Recherche d'information : contours et bonnes pratiques

Pr. Mohand Boughanem
bougha@irit.fr
http://www.irit.fr/~Mohand.Boughanem
Université Paul Sabatier de Toulouse
Laboratoire IRIT, UMR5055
118 Route de Narbonne F-31062 Toulouse

Plan

- Comprendre ce qu'est la Recherche d'information (RI)
  - Intérêt et contours
  - Fonctionnement «interne» d'un système de RI
- Cas pratique : Recherche d'information sur le Web
  - Bonnes pratiques en tant que producteur d'information
    - → Valoriser son site
  - Bonnes pratiques en tant que consommateur de l'information
    - → Utiliser le bon outil
- Conclusion : nos activités de recherche en RI

Qu'est ce que la RI ?

- Recherche d'information (RI) Ensemble des méthodes, procédures et techniques pour l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information (données, texte, son, images, vidéo).

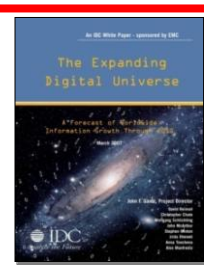
Exemples de Systèmes de RI

Volume sans cesse croissant

- Gros volumes d'informations (numériques) créées toutes les minutes
- 1% sur l'Internet (localisé par les moteurs)
- 99% dans des Intranet (entreprises, laboratoires, …)

.. en perpétuelle croissance

KiloOctets	10^3
MegaOctets	10^6
GigaOctets	10^9
TeraOctets	10^{12}
PetaOctets	10^{15}
ExaOctets	10^{18}
ZettaOctets	10^{21}



Facteur de 10 en 5 ans!

Source: IDC, 2008

.. produite par tout le monde

.. L'information (numérique) est disponible partout

… dans tous les domaines d'activités

- Domaines d'application
  - Internet (Web, *Forum/Blog search*, News)
  - Entreprises
  - Bibliothèques numériques «*digital library*»
  - Domaine spécialisé (médecine, droit, littérature, …)
  - Nos propres PC (*Yahoo! Desktop search*)

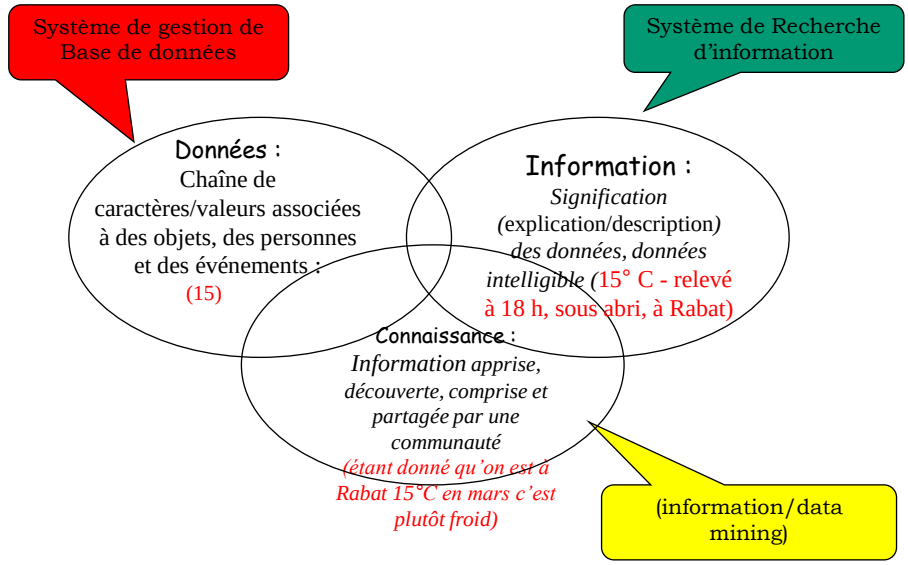
… la RI a un coût

- Rechercher une information a un coût
  - « On» passe (en moyenne) 35% de son temps à rechercher des informations
  - Les *managers* y consacrent 17% de leur temps
  - Les 1000 grandes entreprises (US) perdent jusqu'à $2.5 milliards par an en raison de leur incapacité à récupérer les bonnes informations
- Nécessité de développer des systèmes automatisés efficaces permettant :
  - Collecter, Organiser, Rechercher, Sélectionner

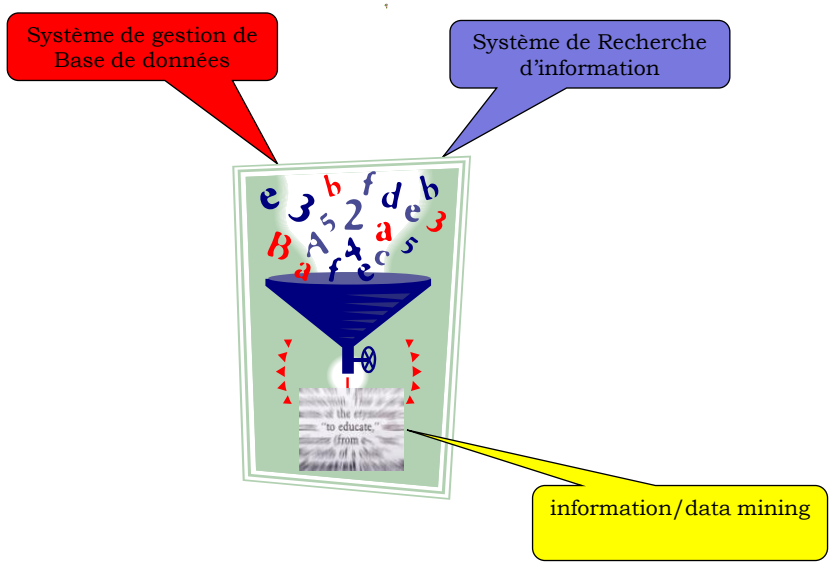
Contours de la RI

- Données, information et connaissance
- Tâches de recherche d'information

Contours:
Données-Information-Connaissance

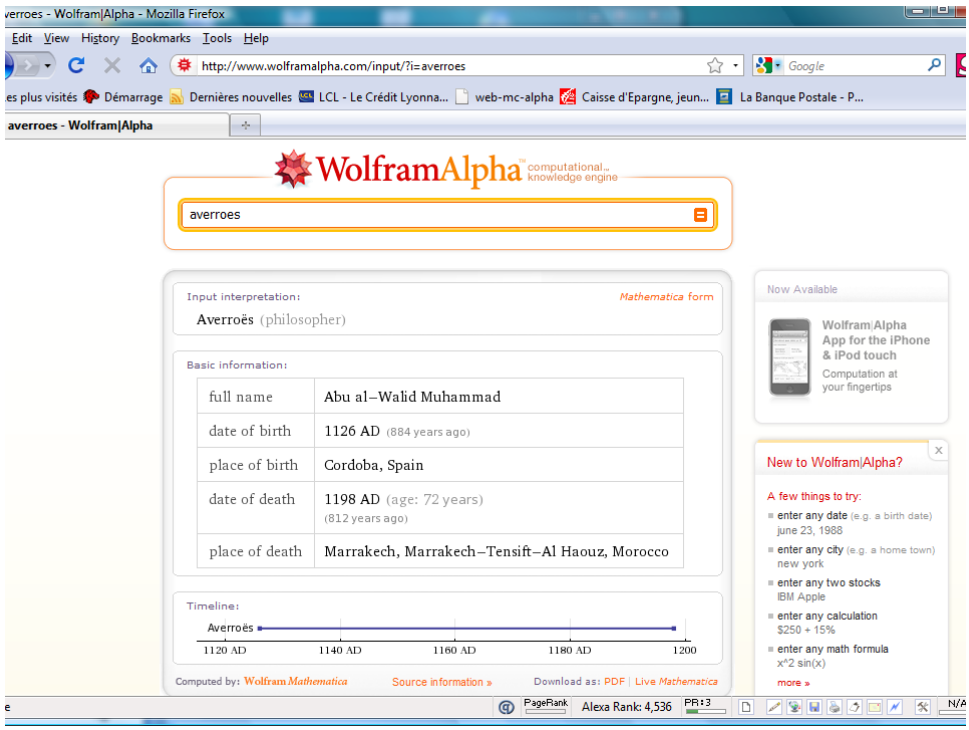


Veille a besoin des trois domaines



Contours :
Tâches de recherche d'information (1/2)

- Recherche adhoc (classique)
  - Je cherche des infos (pages web) sur «veille économique »
  - Requête «veille économique»→ SRI → renvoie une liste de documents
- Classification /catégorisation (*clustering*)
  - Regrouper les informations (documents) selon un ou plusieurs
- Question-réponses (*Query answering*)
  - Chercher des réponses à des questions
  - par exemple «qui est averroes ? »
  - « Quel est la longueur du Nil ? »



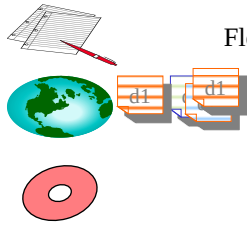
Contours :
Tâches de recherche d'information (2/2)

- Filtrage d'information (*filtering/recommendation*)
  - Dissémination sélective d'information

Filtrage d'information

Sources

Utilisateurs ayant des besoins définis au préalable
= Profils



Flot de documents

SFI/push

Terminologie :
- Push
- Filtrage d'information
- Recommandation
- Dissémination sélective d'information
- Système d'alerte

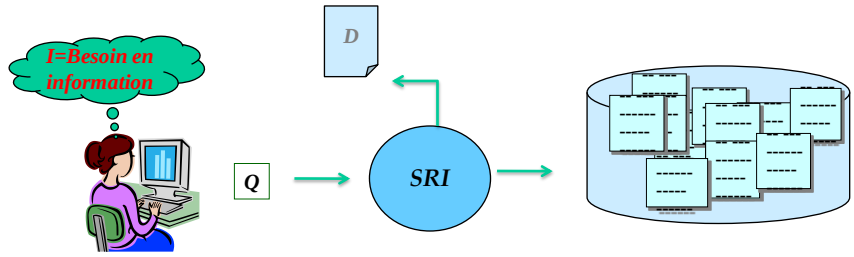
Contours : Tâches de recherche d'information (2/2)

- Filtrage d'information (*filtering/recommendation*)
  - Dissémination sélective d'information
- Croisement de langues (*cross language*)
  - Rechercher des infos écrites dans une langue autre que celle de la requête
  - Requête en français → retour documents en arabe
- Métat-moteurs (*Meta-search*)
  - Moteurs interrogeant plusieurs moteurs
- ….

Plan

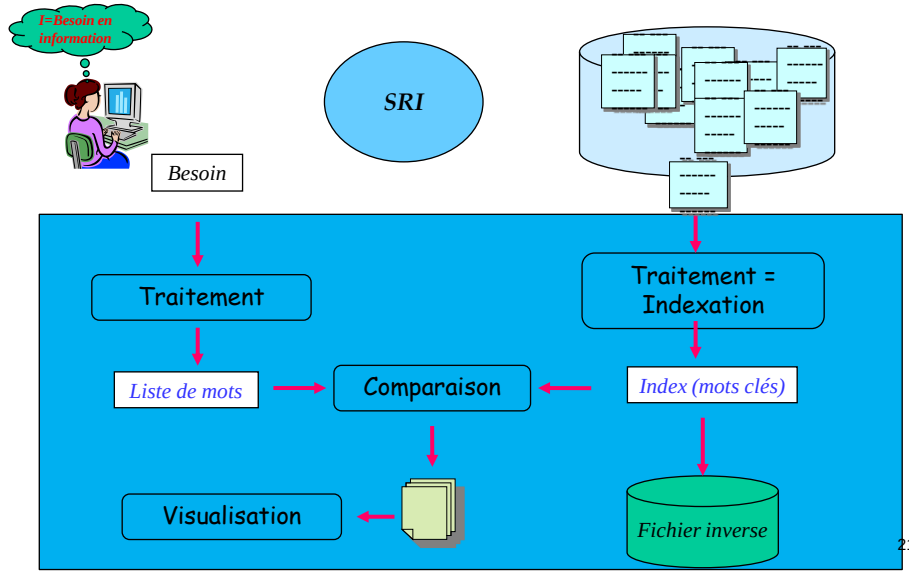
- Comprendre ce qu'est la Recherche d'information (RI)
  - Intérêt et contours
  - Fonctionnement «interne» d'un système de RI
- Cas d'étude : Recherche d'information sur le Web
  - Bonnes pratiques en tant que producteur d'information
  - Bonnes pratiques en tant que consommateur de l'information

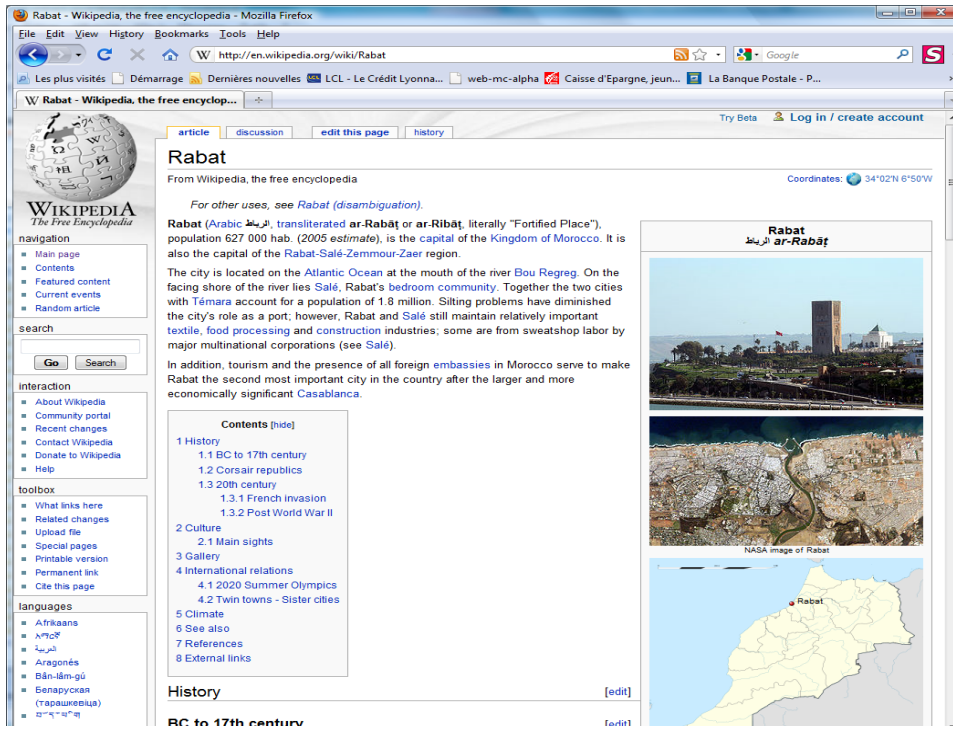
Comprendre le fonctionnement d'un SRI

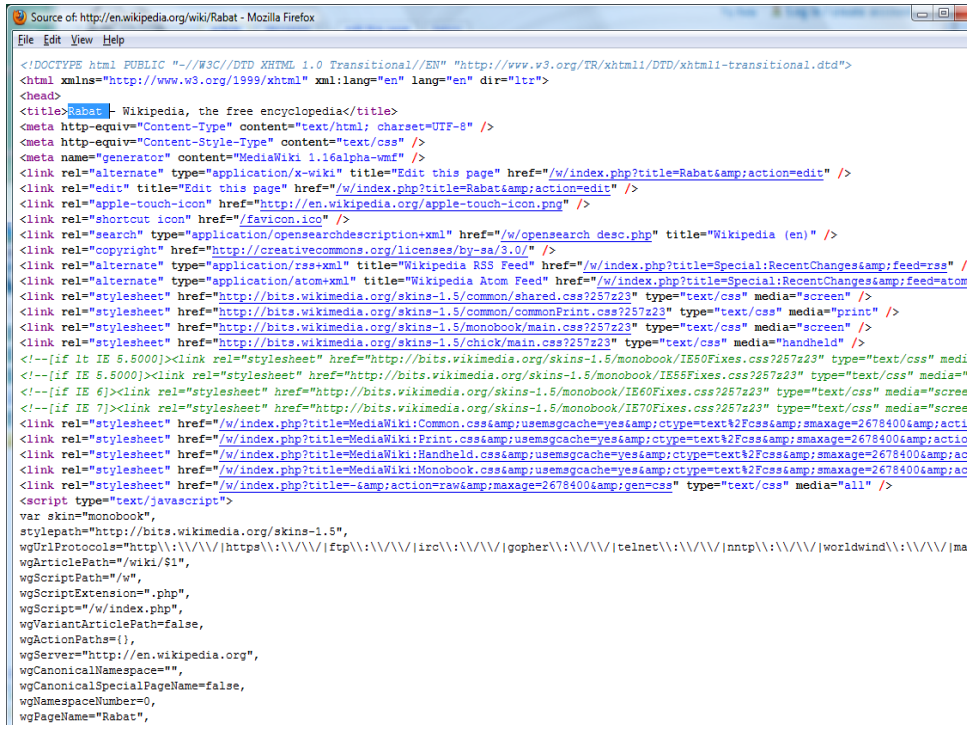


- Sélectionner dans une collection
  - les **informations**
  - **… pertinentes** répondant aux
  - … **besoins en information des utilisateurs**

Fonctionnement d'1 SRI : Ouvrir la boîte noire







Traitement – indexation

- Décomposer le texte
- Décomposer les mots
- Supprimer les mots communs
  - Basé sur une “short list” “the”, “and”, “or”
- Radicaliser les mots
- Regrouper les mots

Title: Rabat
Corps du texte : Rabat or Ribat is located in the atlantic Ocean

Rabat, Rabat, or, Ribat, is, located, in, the, atlantic, Ocean

Rabat, Rabat, Ribat, located, atlantic, Ocean

Rabat, Rabat, Ribat, locate, atlantic, Ocean

Rabat 2, Ribat 1, locate 1, atlantic 1, Ocean 1

Organiser les termes et les documents dans un Fichier Inverse

d1: So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious

d2: I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.

Traitement = Indexation

Term	N docs	Tot Freq	Ptr
ambitious	1	1	1
be	1	1	2
brutus	2	2	3
capitol	1	1	5
caesar	2	3	6
did	1	1	
enact	1	1	
hath	1	1	
I	1	2	
i'	1	1	
it	1	1	
julius	1	1	
killed	1	2	
let	1	1	
me	1	1	
noble	1	1	
so	1	1	
the	2	2	
told	1	1	
you	1	1	
was	2	2	
with	1	1	

Doc #	Freq
1	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
1	1
2	1
1	2
1	1
2	1
1	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1

Répondre à une demande (requête)

caesar

Term	N docs	Tot Freq	Ptr
ambitious	1	1	1
be	1	1	2
brutus	2	2	3
capitol	1	1	5
caesar	2	3	6
did	1	1	
enact	1	1	
hath	1	1	
I	1	2	
i'	1	1	
it	1	1	
julius	1	1	
killed	1	2	
let	1	1	
me	1	1	
noble	1	1	
so	1	1	
the	2	2	
told	1	1	
you	1	1	
was	2	2	
with	1	1	

Doc #	Freq
2	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
2	1
1	2
1	1
2	1
1	1
2	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
2	1
1	1
2	1
2	1

Répondre à une demande : trier les documents

- Calculer un score de pertinence pour chaque document
  - Prendre en compte plusieurs indicateurs :
    - Fréquence du terme dans le doument (*tf*), sa fréquence dans la collection (*idf*), sa position dans le texte(p), taille du document (*dl*) ...

$$Score(D) = fonction(tf, idf, dl)$$

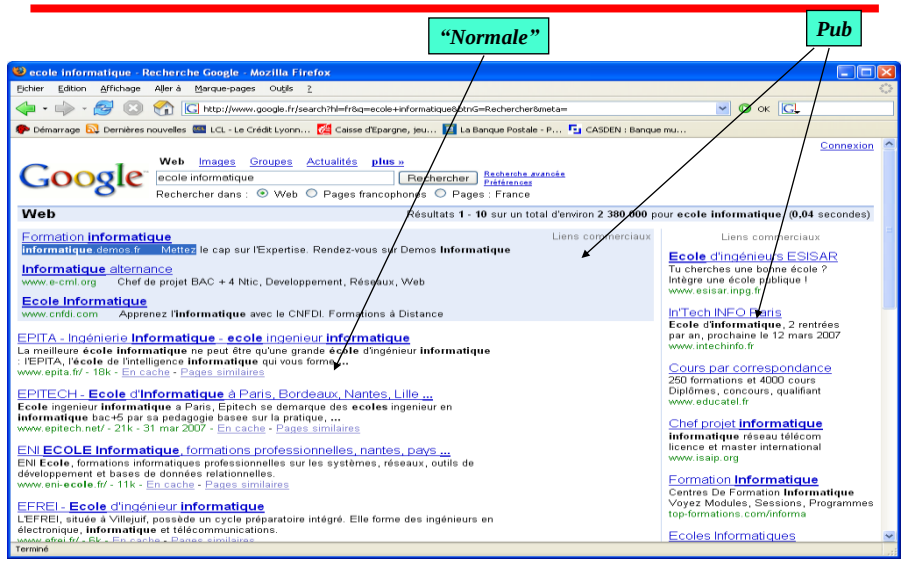
Plan

- Comprendre ce qu'est la Recherche d'information (RI)
  - Intérêt et contours
  - Fonctionnement «interne» d'un système de RI
- Cas d'étude : Recherche d'information sur le Web
  - Bonnes pratiques en tant que producteur d'information
  - Bonnes pratiques en tant que consommateur de l'information

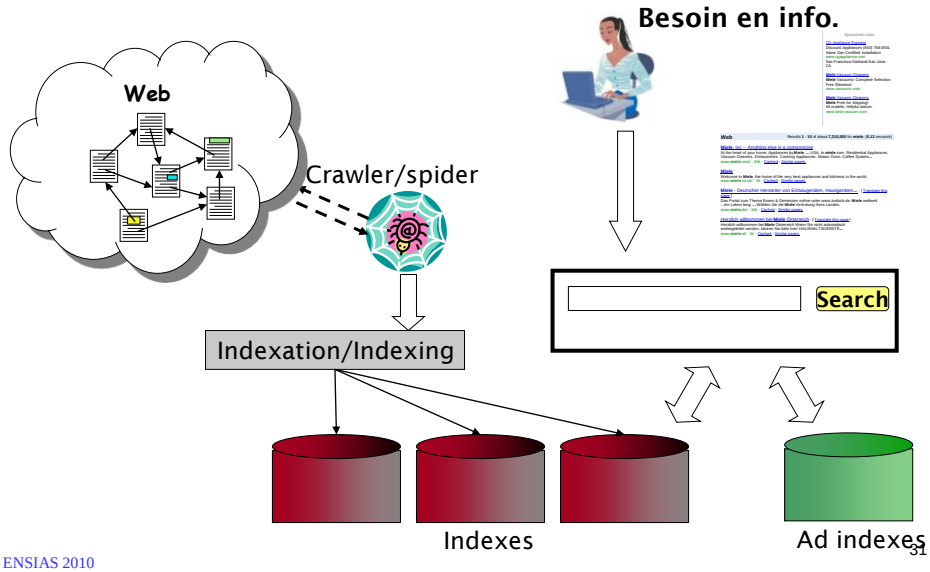
Web

- **World Wide Web**, communément appelé le **Web**, **Toile**, littéralement la « toile (d'araignée) mondiale »
- Ensemble de Pages Web connectées (reliées)
  - Statique/Dynamique
  - Distribué
  - Documents sont volatiles (ERROR 404)
- Comporte de tout
  - information avérée, contradictoire, fausse, obsolète, spam …
  - Structurée , semi-structurée (<TITLE>, <B>, <H1>, <H2>, etc.)
- Collections très volumineuses et hétérogènes
  - Web : plus de 170 TO (10^{12} octets)
  - Google : Dizaine de milliards de Pages
- Millions d'utilisateurs, efficacité des accès

Exemple de résultats d'un moteur de recherche



Approche de base de la recherche sur le Web



Bonnes pratiques

- Producteurs de l'information
  - Savoir référencer son site
  - Mettre de bons indicateurs (mots clés) pour que son site (page) soit sélectionné(e) en tête des résultats ? →Optimiser son site (Search engine optimization (SEO))
- Consommateurs
  - Choisir le bon outil pour ma recherche
  - Savoir l'utiliser (interrogation)

Producteurs d'info. : référencer son site

- Plusieurs moteurs proposent des liens gratuits
  - Google
    - http://www.google.com/addurl
  - Bing
    - http://www.bing.com/webmaster/SubmitSitePage.aspx
  - DMOZ (Open Directory)
    - http://www.dmoz.org/World/
  - Yahoo Directory
    - http://fr.docs.yahoo.com/info/ajouter.html

Producteurs d'info. : Optimiser son site

- Optimiser son site → Mettre de bons indicateurs (mots clés) pour que son site (page) soit sélectionné(e) en tête des résultats
  - Search engine optimization (SEO)
  - Aussi : Search Engine Marketing (SEM)
  - Sites à visiter
    - http://en.wikipedia.org/wiki/Search_engine_optimization
    - http://fr.wikipedia.org/wiki/Optimisation_pour_les_moteurs_de_recherche
    - http://searchenginewatch.com
    - http://www.abondance.com
    - http://googleblog.blogspot.com
- Spamdexing vs. SEO

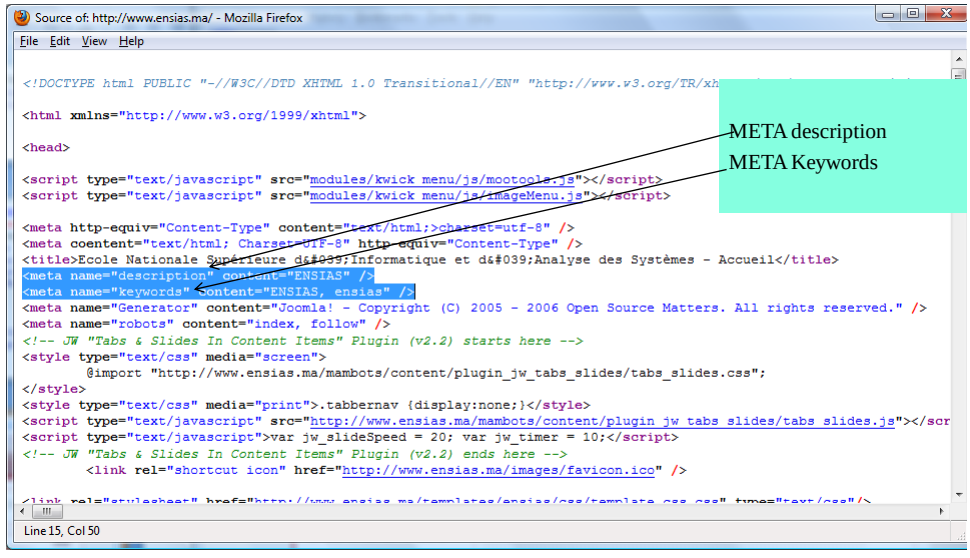
Producteurs d'info. : optimiser son site

- Quels sont les facteurs/indicateurs utilisés par les moteurs de recherche pour trier (sélectionner) les pages Web ?
- Beaucoup de facteurs plus de 200
- La manière dont ces facteurs sont combinés est spécifique à chaque moteur de recherche, généralement « secrete »

Facteurs utilisés par les moteurs pour trier des pages (1/5)

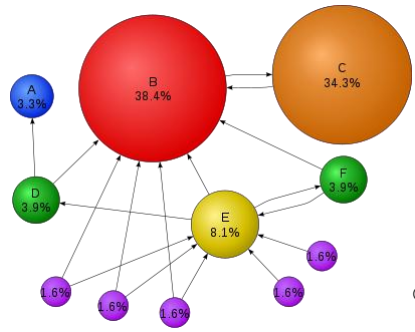
- Deux groupes de facteurs
  - Contenu de la page web (mots-clés dans la page web)
    - Classement des pages dépend fortement de la localisation (position) du mot dans la page
    - → Titre page, mots dans ancre (anchor), domaine, URL, Meta…


Mots dans le titre
Mots dans le domaine
Mots dans un lien(anchor)
ENSIAS
LA GRANDE ÉCOLE
DU GÉNIE INFORMATIQUE


Source of: http://www.ensias.ma/ - Mozilla Firefox
META description
META Keywords
Line 15, Col 50

Facteurs utilisés par le moteurs pour trier les pages (2/5)

- Deux groupes de facteurs
  - Contenu de la page web (mots-clés dans la page web)
  - Popularité du site/page web (*PageRank)*)
    - PR(p)= f(nombre de sites pointant p, popularité de ces sites, …)



© ***wikipédia***

- http://www.seomoz.org/
- http://www.seomoz.org/article/search-ranking-factors

Facteurs utilisés par les moteurs pour trier les pages (3/5)

- Mots dans le titre de la page (Keyword Use in Title Tag)
- Popularité du site (PageRank)
- Mots dans les liens (‘anchor’) provenant de l’extérieur (Keyword Focused Anchor Text from External Links)
- Diversité des liens externes (Diversity of Link Sources)
- Fiabilité du domaine (fonction de la distance de votre site vis-à-vis d’un site de confiance, cette liste est inconnue)
- Mots dans le nom du domaine (www.ensias.ma) (Keyword Use in the Root Domain Name)
- Mots dans le nom du sous-domaine (mots.domaine.ma) (Keyword Use in the Root Domain Name)

Facteurs utilisés par les moteurs pour trier les pages (4/5)

- Mots clés dans la balise H1 (Keyword Use in H1)
- Mots dans les balises H2, H3 Page URL
- Mots clés dans les Meta Description / Keywords Tags
- Mots dans le Body (Keyword Use in Body Text)
- HTML Validation

Facteurs utilisés par les moteurs pour trier les pages (5/5)

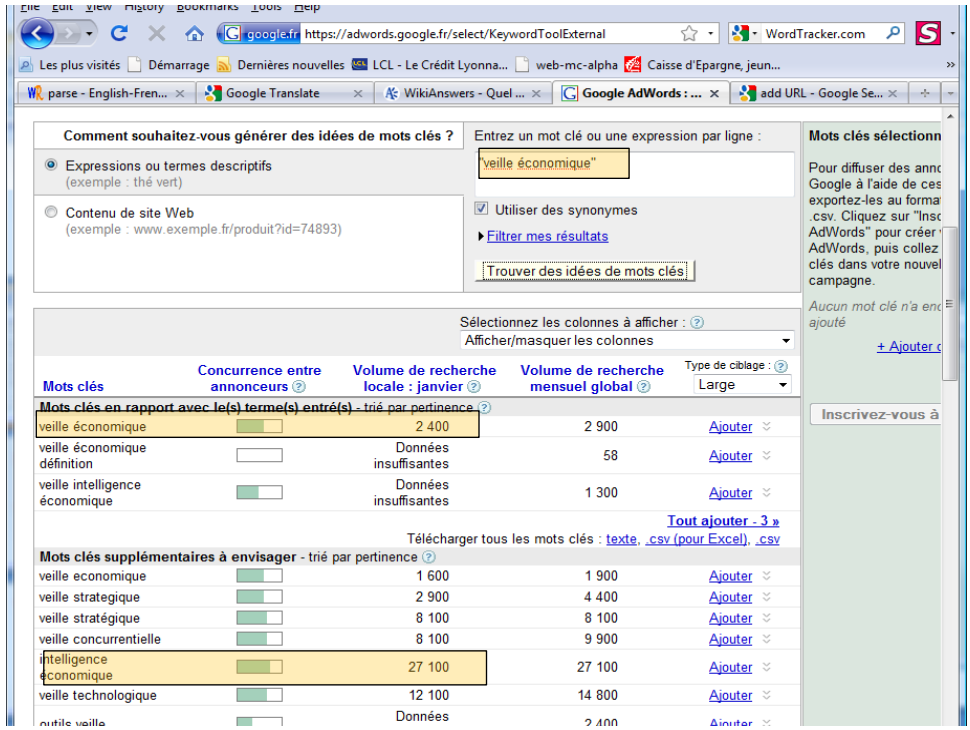
- **Top 5 facteurs négatifs**
  - Serveur souvent inaccessible
  - Contenu similaire ou dupliqué (copie de site)
  - Liens vers des pages de qualité douteuse (Spam)
  - Liens provenant de sites (courtier/"Vendeurs" de liens)
  - Sur-utilisation de mots clés

D'après Amit Singhal (Architecte de Google) lors de sa conférence à ECIR '2008

Google utilise 263 facteurs

Choix des mots → selon leur fréquence d'utilisation par les internautes

- Outils mesurant la popularité des mots
  - Overture's Keyword Selector Tool (http://inventory.overture.com)
  - WordTracker.com
  - Trellian's KeywordDiscovery.com
  - Google Adwords
  - Google Trends
  - Google Suggest

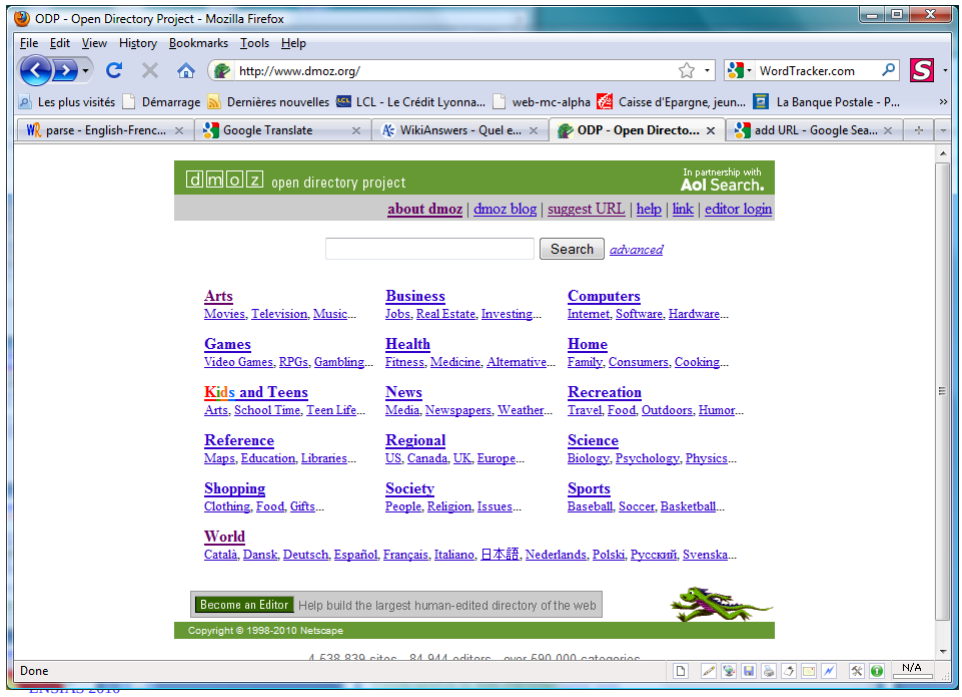


Bonnes pratiques : consommateur de l'info

- Quel est le meilleur moteur de recherche ?
  - Il ya plus de 300 moteurs de recherche disponibles pour une utilisation libre. Chacun possède ses propres atouts
- Choisir son moteur selon ses besoins et sa tâche
- Utiliser les différentes possibilités du moteur (en particulier en terme d'interrogation)

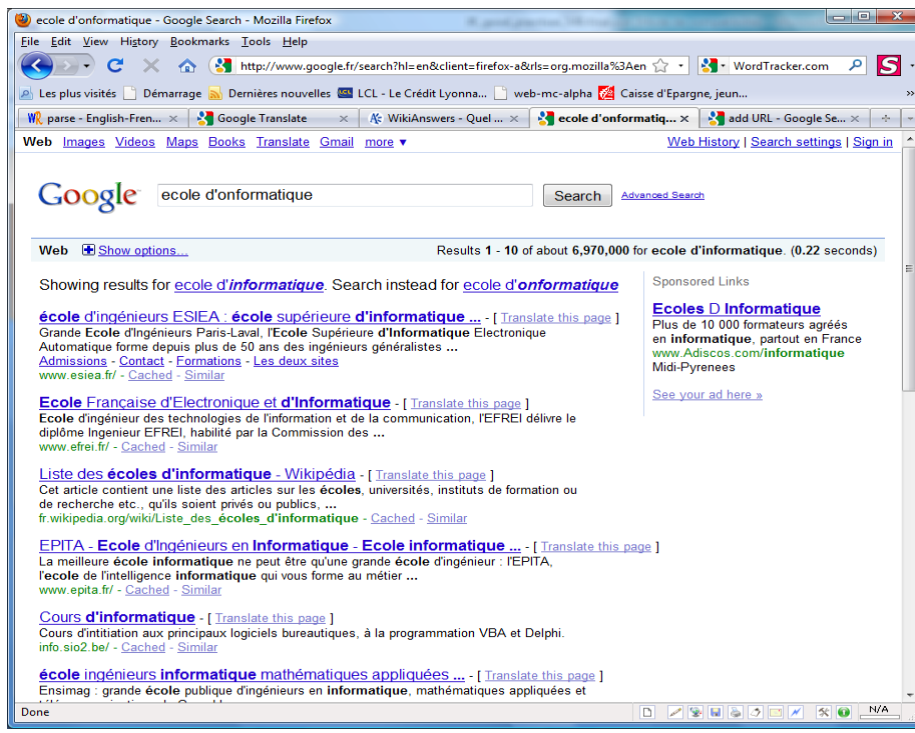
Choisir son moteur

- Mode d'accès
  - Mots (clés) : Google, Bing, Exalead, …
  - Annuaires : Yahoo, DMOZ, About , …
- Mode de visualisation/présentation
  - Liste linéaire "à la google"
  - Classification (Clusty/ search-cube)
- Type de moteurs
  - Généralistes vs. Verticaux (type de médias/thématiques)
- Autres facteurs :
  - Nombre de pages/sites indexés, fraîcheur, type de documents: (PDF, Word, Excel, PowerPoint, …), temps de réponse et consistance (pas de spam!!)

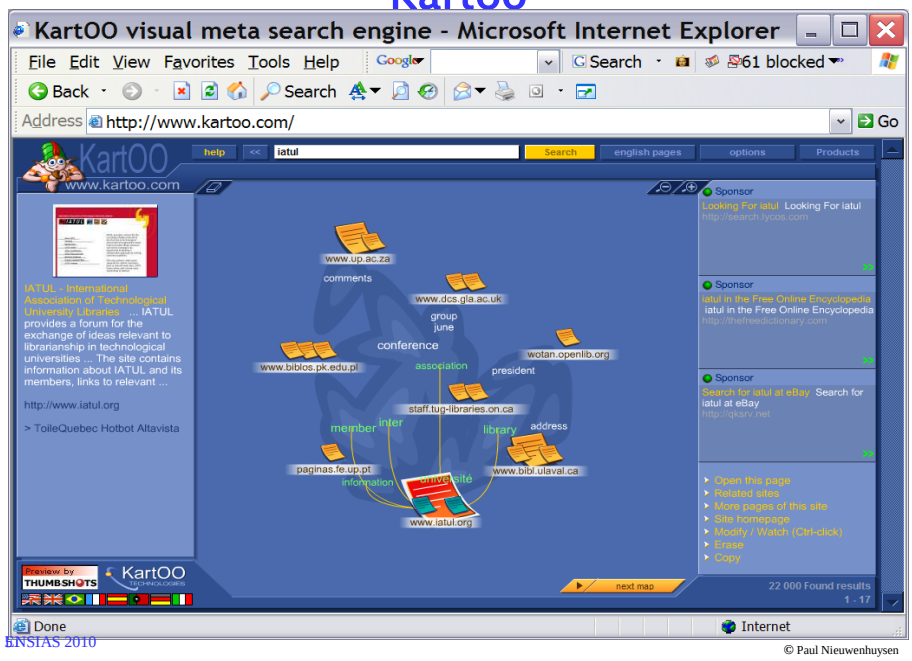


Choisir son moteur

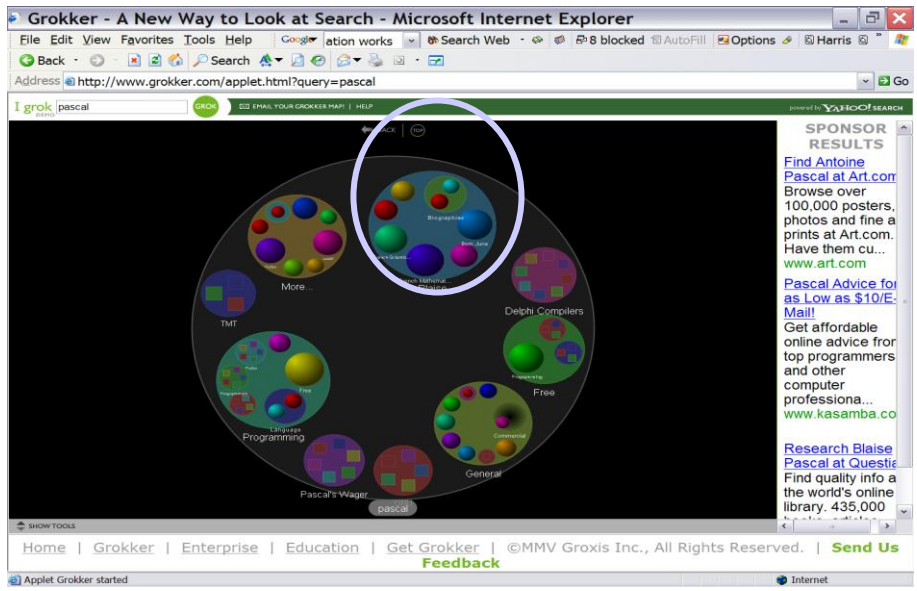
- Mode d'accès
  - Mots (clés) : Google, Bing, Exalead, …
  - Annuaires : Yahoo, DMOZ, About , …
- Mode de visualisation/présentation
  - Liste linéaire "à la google"
  - Classification (Clusty/ search-cube)
- Type de moteurs
  - Généralistes vs. Verticaux (type de médias/thématiques)
- Autres facteurs :
  - Nombre de pages/sites indexés, fraîcheur, type de documents: (PDF, Word, Excel, PowerPoint, …), temps de réponse et consistance (pas de spam!!)



Kartoo



Grokker

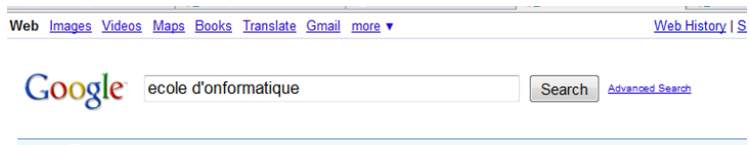


Choisir son moteur

- Mode d'accès
  - Mots (clés) : Google, Bing, Exalead, …
  - Annuaires : Yahoo, DMOZ, About , …
- Mode de visualisation/présentation
  - Liste linéaire "à la google"
  - Classification (Clusty/ search-cube)
- Type de moteurs
  - Généralistes vs. Verticaux (type de médias/thématiques)
- Autres facteurs :
  - Nombre de pages/sites indexés, fraîcheur, type de documents: (PDF, Word, Excel, PowerPoint, …), temps de réponse et consistance (pas de spam!!)

Type de moteurs (1/3)

- Généralistes
  - Google, Yahoo, Bing, ask, AllTheWeb.com (http://www.alltheweb.com)
  - Hybrid Engines -Métamoteurs : DogPile, MetaCrawler
- Multimedia Search Engines
  - Image, Audio & Video Searching



Type de moteurs (2/3)

- Moteurs spécialisés (vertical search Engine) : « focus » sur un sujet, domaine, …
  - Finance (www.searchFinance.com)
  - Technologie (www.knowledge storm.com)
  - Droit (Legal Search Engines)
  - Science : **Scirus ,** CiteSeerX (citeseerx.ist.psu.edu) pour l'informatique et IT
  - *Medical Search Engines (Medhunt (www.hon.ch/MedHunt))*
  - *Travel Search Engines*
  - *Chemfinder (www.chemfinder.com)*
  - *Expert search finding*
  - *Genomic information retrieval*
  - *Geographic information retrieval*
  - …

Type de moteurs (3/3)

- News Search Engines
  - Google News
  - Yahoo News (http://news.yahoo.com/)
  - AltaVista News http://news.altavista.com/

- Moteurs "temps réel"
  - Twitter
  - Topsy
  - OneRiot
  - Wowd

Trouver son moteur de recherche

- Search Engine Colossus
  - http://www.searchenginecolossus.com
  - Bonne liste de moteurs listés par pays et langue
- Search Engine Watch
  - http://searchenginewatch.com/links/
  - Listes de moteurs selon leur domaine.
- Tous les moteurs de recherche
  - http://www.allsearchengines.com/ and http://www.allsearchengines.com/complete.html

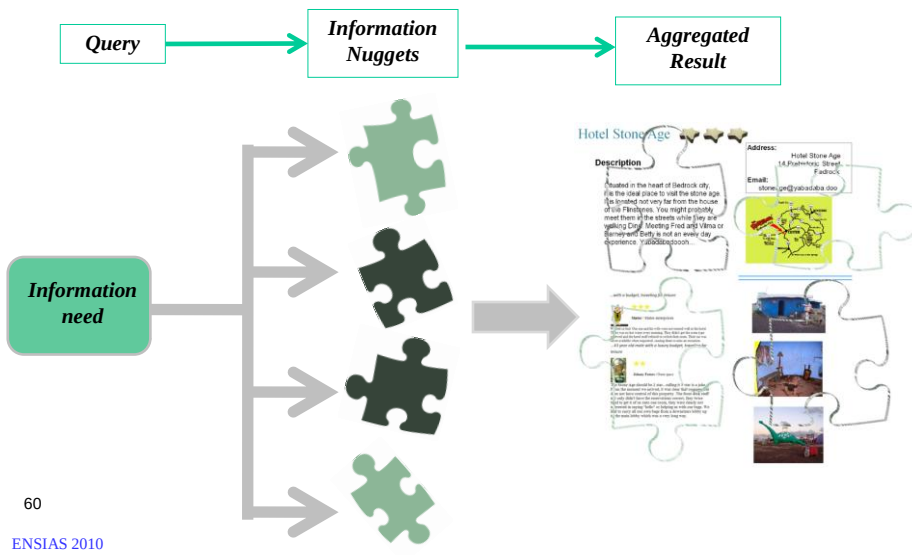
Utiliser les différentes les possibilités d'interrogation

- Recherche d'expression ("…" double quotes)
- Opérateurs booléens : AND, OR, NOT
- Inclusion terme (+), exclusion terme (-)
- Connecteurs With , Near, ADJ
- Recherche sur des champs spécifiques
  - Title: "information retrieval"
  - url, filetype
  - Title: "veille économique" filetype:pdf

Conclusion : Nos travaux actuels

- Recherche d'information personnalisée/contextuelle
  - Prendre en compte le contexte de l'utilisateur (ses centres d'intérêts, sa tâche, sa géo-localisation, temporelle)
- Détection d'opinions
  - Identifier les pages de type opinion puis classer les opinions (+) (-)
- Recherche d'information sociale
  - Prendre en compte annotation/jugements, distance sociale approbation/désapprobation des autres membres
- Recherche agrégée
  - Construire la réponse à la question en s'appuyant sur plusieurs sources
- Recherche dans les document XML

Recherche agrégée



Merci

Open-source

- Smart (Cornell)
- MG (RMIT & Melbourne, Australia; Waikato, New Zealand),
- Lemur (CMU/Univ. of Massachusetts)
- Lucene (Nutch)
- Terrier (Univ Glasgow)