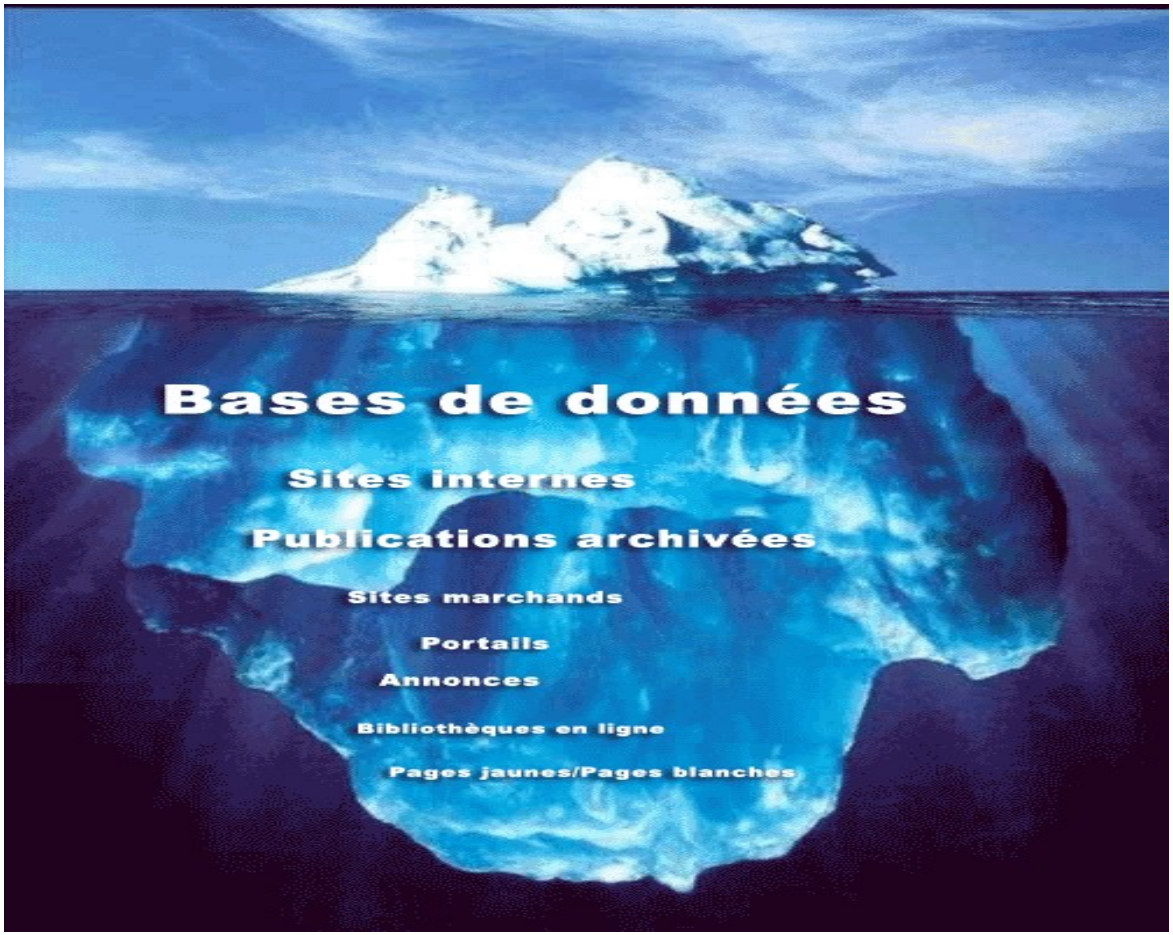
Biblio SHS (plus de 7000 revues)

http://biblioshs.inist.fr/

Accès aux articles en SHS

- Revues.org : http://www.revues.org
- Persée : http://www.persee.fr
- Cairn : http://www.cairn.info
- Erudit : http://www.erudit.org/revue
- HAL SHS : http://halshs.ccsd.cnrs.fr/

Le web invisible



Les 4 types de web invisible

- The ***Opaque Web*** : les pages qui pourraient être indexées par les moteurs mais qui ne le sont pas (limitation d'indexation du nombre de pages d'un site, fréquence d'indexation, liens absents vers des pages…)
- The ***Private Web*** : les pages web volontairement exclues par les webmasters.
- The ***Proprietary web*** : pages seulement accessibles pour les personnes qui s'identifient. Le robot ne peut pas toujours y accéder.
- The ***Truly Invisible Web*** : contenu qui ne peut être indexé pour des raisons techniques, comme un format inconnu par le moteur (même si Google reconnaît de plus en plus de formats) et surtout de nombreuses des bases de données.

Source : Chris Sherman et Gary Price, The Invisible Web : Finding Hidden Internet Resources Search Engines Can't See

Un exemple de Web invisible

La phrase « Eindhoven et Bois-le-Duc envoyaient leurs toiles au blanchissage à Harlem » prise dans l'article Z.-W. Sneller. La naissance de l'industrie rurale dans les Pays-Bas aux XVIIe et XVIIIe siècles, *Annales*, 1929, n° 2, pp. 193-202.

Un exemple de Web invisible

Se retrouve bien dans Google sur Jstor mais pas sur Persée où pourtant elle est librement accessible

http://www.persee.fr/web/revues/home/prescript/article/ahess_0003-441x_1929_num_1_2_1064

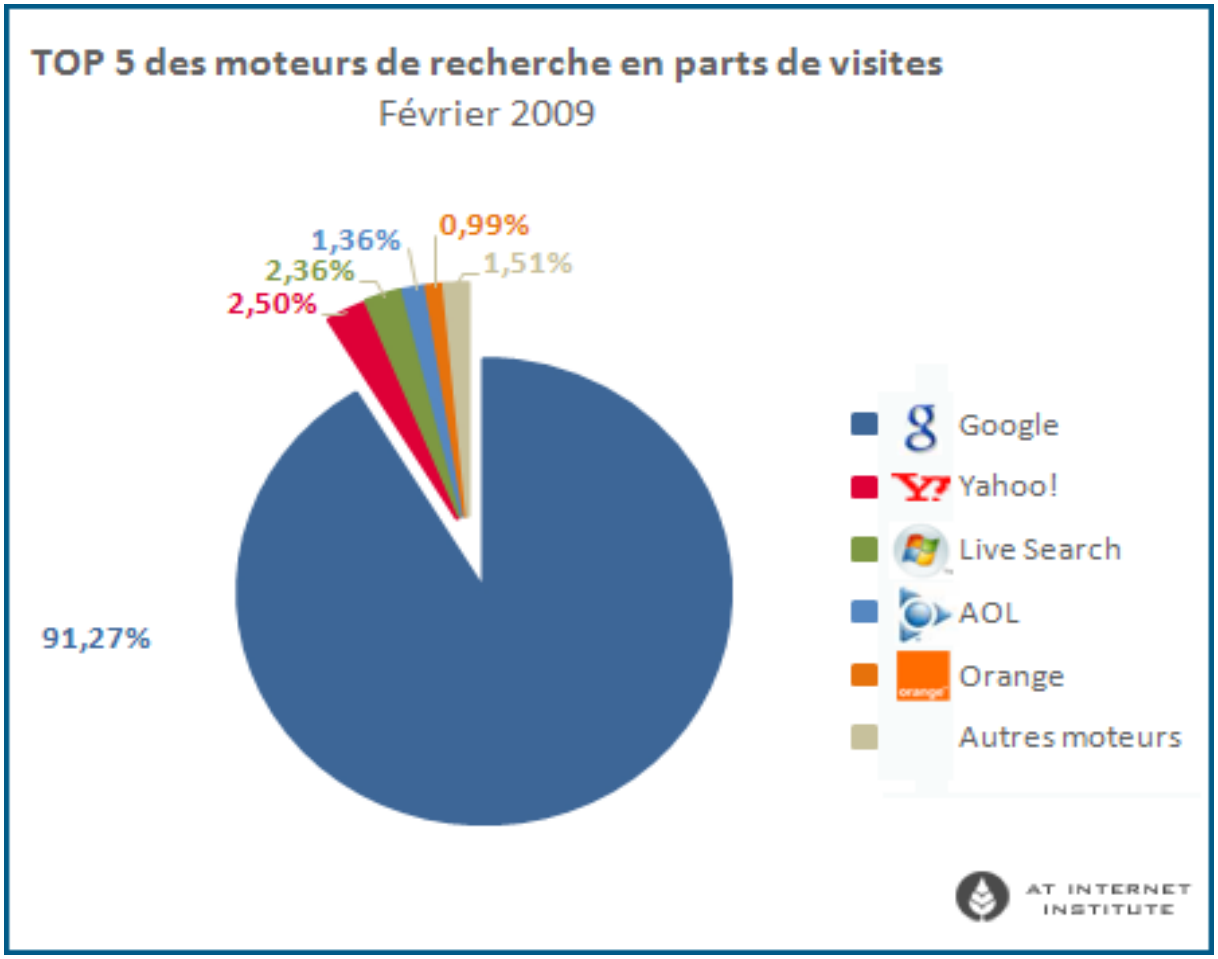
Prédominance de Google

- **Google représente plus de 90 % du trafic généré par les moteurs de recherche francophones**

AT Internet Institute, février 2009 ;

- **Plus de la moitié des recherches Google proviennent de pays autres que les États-Unis.**
  (Données internes à Google)

Prédominance de Google



Relativiser les modes de recherche

Sur Google 34% des internautes se contentent de requêtes en un mot,

30% en 2 mots,

18% en 3 mots

Seuls 4% des internautes utilisent les catégories d'annuaire.

Source : http://www.journaldunet.com/chiffres-cles.shtml (*Journal du net*, décembre 2005)

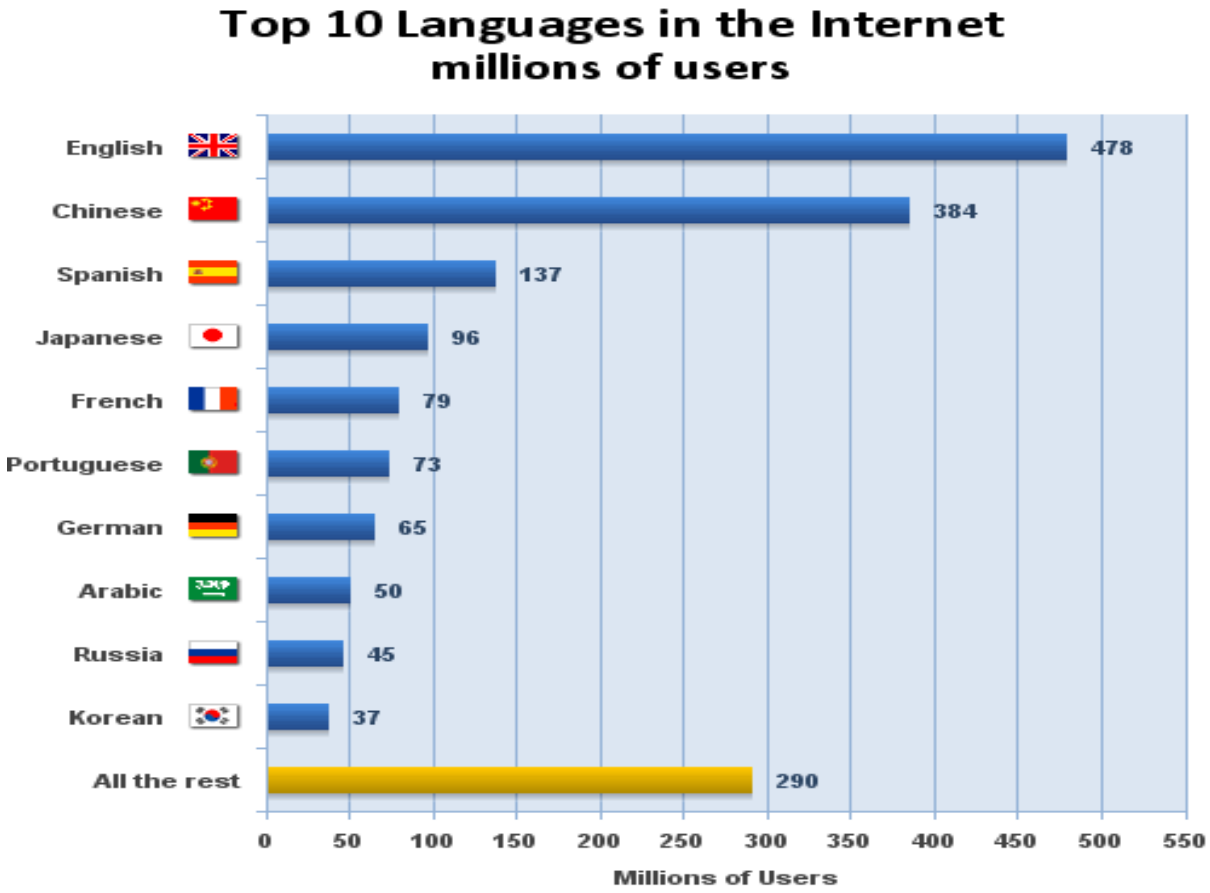
Relativiser les modes de recherche

42% des utilisateurs cliquent uniquement sur le premier résultat de la recherche.

8% cliquent ensuite sur le deuxième résultat.

Source : http://www.useit.com/alertbox/defaults.html

Relativiser l'emploi des langues sur l'internet



Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users are 1,733,993,741 for September 30, 2009

Validité - Citabilité

- Validité de l'information : évaluer la qualité d'une ressource sur Internet en appliquant une série de critères d'analyse
http://www.fl.ulaval.ca/icarish/guide/module

- Citabilité : citer les références électroniques
http://www.bibenligne.org/index69.html

Organiser sa veille

- En créant des alertes sur des mots clés :

http://www.google.fr/alerts

- En utilisant la syndication de contenu ou fils RSS, exemples :

https://bloglines.com/public/Bagolina

et

http://www.bibenligne.org/index70.html

Bonnes recherches...

N'hésitez pas à contacter les professionnels de la bibliothèque à la médiathèque de la MMSH

mediatheque@mmsh.univ-aix.fr

Mon courriel :

ginouves@mmsh.univ-aix.fr

Le diaporama :

http://www.bibenligne.org/index5822.html