

UNIVERSITÉ DE
BORDEAUX

Département Documentation-URFIST

RECHERCHE DOCUMENTAIRE ET RECHERCHE D'INFORMATION

Bien distinguer :

Recherche documentaire : rechercher des références de documents ou les documents eux-mêmes (ouvrages, thèses et mémoires, articles, etc)

ET

Recherche d'informations : rechercher de l'information via un moteur de recherche sur Internet (et non plus seulement des documents)

L'information est multiforme

- **L'information « orale » (médias audiovisuels notamment)**

- **L'information écrite :**
 - **l'édition commerciale donc payante ou littérature blanche (livres , revues, bases de données essentiellement)**

- **l'édition non commerciale ou littérature grise (thèses, mémoires, rapports, articles en prépublication, actes de congrès non publiés, etc)**

Définition « littérature grise » (AFNOR)

tout document dactylographié ou imprimé, produit à l'intention d'un public restreint, en dehors des circuits commerciaux de l'édition et de la diffusion et en marge des dispositifs de contrôle bibliographiques

Voir [Gobert, B. Les dessous de la publication scientifique](#)
IRD, Brest, 2008

L'IST - Information Scientifique et Technique

« ensemble des connaissances générées par les activités des groupes de recherche et qui, après validation par la communauté internationale, (**validations par les pairs**) est transformée et exploitée selon les besoins de différents publics »

Extrait du PLF 2006, rapport annexe sur l'état de la recherche et du développement technologique.

Ensemble des informations destinées aux secteurs de la recherche, de l'enseignement et de l'industrie.

L'IST se caractérise donc par :

sa **portée internationale**

sa **validation** au sein d'un collège de
spécialistes

la mise en œuvre de **techniques de
médiation et d'outils informatiques.**
Les deux derniers points induisent
des coûts....

Essentielle à la production des savoirs, l'IST est un **enjeu** majeur dans la concurrence **économique et scientifique**.

Par conséquent, c'est un secteur d'activité économique (on vend de l'information et souvent cher).

Donc toute l'IST n'est pas en accès gratuit sur le web. Si vous ne regardez que là, vous ne voyez qu'une partie de l'information existante.

- **Marché de l'IST (d'après Remy Bilbault, Wolters KluwerFrance, Les dernières tendances des modèles économiques de l'information professionnelle, 2009)**
- **Marché mondial = 14 milliards d'€ de CA**
Dont Marché européen de l'IST = 2 milliards d'euros (dont 1,3 milliard d'euros pour l'IST en ligne)
- **2000 éditeurs de revues scientifiques dans le monde**
- **25.000 titres de revues dont 8 000 sont considérées comme importantes et**
- **4000 ont un fort facteur d'impact**

- **Facteur d'impact** = indicateur statistique destiné à l'évaluation de la notoriété d'un périodique, publié par l'ISI (Information Science Institute) dans le Journal of Citation Report (JCR) = rapport pour une année donnée, entre le nombre de citations des articles publiés par un périodique et le nombre d'articles publiés, le tout sur une période de deux ans.

- **1,5 M d'articles publiés par an**

- **780 éditeurs dans l'UE, 49% de la production d'articles**
- **90% des revues scientifiques sont accessibles en ligne**
- **Editeurs= 66% , Sociétés savantes= 30%
Universités= 4%**
- **Accès aux articles par abonnement à 90%**

Voir le marche mondial (papier et électronique) par le GFII

- **Et le rapport « Salençon » sur l'IST**

- **Classement mondial des groupes d'édition d'information scientifique et technique (et professionnelle), les « big three » ou « four » (données 2011, en chiffre d'affaires):**
- **1-Thomson Reuters : CA de 10,4 Md. € (dont 90% sur produits en ligne)**
- **2- Reed Elsevier : 6,9 Md. € (75%)**
- **3- Wolters Kluwer : 3,35 Md. € (70%).**
Ce groupe inclut Springer. Rentabilité 2010=21% sur CA

- **Ces trois grands groupes se partagent près de 50% (en valeur) du marché mondial**

25% des revues

et surtout 50% de celles à facteur d'impact

- **Augmentation du prix des périodiques : de 1975 à 1995 = +300%**

Depuis 2000, augmentation de 6 à 8% par an

- **Dépenses des bibliothèques pour la documentation électronique:**
56 millions d'euros en 2008 (33 pour les universités, 28 pour les organismes de recherche)
- + de 70% des abonnements souscrits via le consortium d'achat Couperin**
Ceci permet de mieux négocier les prix avec les éditeurs

- **Licences nationales = achats de corpus clos = archives de revues, dictionnaires, corpus de textes, etc**
- **par l'ABES (Agence Bibliographique de l'Enseignement Supérieur) pour un accès pour toutes les BU et à terme pour la BNF et les Bibliothèques Municipales.**

Les composantes de l'IST

Publications

Revue, thèses, actes
de congrès, ouvrages,
brevets

Données factuelles

Données d'expérience, statistiques
enregistrements vidéos
ou sonores...

Référencement

- bases de données
bibliographiques, catalographiques
et factuelles
- moteurs de recherche
généralistes & spécialisés

Outils d'analyse

Bibliométrie-infométrie, outils
d'évaluation (Journal Citation Reports,
F1000, PLoS One...)

L'IST est essentiellement produite par les pays développés (l'UE et les USA)

marché mondial de l'information scientifique, technique et médicale (papier et électronique) = près de 20 milliards de dollars (source Outsell Inc., 2007)

Une des activités les plus rentables au sein de l'édition professionnelle : croissance annuelle d'environ 8%.

Outils de référencement et de recherche de l'IST (bases de données) :

- **Situation de quasi-monopole des États-Unis sur la production de BdD**
- **Peu de poids du marché européen (en valeur, le 10ème du marché américain environ)**

- L'ISI (Institute for Scientific Information) produit les outils bibliométriques dominants : Journal of Citation Report
- L'information est majoritairement produite en **langue anglaise** quelle que soit la nationalité des chercheurs qui la publient
- Nouvelles modalités de l'Information Scientifique M.L Malingre, J. Sicot, *URFIST Rennes, 2010*

Les vecteurs de l'information en constante évolution :

Le **papier** : encore irremplaçable

Les **supports informatiques** : après les disquettes et les cassettes VHS, les **disques durs**, **cd-roms** (apparus au début des années 80, stockage 750 Méga octets), les **DVD** (stockage jusqu'à 18 Giga octets), les **clés USB**, **l'encre ou papier électronique**, ...

Des appareils de lecture : les liseuses ou tablettes de lecture , les IPAD, etc...

Et Internet bien sûr depuis 1989 (le web)

DOCUMENT PRIMAIRE

livre, thèse, article de revue,
etc

Contient :

TEXTE ORIGINAL

RESUME (éventuellement)

Et des informations:

Titre

Auteur

Lieu de publication

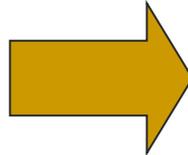
Editeur

Adresse

Date

ISBN ou ISSN

ETC



le document
est

ANALYSE
et
INDEXE

par des
spécialistes

REFERENCE BIBLIOGRAPHIQUE

Auteur

Titre

Lieu de publication

Editeur

Date

MOTS-CLES

Et éventuellement
un résumé, la page de titre,
La table des matières, etc

Les références bibliographiques permettent de produire deux grands types d'outils bibliographiques dont la finalité est différente :

- **Les bases de données bibliographiques :** signalement de **l'information produite**, quelque soit la langue, le pays, en général **dans un domaine scientifique** (Historical Abstracts, Mathscinet, Business Source Premier, Lexbase, etc) , **parfois pluridisciplinaires** (Francis, ASP)

- **Les catalogues** : signalement de l'information primaire possédée par un ou des organisme(s) documentaire(s)
- **BABORD PLUS**: catalogue des bibliothèques des universités de Bordeaux et plus qu'un catalogue, une « interface de recherche documentaire »
- **Sudoc**: catalogue national des BU
- **CCFR** Catalogue des BU, BNF et certaines BM

- **Typologie des outils de recherche :**
- **Les catalogues** de bibliothèque
- **Les bases de données** bibliographiques spécialisées (gratuites ou payantes)
- **Les outils de recherche sur Internet** (ou plutôt sur le web) :
des annuaires, des listes de signets ou répertoires de sites spécialisés ou des portails, des méta moteurs, des moteurs, des moteurs spécialisés, des agents intelligents.

Principes généraux de la recherche documentaire, méthodologie

Bien définir quel est l'**objectif** de la
recherche : un exposé, une note de
synthèse de une à deux pages, un
mémoire de 20 pages, une thèse...

Donc **le temps** que l'on peut et doit y
consacrer

Le niveau de la recherche bibliographique, son volume seront tout à fait différents

Les outils documentaires que l'on décidera d'utiliser également.

Selon le type de document que l'on cherche (ouvrage ou article), l'outil documentaire ne sera pas forcément le même.

**Tout l'art de la recherche
documentaire consiste à savoir **se
poser des questions pour cerner
son sujet****

**On ne dit jamais ce qu'on veut vraiment
du 1er coup**

Donc reformuler

**Bien lire et relire le sujet comme dans la
dissertation**

Que sais-je déjà du sujet ? jeter tous les mots sur un papier en vrac, tout ce qui vient en tête sur le thème

Utiliser le système 3QOCP (ou système de Quintilien ou modèle de Laswel)

qui, quoi, quand, où, comment, pourquoi
(5W en anglais)

Définir des **mots-clefs** = mots indispensables pour le sujet, 2 à 4

A supprimer :

- **Les mots vides : articles, prépositions, pronoms, etc**

- **Les mots inutiles :**

Problème, problématique, incidence de, influence de, comparaison, relation, importance, etc

▪ Les mots qui seront induits de toute façon par la combinaison des deux ou trois mots qu'on va garder dans l'énoncé

Ex: seuls les mots en vert sont à garder
stress professionnel des **infirmières**
difficultés d'insertion de l'enfant **handicapé**
à l'**école**
influence de la montée en puissance (on mettra par ex. **croissance**) de la **Chine** sur
le **commerce international**

Chercher des synonymes aux différents termes que l'on garde et leur traduction.

Ne pas oublier l'utilité des ouvrages de référence traditionnels (papier ou en ligne): dictionnaires, encyclopédies, etc pour définir les termes retenus, trouver des mots clés, des auteurs et traduire en anglais (beaucoup de bases de données sont américaines et s'interrogent en anglais)

Certaines de ces ressources sont accessibles en ligne de partout via l'ENT de votre université (abonnements pour toutes les universités de Bordeaux)

- **Encyclopédia Universalis**
- **Dictionnaire Robert (langue française)**
- **Dictionnaire Robert et Collins anglais français/français-anglais**

On peut aussi trouver des dictionnaires gratuits répertoriés dans ces sites : le grand dictionnaire terminologique, le Trésor de la Langue Française et bien d'autres

<http://www.dicorama.com/>

www.yourdictionary.com

Des encyclopédies gratuites (tout ou partie)

La plus connue, l'encyclopédie
Wikipedia

Wikipedia est un projet collaboratif gratuit, passionnant **mais en constante évolution, tout le monde peut modifier un article (à condition d'être utilisateur enregistré), vous pouvez d'ailleurs voir les modifications (dans l'onglet « historique » d'un article).**

**Attention à la validité des informations
pour un niveau universitaire , voir
les avertissements dans le site
Et les débats sur l'évaluation de
l'information (débats déjà anciens)**

**Donc préférez les « articles de
qualité » (articles finis, validés),
ceux-là peuvent être cités en
confiance. Autre catégorie : les bons
articles**

- Organisation de Wikipedia, statuts (de certains membres choisis par la communauté) :
- Administrateur, Bureaucrate, Stewart, Vérificateur d'adresse, Patrouilleur, Masqueur de modification, etc
- **Sylvain Machefert. Wikis & wikipédia du point de vue du bibliothécaire, janvier 2012**

Pour chercher de la documentation ou de l'information, on utilise des outils informatisés :

Spécifiques à la recherche documentaire

- **Catalogues**
- **Bases de Données (pluridisciplinaires ou disciplinaires)**

Destinés à la recherche d'information en général

- **Moteurs de recherche sur le web (généralistes ou spécialisés)**

- **Les logiciels de recherche de ces outils ne comprennent pas le langage naturel humain, n'établissent aucune relation de sens, ne traitent que des chaînes de caractères (ce sont des « machines »).**
- **Par conséquent, ils ne pardonnent aucune inexactitude (erreur de frappe, d'orthographe, etc). Cependant, ils font souvent des suggestions (par autocomplétion).**

La recherche avec des outils documentaires (catalogues, bases de données)

Principes de base :

Pas de casse, c'est-à-dire de différence entre les caractères minuscules et majuscules, mettre tout en minuscules

Pas d'accent

Supprimer les mots vides (articles, d', prépositions, etc)

Utiliser les **troncatures** (illimitée et limitée) :

Troncature illimitée : dans tous les systèmes, c'est l'astérisque ou étoile *

Très utile au moins pour prendre le singulier et le pluriel d'un mot mais aussi le radical d'un terme :

enfant* permet de retrouver enfant, enfants, enfance, enfantin(s), enfantine(s)

Energi* -> energie(s), energetique(s), energy, etc

- **Attention à ne pas tronquer trop court et générer du bruit (trop de réponses et/ou pas pertinentes) par ex: Bio* , psy***
- **Troncature limitée (ou masque ou joker) : ? (point d'interrogation, parfois un autre signe, le dièse #)**
- **remplace *une* ou *zéro* lettres (wom?n)**

Les **opérateurs booléens**

George Boole, Traité d'algèbre, 19ème siècle

Les opérateurs booléens permettent de croiser entre eux les termes choisis et de créer des intersections, des réunions, des exclusions d'ensembles (opérations algébriques, logiques)

- **ET (AND)** : réduit la question, permet de rajouter un deuxième élément de précision, de combiner deux mots, deux ensembles (ou plus) = **intersection** d'ensembles
- **Exemples :**
handicap* et ecole*
adolescen* et anorex*
violence et banlieu*
lesion AND pancreatic

OU (OR) : élargit la question, juxtapose plusieurs termes , des synonymes ou des mots mis sur le même plan par exemple différents groupes sur lesquels porte l'étude, etc = réunion d'ensembles
Ex : seisme ou tremblement de terre
ecole ou enseignement
infirmier* ou medecin* ou soignant*
aids ou sida
epilepsi* ou convulsi*
kidney OR renal

SAUF (NOT ou AND NOT) : enlève un terme, un ensemble de réponses = **exclusion d'ensembles**

Ex : chimie sauf organique

depressi* sauf adolescen*

malade sauf imaginaire

ganglia OR tumor AND NOT malignant

On peut, bien sûr, créer des équations

complexes avec des parenthèses

(comme en algèbre) :

(a ou b) et (c ou d) et e sauf f

- **Deux grands écueils à éviter en recherche documentaire**
- **le « bruit » : trop de réponses et non pertinentes par manque de précision ou mauvaise stratégie**
- **le « silence » : pas assez ou absence de réponses à cause d'une mauvaise stratégie (ou d'une faute d'orthographe ou faute de frappe...) ou de trop de précision, trop de mots clés**

- **Autres opérateurs**
- **Les opérateurs numériques**
- **Plus grand que : >, plus petit que : <**
- **Egal : =**

Les **opérateurs de proximité** :

ADJ : permet de retrouver les références où deux mots liés par l'opérateur ADJ se suivent immédiatement et dans l'ordre fixé

NEAR : les deux mots se trouvent dans la même phrase (indépendamment de l'ordre)

D'autres opérateurs permettent, selon les logiciels d'interrogation, de retrouver des termes dans la même phrase, le même paragraphe (PRH, PRG) ou à X mots de distance

- **Il faut, au fur et à mesure de sa recherche, conserver les références bibliographiques trouvées, la mémoire des outils utilisés, voire des stratégies de recherche**
- **Donc utiliser les fonctionnalités des outils de recherche: envoi par mail, sauvegarde des recherches , alertes, etc**
- **Et un outil de gestion des références bibliographiques : Endnote (payant), Zotero (gratuit) ou autre**

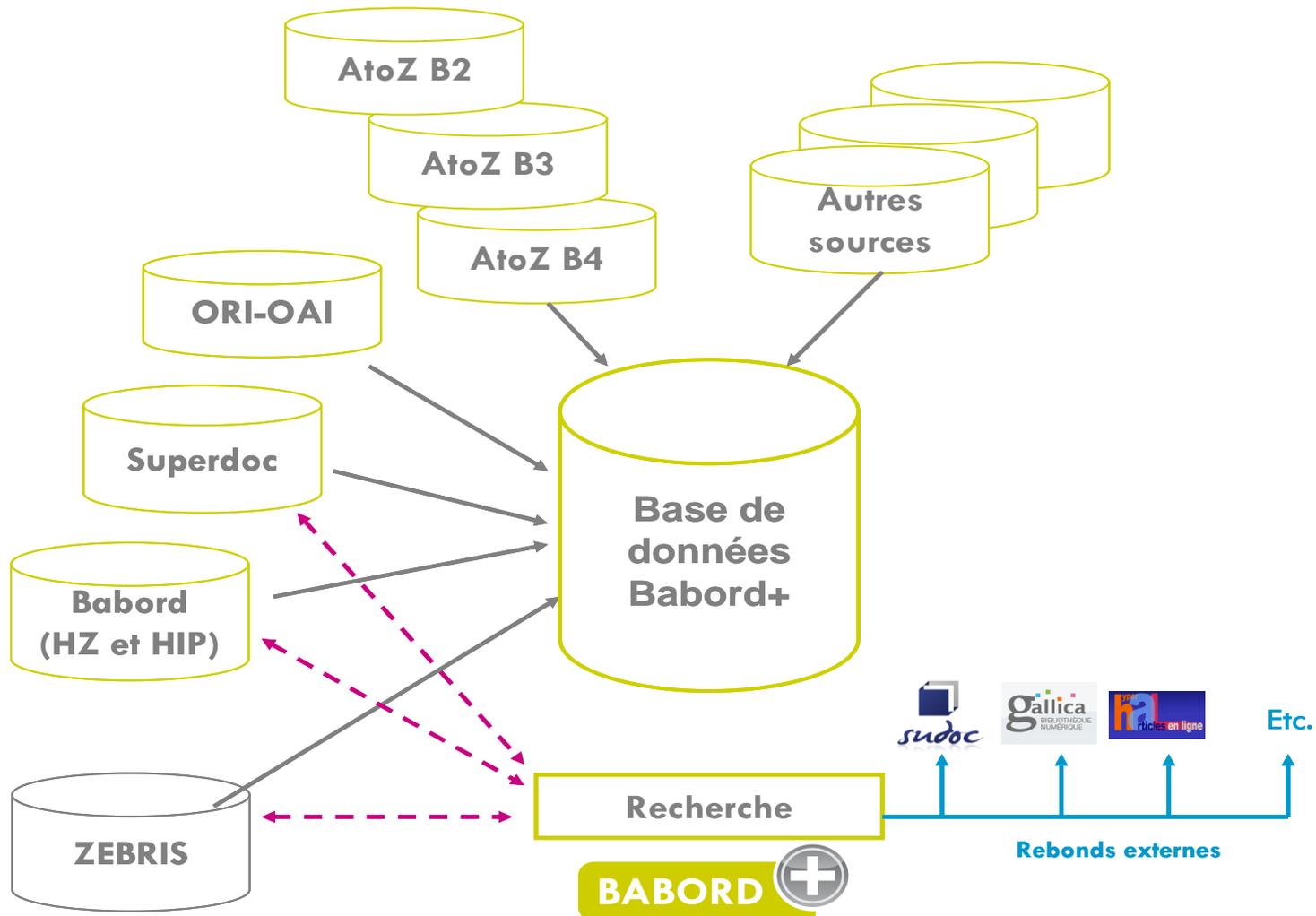
Les catalogues de bibliothèque:

On y trouve certains types de documents : des livres, des mémoires, des thèses, des comptes-rendus de congrès, des numéros spéciaux de revues, les titres des revues, **peu d'articles** (ce sont les bases de données qui les recensent), des rebonds (sans reposer la question) vers d'autres sources (autres catalogues, bases de données, etc)

Babord + : nouveau catalogue local des bibliothèques des universités bordelaises (septembre 2011), plus large qu'un simple catalogue, rebond vers d'autres ressources (BdD ou archives ouvertes ou autres catalogues, selon les choix des SCD)

Bon tutoriel pour chercher une revue dans Babord+ (il s'agit bien de trouver un titre de revue et non des articles)

- BabordPlus schéma de fonctionnement (D. Coudrin, DDOC)



- **Sudoc: catalogue national des BU et bibliothèques de l'enseignement sup.**
- **CCFR : catalogue national englobant la BNF, les B.U, certaines BM**

- Pour indexer des documents dans les catalogues = dire de quoi ils parlent donc mettre des mots-clefs (des vedettes-matière en jargon de bibliothécaire!), on utilise une liste d'autorité
- La plus utilisée dans les bibliothèques françaises RAMEAU
- Les autorités RAMEAU sont accessibles en ligne à partir du [catalogue général de la BnF](#) (bouton "RAMEAU" en bas à gauche de l'écran).

Les catalogues sont des outils documentaires comme les autres, on y trouve les mêmes fonctionnalités que dans les Bases de Données :

Une recherche simple

Une recherche avancée

Des limitations

Un historique de recherche

Des possibilités de sauvegarde, envoi par mail, alerte, etc...

**Un cas un peu particulier de document :
les thèses**

Pour avoir les références :

**thèses locales : dans le catalogue de
votre bibliothèque ou du réseau local**

**thèses partout en France : dans le Sudoc
ou dans [Theses.fr](http://theses.fr) (utile aussi pour voir
les thèses en préparation)**

Pour avoir des thèses en texte intégral en ligne :

Theses.fr : environ 6000 thèses (cliquer sur Uniquement les thèses soutenues accessibles en ligne)

Tel (CCSD du CNRS) environ 26.000 thèses

- Cyberthèses (thèses de Lyon 2, du Chili, d'Amérique latine)
- Cither (thèses de l'Insa)
- Pastel (thèses des grandes écoles de Paris)
- Au niveau européen DART

Les bases de données:

**Pour chercher des références d'articles
essentiellement**

**Comment connaître les bases de données
à utiliser : le plus simple est de regarder
ce que votre SCD vous propose**

**Il sélectionne et s'abonne pour vous à
certaines bases de données – vous
pouvez les consulter soit sur place, en
intranet ou de partout via l'ENT = accès
sur authentification**

La plupart des bases de données utiles pour un travail universitaire sont payantes (abonnement par le SCD donc « gratuites » pour l'utilisateur)

Pour connaître les BdD (au sens large)

gratuites : le répertoire DADI (Urfist de Lyon mais n'est plus actualisé)

Utilisez d'abord les ressources de votre SCD mais pensez à celles des autres SCD du site qui peuvent aussi vous être utiles (accessibles en vous déplaçant dans les différentes bibliothèques)

SCD Bordeaux 1 : Pascal, Inspec, Georef, SciFinder Scholar, MathSciNet, Scopus, etc

SCD Bordeaux Segalen (ex Bx2) : Francis, Psycinfo, Psychology and Behavioral Sciences Collection, Socindex, Scopus, etc

SCD Bordeaux 3 : Francis, MLA, Historical Abstracts, America History and Life, Index Islamicus, etc

SCD Bordeaux 4 : Business Source Premier, Econlit, Regional Business News, Delphes, Vente et gestion , Source OCDE, Dalloz, Jurisclasseur, Lextenso, Cairn, Science Direct, etc

toujours se demander quel est le contenu de la base de données que l'on utilise :
champ(s) disciplinaire(s) couvert(s), texte intégral des articles ou références bibliographiques seulement (ex : Francis), période chronologique couverte, etc.

Le SCD en fait une description ou la base elle-même donne ces informations.

- **élaborer sa stratégie de recherche :**
chercher des mots clefs (2 à 4)
- **Utiliser les possibilités d'affiner par les facettes (de + en + souvent proposé)**
- **appliquer les principes d'interrogation :**
troncature, opérateurs booléens (en recherche avancée), limitations, etc
- **Pour connaître l'indexation utilisée**
dans la base, consulter quand il existe le **thesaurus = liste des mots clefs**

Recherche d'information sur Internet (Interconnected Network ou International Network) ou plutôt sur le web :

**Le web n'a pas été conçu comme un outil
documentaire**

**Lorsqu'on a une recherche documentaire
à faire, la première chose à faire, c'est de
ne pas aller sur le WEB !! utiliser d'abord
les outils spécialisés (catalogues, BdD)**

Que trouve-t-on sur Internet (sur le web) ?

Tout et n'importe quoi ou plus exactement, de tout, pas tout (web invisible où se trouve notamment l'information scientifique payante) et parfois n'importe quoi...

- **Ce qui caractérise les informations disponibles sur le WEB :**
 - la **diversité de nature des informations** (techniques, commerciales, de divertissement, politiques, etc...)
 - la **diversité de forme** (textes, vidéos)
 - la **quasi absence de structuration des informations**
 - la **difficulté de connaître l'auteur et la date des informations**

Internet remet en cause les schémas économiques du marché de l'information

- favorise et **accélère la diffusion** des informations

MAIS pose le **problème de:**

- la **validation** et de la **pérennité** des informations

- Et l'aspect politique d'Internet ? Un article indispensable de M. Dacos résumant le livre de Dominique Cardon. **La démocratie Internet. Promesses et limites. Seuil, 2010**

Internet, cinquième puissance économique mondiale en 2016 (derrière seulement les Etats-Unis, la Chine, le Japon et l'Inde, et devant l'Allemagne)

Etude du Boston Consulting Group : poids de l'économie Internet en 2016 = 4.200 milliards (contre 2.300 milliards en 2010)

Essor de 10% par an, porté par deux tendances de fond : l'accès à Internet sur les appareils portables et l'Internet «social», où la navigation est largement guidée par affinités.

**Domination des USA dans tous les secteurs :
routeurs de réseau = appareils pour
interconnecter les réseaux informatiques
(Cisco), moteurs (Google) vidéos en ligne
(Youtube appartenant à Google), réseaux
sociaux (Facebook), logiciels (Apple),
systèmes d'exploitation (Microsoft), semi-
conducteurs (Intel)
En 2005, 19 entreprises sur les 25 qui
dominaient le marché du logiciel et d'Internet
étaient américaines. Les USA ont plus de la
moitié des satellites**

Internet remet aussi en cause le statut classique de l'auteur, l'autorité ou autoritativité

Tout le monde peut publier sur le web, créer du contenu et ceci s'est accentué avec les réseaux sociaux, le web 2.0.

Les types d'outils de recherche sur le web:

- **annuaires ou listes de signets ou répertoires de sites spécialisés ou portails ou guides**
- **métamoteurs**
- **moteurs généralistes**
- **moteurs spécialisés**
- **« agents intelligents »**

Bien distinguer les annuaires des moteurs :

Les annuaires ont précédé les moteurs (1992, Yahoo était au départ un annuaire) mais ils ont perdu beaucoup d'importance ou parfois disparu.

Un annuaire classe les sites dans des rubriques, des catégories, des thèmes : sciences, culture puis plus finement en sous-catégories.

- **Pertinence des sites sélectionnés (surtout s'il est fabriqué par des humains...)**
- **Utiles surtout pour des recherches assez vastes où tout un site peut nous intéresser**

On peut classer dans la même famille que les annuaires, les répertoires de sites ou les signets, les portails spécialisés, les guides (terminologie diverse et un peu flottante).

Ces outils sont fabriqués par des spécialistes du domaine, de la discipline, du thème qui sélectionnent les sites qui leur paraissent pertinents, de qualité, utiles.

Annuaire grand public:

-de Yahoo

-Celui de Google a été supprimé en 2011, ils renvoient sur celui qu'ils avaient racheté auparavant...open directory

Open directory meilleur surtout pour un niveau universitaire

Préférez les annuaires thématiques et/ou spécialisés (fabriqués souvent des professionnels de la documentation donc sélection des sites plus riche, plus pertinente, commentaires sur les contenus, l'intérêt du site)

Répertoire d'annuaires (en français)
<http://www.lesannuaires.com/>

Bons répertoires mais en anglais :

InfoMine

Intute

The WWW Virtual Library

IPL2

Annuaire spécialisés, disciplinaires de
niveau universitaire :

Sciseek

Cyberstacks

En chimie: Chemdex

En lettres : Wessweb

et bien d'autres

Des listes de signets choisis par les bibliothèques (en français):

Signets de la BNF

Signets des universités (par l'Abes et le Cerimes) : sites bien décrits, choisis avec soin, pas trop nombreux, avec flux RSS

- Signets de Paris 8
- Signets de l'IEP de Lyon
- Les signets sont maintenant souvent sous des univers ou portails Netvibes par ex Lyon 1, la BNF
- En Droit = Cujas ou Precisement.org

- **Des Portails :**
- **SciCentral : + de 50 000 sites sur plus de 120 spécialités scientifiques**
- **En maths, informatique, sci. ing.**
Techxtra
- **En Histoire :**
History on line

En droit :

- JuridicOnline
- Droit en ligne
- Droit.org

Des guides comme Jurisguide (droit),
Rime (économie, management)

- Une veille qui récapitule beaucoup de sites en droit et sciences de l'information: Silex

Les UNT (Universités Numériques Thématiques) dans 7 disciplines **Video présentation**

Lettres Sci Hum : UOH <http://www.uoh.fr/>

Droit : UNJF <http://www.unjf.fr/>

Sci Eco : Aunege <http://www.aunege.org/>

Sciences : Unit et Unisciel

**Environnement et Développement Durable:
UVED**

Sciences de la Santé et du Sport : UNF3S

- **Les télévisions universitaires :**
 - **Canal U, vidéothèque numérique de l'enseignement supérieur (conférences, cours)**
ou Canal2c pour des colloques et conférences
- Voir aussi Youtube Edu**
- Les cours du Collège de France
 - De l'école nationale supérieure (ENS) depuis 2011 et archives précédentes

- **Des outils particuliers pour les scientifiques, normes et brevets:**
- **Normes**
- **AFNOR**
[http://www.boutique.afnor.org/NE1
AccueilNormeEnLigne.aspx](http://www.boutique.afnor.org/NE1AccueilNormeEnLigne.aspx)
- **ISO**
http://www.iso.org/iso/fr/iso_catalogue.htm

- Les sources grand public et gratuites : les bases de données des offices de propriété industrielle

[d'après la Diapositive de Philippe Gorry (MCU-PH, Université Bordeaux2). Les bases de données brevets]

- Français : <http://fr.espacenet.com/>
- Inpi (Plutarque) marques et brevets
http://www.plutarque.com/plutarque_accueil/php/contenu.php?action=tester_brevets
- Européen : <http://ep.espacenet.com/>
- USA : www.uspto.gov/patft/index.htm *Plus de 6,800,000 brevets U.S. classés dans 462 catégories, et 154,409 sous-classes*

- ◉ Japon: <http://www.jpo.go.jp/index.htm>
- Corée:
<http://patent2.kipris.or.kr/pateng/searchLogina.do?next=GeneralSearch>
- ◉ Chine:
<http://ensearch.sipo.gov.cn/sipoensearch/search/tabSearch.do?method=init>
- ◉ Inde:
<http://ipindia.nic.in/ipr/patent/patents.htm>
- ◉ OMPI:
<http://www.wipo.int/portal/index.html.fr>
- ◉ Autre: <http://www.google.com/patents>

Meta moteurs :

Ils interrogent plusieurs moteurs ou annuaires à la fois et dédoublonnent les résultats.

Outils très inégaux qui donnent beaucoup de résultats

Ne peuvent faire de recherche complexe : ils n'envoient qu'une requête à tous les moteurs et ceux-ci ont des syntaxes différentes

Ixquick interface en français (mais on ne sait pas quels moteurs sont interrogés)

Dogpile affiche sur une même page les meilleurs résultats de Google, Yahoo, Bing en une liste unique et sans doublon

Metacrawler (idem, résultats de Google, Yahoo, Bing)

Mamma

Yippy

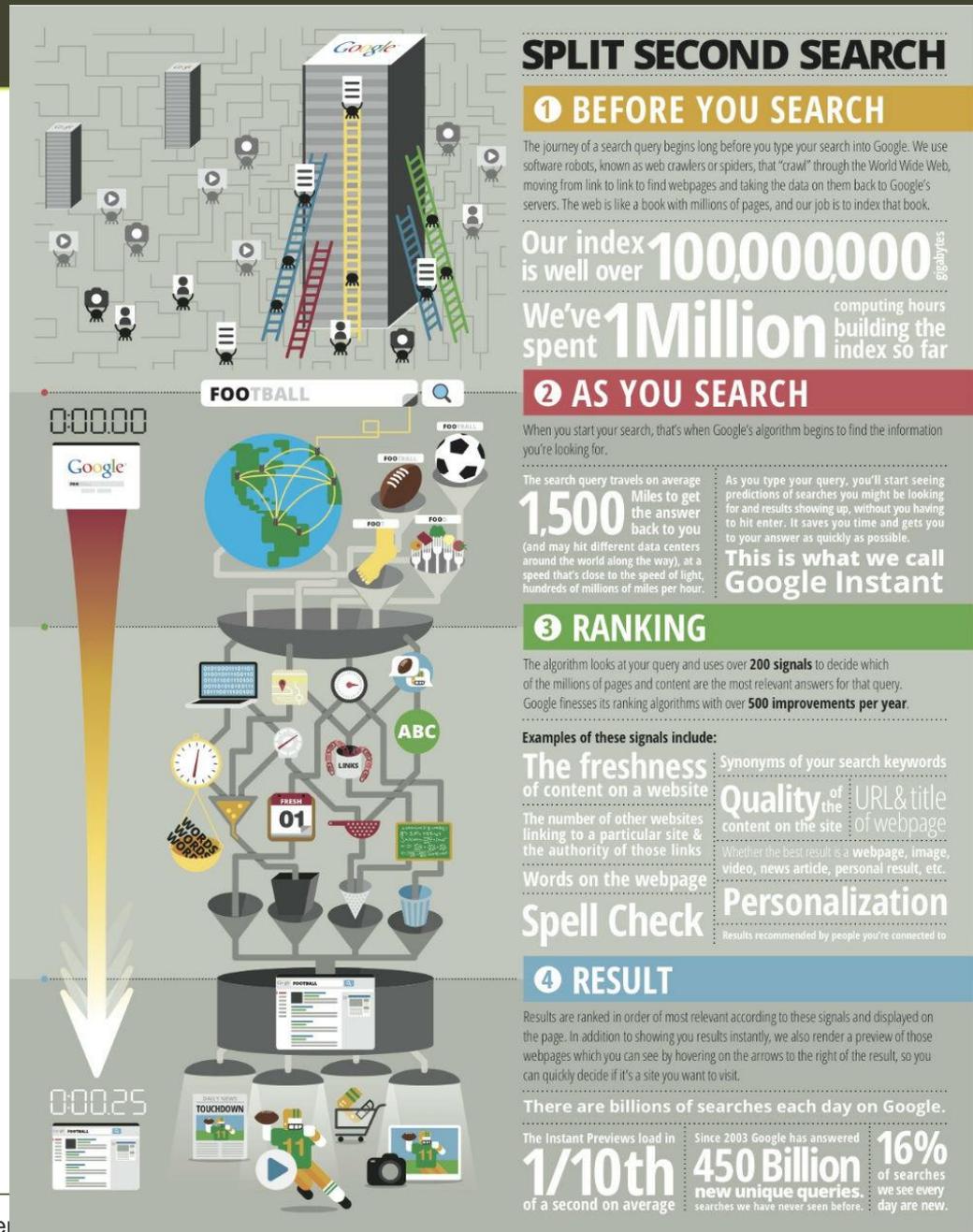
**Les moteurs:
tout est automatisé, sans intervention
humaine : un robot (ou spider ou crawler)
balaye le web à intervalle régulier
cherche les pages, les rapatrie dans une
base de données énorme et les indexe
automatiquement**

**Vidéo en anglais : Matt Cutts explique le
fonctionnement du moteur Google**

Mais dernière innovation et bataille concurrentielle : les moteurs essaient d'indexer en temps réel pour tenir compte des réseaux sociaux de Twitter, des blogs (descriptif de 8 moteurs ou liste)

Par conséquent infobesité de plus en plus importante aggravée par les outils web 2.0, les réseaux sociaux (Facebook, Twitter par ex) Carte

Sur le moteur de recherche Google la recherche s'effectue dans un index de 100 milliards de Go, la requête effectuée en moyenne 1 500 miles (2 500 km) pour revenir vers vous avec la page de résultats, etc. (juin 2012 site Abondance)



SPLIT SECOND SEARCH

1 BEFORE YOU SEARCH

The journey of a search query begins long before you type your search into Google. We use software robots, known as web crawlers or spiders, that "crawl" through the World Wide Web, moving from link to link to find webpages and taking the data on them back to Google's servers. The web is like a book with millions of pages, and our job is to index that book.

Our index is well over **100,000,000** gigabytes

We've spent **1 Million** computing hours building the index so far

2 AS YOU SEARCH

When you start your search, that's when Google's algorithm begins to find the information you're looking for.

The search query travels on average **1,500** Miles to get the answer back to you (and may hit different data centers around the world along the way), at a speed that's close to the speed of light, hundreds of millions of miles per hour.

As you type your query, you'll start seeing predictions of searches you might be looking for and results showing up, without you having to hit enter. It saves you time and gets you to your answer as quickly as possible.

This is what we call Google Instant

3 RANKING

The algorithm looks at your query and uses over **200 signals** to decide which of the millions of pages and content are the most relevant answers for that query. Google finesses its ranking algorithms with over **500 improvements per year**.

Examples of these signals include:

- The freshness of content on a website** (Synonyms of your search keywords)
- Quality of the content on the site** (Quality of the URL & title of webpage)
- The number of other websites linking to a particular site & the authority of those links** (Whether the best result is a webpage, image, video, news article, personal result, etc.)
- Words on the webpage** (Personalization)
- Spell Check** (Results recommended by people you're connected to)

4 RESULT

Results are ranked in order of most relevant according to these signals and displayed on the page. In addition to showing you results instantly, we also render a preview of those webpages which you can see by hovering on the arrows to the right of the result, so you can quickly decide if it's a site you want to visit.

There are billions of searches each day on Google.

The Instant Previews load in **1/10th** of a second on average

Since 2003 Google has answered **450 Billion** new unique queries. searches we have never seen before.

16% of searches we see every day are new.

Web visible et invisible :

Les robots (Googlebot ou autre) qui balayent tout le web ne savent pas récupérer certaines pages (on parle alors de web invisible ou caché ou profond, la partie immergée de l'iceberg)

Une partie seulement du web est visible

En 2007, 12%, beaucoup plus maintenant, 40, 60%?. Et dans ce web visible, les meilleurs moteurs n'indexent que 75%

Le web visible = tout ce qui peut être indexé par un moteur

**Index de Google entre 1998 et 2005 multiplié par 1000 (de 24 millions à 24 milliards de pages)
+ de 550 millions de sites web**

Les moteurs ont arrêté de donner la taille de leur index (Google en 2005)

Web invisible = ce à quoi que les moteurs ne peuvent accéder

Sites payants avec authentification (par ex Bases de données), pages dynamiques, formats spécifiques, etc





- **Les moteurs n'ont pas la même couverture du web (en utiliser plusieurs pour une recherche exhaustive)**
- **Les moteurs ne classent pas les résultats de la même façon (algorithmes différents et évoluant, Pagerank et autre) voir Thumbshots Ranking**
- **Les moteurs annoncent un nombre de résultats supérieur à ce qu'ils peuvent fournir**

Historique : (cours de l'Urfist de Rennes ou une carte):

Les premiers moteurs sont apparus à partir de 1990 : Archie, Wanderer

puis 2^{ème} génération vers 1994-95 : Yahoo, Voila, Altavista et en 1998 Google

3^{ème} génération en 2001, des moteurs plus « graphiques » (Exalead, Kartoo et Ujiko mais ces deux derniers ont disparus en 2010, etc)

Des nouveautés toujours (avec plus de spécialisation par type de ressource, de contenu, etc)

- Recherche de blogs [Technorati](#), Icerocket
- De forums [1001forums](#)
- De personnes [123People](#), [Pipl](#)
- D'images fixes ou animées [The Picture Factory](#)
- De vidéos [Blinkx](#)
- De sons [FindSounds](#)
- etc

- **D'articles de presse et d'actualités:**
Eufeds, Pickanews, Pearltrees Media France, Webplanète (811 sources francophones)
- Google actualités (avec la recherche avancée)
- **Outils d'analyse de l'actualité:**
- News Explorer
- Silobreaker

- Des moteurs intégrant des fonctionnalités web 2.0 (collaboratifs comme Wikia, Yoono, personnalisables comme Rollyo)
- très « visuels » comme Spezify ou Oamos
- Des moteurs sémantiques Wolfram Alpha ou Hakia
- On cherche.....gros enjeux pour capter l'audience dans la surabondance (économie de l'attention)

- **Les moteurs leaders (les plus utilisés) dans le monde sont :**
- **Google (88 milliards de requêtes), Yahoo (9, 4), mais aussi Baidu (chinois, 8,5), Bing de Microsoft (ex Live Search, ex MSN, 4)**

Baromètre des moteurs en France et pays européens

- **Classement des moteurs en France et dans le monde**

Les principes de base de la recherche documentaire s'appliquent aussi sur le web:

Les questions à se poser

- **Combien de temps pour ma recherche**
- **Type d'information souhaité (textes, statistiques, images, son, etc)**
- **Quantité d'information (quelques références, exhaustif, etc)**

- **Limites de ma recherche (langue, période, région ou zone géographique)**
- **Définir en quelques mots le sujet, les synonymes, traduire en anglais**

Principes propres aux moteurs de recherche sur le web:

Voir par ex les [principes de base de Google](#)

[Google: trucs et astuces](#) Béatrice Foenix-Rioux , ABBS

L'ordre des mots est important et conditionne l'affichage des résultats : le premier mot est le plus important
Pour mieux cibler sa recherche et éviter le « bruit », mettre aussi les mots les plus précis, beaucoup des mots, ceux qu'on s'attend à trouver dans les réponses
(dans le titre mais aussi le texte) !

- **Tris dans un moteur pour afficher :**
- **Place du mot dans le document (titre, début, majuscules)**
- **Fréquence// taille du document**
- **Poids du mot dans l'ensemble de la base (mot rare=poids, mots vides sous-évalués ou éliminés)**
- **Correspondance d'expression**
- **Proximité des termes demandés**

Et beaucoup d'autres critères secrets..

- **Le moteur interrogé a une incidence sur les résultats obtenus :**
Etude de 2007 sur les 4 moteurs leaders :
88% des résultats sont uniques à un moteur, et seulement 0,6% identiques sur les 4 moteurs
7 résultats sur 10 figurant sur la première page des moteurs ne figurent pas sur la première page des autres moteurs.
Le 1^{er} résultat n'est le même que dans 3,6% des cas

- **minuscules** : les moteurs ne tiennent pas compte de la casse (pas de différence entre majuscules et minuscules) pour les y forcer utiliser le + (ex pour un sigle:+IBM)
- **pas d'accent théoriquement** mais Google en tient compte
- Pour avoir la forme accentuée , utiliser le + (ex: +marché ou «marché» pour éviter marche)

- **pas de troncature** : la plupart des grands moteurs ne l'acceptent pas, sauf Exalead
- **Google en fait une quand il le juge utile....une lemmatisation en fait**
- **ne pas utiliser * qui a un autre sens=masque=remplacement d'un mot par ex bibliotheque*numérique recherche bibliotheque européenne ou mondiale numérique**
- **Taper donc toutes les formes singulier OR pluriel**

- **guillemets** pour obliger le moteur à chercher dans l'ordre donné une expression consacrée, un prénom et nom, etc.
- le **+** : oblige le moteur à prendre en compte l'orthographe exacte, utile notamment quand on veut chercher des mots vides (stop words = articles, prépositions) ou une forme accentuée

▪ **Opérateurs booléens:**

ET (intersection) : par défaut

OU (réunion) : OR en majuscule (ou |)

SAUF (exclusion) : – collé au mot à éliminer

Ex (limite) de recherche complexe avec des opérateurs booléens et parenthèses : ("matériaux composites" OR "materiaux composites") AND (acteurs OR sociétés OR fusion OR acquisition OR concentration OR acteur OR fusions OR acquisitions OR concentrations OR délocalisations)

Recherche d'un mot dans le titre

intitle:paris

Recherche de plusieurs mots dans le
titre

allintitle:tour eiffel

Recherche d'un mot ou plusieurs dans
l'adresse

inurl : ou **allinurl:**

Recherche dans un site ou un type de site
site:www.... et un mot : ne cherche ce mot
que dans ce site

Le plus utile **site:edu** cherche des sites
académiques (éducatifs, universitaires)
sur un sujet

On peut aussi chercher des .gouv ou .org,
sites de confiance, vu les producteurs

link:www... et le nom du site permet de trouver les liens de ce site donc des sites identiques, sur les mêmes thèmes

define: recherche de définitions

- **filetype:xxx** permet de chercher des fichiers du certain format pdf, powerpoint, doc, etc
- Très utile
- - quand on cherche un cours ou une conférence, ce sera surement un format de présentation powerpoint (.ppt)
- - quand on cherche un article , ce sera sûrement un format de document (.pdf ou .doc)

Il existe une recherche avancée dans les moteurs :

Il faut l'utiliser, elle permet d'affiner la recherche et de restreindre (au moins un peu!) la masse d'information, d'obtenir des réponses plus pertinentes donc de gagner du temps.

- Dans Google par ex., la recherche avancée permet d'affiner la recherche sans avoir besoin de connaître les commandes déjà citées
- À droite, au dessus de la barre de recherche, cliquer sur la roue dentée et sur Recherche avancée
- Consulter l'aide en ligne (en anglais)

Pages contenant

- **tous les mots suivants** : comme dans la barre de recherche simple, on met des mots les uns derrière les autres et l'opérateur booléen **ET** est mis automatiquement (=intersection = précision, restriction de la question)
- cette **expression exacte** : ceci équivaut à mettre des guillemets

- **au moins un des mots suivants** : ceci équivaut à mettre l'opérateur booléen OU (ou OR) = réunion = élargissement de la question, synonymes ou termes équivalents
- **aucun des mots suivants** : ceci équivaut à mettre l'opérateur booléen SAUF (ou NOT) = exclusion = supprimer un terme (ou plusieurs)

- **Limitations par langue ou région** : en sélectionnant dans la liste, utile surtout pour les langues
- **Format de fichier** : très utile, pour n'avoir que des PDF par ex., donc des documents à coup sûr
Idem pour le format .doc
Très utile aussi pour sélectionner des présentations Powerpoint (donc souvent des cours)

- **Date** : permet de rechercher les pages Web découvertes (par le robot) durant une période : pas très performant, ce n'est donc pas une limitation par date des documents ou des informations mais par date de leur apparition sur le web
- **Emplacement** : permet de trier le ou les termes à un endroit dans les pages, seulement dans le titre par ex, équivaut alors à intitle:xxx, ceci limite beaucoup la question mais attention trop parfois.
- **Domaines** : permet d'afficher les pages du site ou du domaine par exemple .org, .fr, etc. équivaut à site:xxx Très utile de trier les sites edu c'est-à-dire les sites éducatifs donc de niveau universitaire, de confiance

- **Droits d'utilisation** : permet de savoir si vous pouvez réutiliser ou non

- **SafeSearch** : ce filtre évite les sites à caractère sexuel, pour les enfants

- **Recherche de pages spécifiques**
 - Similaires** Rechercher les pages similaires à cette page
 - Liens** Rechercher les pages liées à cette page (donc de contenus sans doute similaires ou approchants)

Comment sont affichés les résultats dans un moteur ?

Le **page-rank de Google:**

système de classement des résultats assez opaque selon un algorithme non explicité par le moteur, 2 grands principes mélangés (pertinence et popularité):

- **indice de pertinence : fréquence d'apparition des termes demandés et leur place dans les pages**

**1ers résultats = documents (pages)
contenant le plus de fois les mots de la
requête et le plus haut dans les pages
(titre, début)**

- **Indice de popularité :**
**principe de la citation : liens que
d'autres sites font vers le mien et
ceux que je fais vers d'autres (bien
côtés aussi)**

- **D'autres classements dans d'autres moteurs , Exalead par ex:**
- **Calcul dynamique de catégories (facettes)**

Organisation d'un lot de résultats pour affiner ou étendre la recherche (par « mots clefs », auteurs, type de docs, zone géographique, etc

Mais aussi positionnement payant (les 1ers résultats de couleur différente)

Les moteurs spécialisés

Nous avons vu qu'il existe des moteurs de tous types.

Ceux qui nous intéressent ici sont les moteurs spécialisés académiques (= de niveau universitaire) pour faire de la recherche documentaire

Toutes disciplines :

- Google Scholar (2004): On y trouve de la littérature scientifique gratuite : des articles (soit de revues en ligne gratuites, soit en pré-print = avant publication), des thèses, rapports, etc la fiche descriptive de Google Scholar (par JP Lardy, Urfist de Lyon)

Approfondir Google Scholar (par J. Sicot Ecole centrale de Lyon)

- Il y a beaucoup de texte intégral, parfois seulement la référence (d'un livre par ex.)
- On sait combien de fois l'article a été cité (son impact, sa valeur) et par qui (renvoi à d'autres articles).
- C'est un bon **complément** aux bases de données spécialisées.

Plus orienté sciences et sciences médicales

- **Scirus (par Elsevier)** 16 000 titres de 4 000 éditeurs, en majorité les domaines médecine, science et technique, seulement 1 600 titres en SHS, soit 10% de son catalogue, en avril 2009 doublement du nombre de revues en sciences humaines .

La **fiche** descriptive et un diaporama Scirus/Google scholar

- Des réservoirs d'articles plus petits (agrégeant plusieurs éditeurs ou d'un seul éditeur, des archives ouvertes, etc)
- CiteUlike (Springer): 5,5 millions d'articles
- Highwire (par l'univ. de Stanford) 1530 publications, 6,7 millions d'articles
- Ingenta Connect : sciences et techniques et biomédical, 4,5 millions de documents
- Techxtra (Sci. Ingénieur, mathématiques, informatique) 4 millions de documents
- Citeseer
- Scitopia (Maths, physique, sci. ingénieur) 3,5 millions de documents
- Scientific commons

EN SHS

- Jurn (arts et humanités) 4348 revues électroniques
- Economics Search engine
- Isidore , moteur spécialisé français cherchant dans 1750 sources (archives ouvertes, revues gratuites), répertoriant 1 800 000 ressources

Des outils entre le portail et le moteur
spécialisé

pour le droit : **Légicité**

<http://www.legicite.com>

Ou [Lawcrawler](#)

Pour la médecine : **Medhunt**

http://www.hon.ch/MedHunt/MedHunt_f.html

Pour la Pharmacie :

<http://www.pharmalivesearch.com/>

- **Trouver des articles dans des réservoirs (ou bouquets) gratuits de revues électroniques (voir le répertoire DOAJ)**
- **Des réservoirs en Lettres, sciences humaines et sociales afin de promouvoir les revues dans ces disciplines (initiatives venant du Ministère ou d'organismes de l'enseignement supérieur)**

- Les réservoirs décrits ci-dessous sont regroupés depuis janvier 2011 dans Isidore
- Cairn, 203 revues, en partie gratuit, les articles les plus récents, 2 à 5 ans selon les revues (barrière mobile) sont sur abonnement
- Revue.org, 238 revues, la moitié environ avec barrière mobile de 1 à 5 ans, le reste en accès libre en totalité
- Persee, 82 revues, accès libre quasi total

Cairn , Persee et Revues.org ont crée une continuité d'accès à certaines revues qu'ils diffusent (par ex le courant sur Revues.org, le rétrospectif sur Persée mais « transparent pour l'utilisateur)

- **Erudit, 66 revues, portail canadien, contient aussi des sciences naturelles, barrière mobile 2 ans**

(ne fait pas partie d'Isidore)

- **Indexsavant : le portail des revues scientifiques de la Francophonie**

- Dans des réservoirs d'archives ouvertes
- Oaister (moissonneur)
- Hal (l'archive française toutes disciplines)

<http://hal.archives-ouvertes.fr/>

- Arxiv (sciences exactes)

<http://arxiv.org/>

- etc

Les agents intelligents (ou métamoteurs clients, à installer sur son poste) :

Recherche intégrée dans l'outil et gestion d'une base de données personnelle

Recherche sur les résultats récupérés

Mise en place d'une veille, etc

Par ex: Copernic (version de base gratuite)

Frédéric Martinet Veille Copernic

ou Strategic Finder

Evaluation de l'information sur Internet

tout le monde peut publier sans filtre éditorial, sans être une autorité dans le domaine, donc liberté mais....

beaucoup d'information (info-pollution, infobésité) mais aussi des informations partisans, déviantes, racistes ou tout simplement fausses, dépassées ou médiocres...

- **Brouillage de diverses frontières : contenu/support, auteur/lecteur, genres documentaires, etc**
- **faire preuve d'esprit critique**
- **on peut utiliser ici aussi le système de questionnement 3QOCP**

Qui ?

**Qui est l'auteur ou les auteurs du site,
l'hébergeur**

**Regarder l'URL , le domaine (.com,.gouv.,
.edu),**

Quoi ?

**le sujet ou thème général du site bien
annoncés**

**Info unique ou sur beaucoup de sites, en
vrac ou organisée**

**Public auquel s'adresse le site
(spécialistes, tout public)
une bibliographie, des références à
d'autres travaux ou sites bien cités**

Quand ?

**Date de création du site, dernière MAJ,
liens à jour vers d'autres sites**

**Où ? nationalité des renseignements,
serveur sur lequel le site est hébergé**

Comment ?

**quelle forme, polémique, point de vue
contradictoire pris en compte, etc**

**Infos présentées de manière logique,
bien organisées**

Site facile à visiter, clair, lisible

**Illustrations utiles, pertinentes, qualité
visuelle des dessins, photos, animations**

Orthographe et niveau de langue

Pourquoi ?

Quelle est la motivation du créateur du site (altruisme, narcissisme, capitalisme, prosélytisme)

publicités sur le site clairement séparées du contenu

Fiche récapitulant les questions à se poser

J'évalue des sites Web

- **Gérer sa bibliographie:**
- Voir le **cours de ML Malingre** Urfist de Rennes
- Avec des outils payants : Endnote, Refworks, etc
- Ou gratuits Zotero
- **CiteULike**

Bon courage pour l'exploration du monde merveilleux de l'information

