

---

# 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

ENSIAS 2010

1

## 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

ENSIAS 2010

2

## 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

---



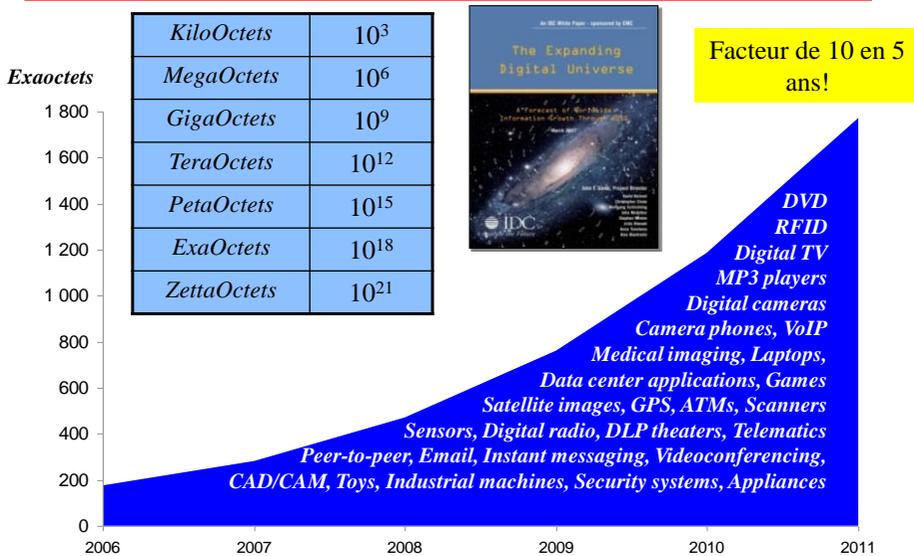
## Volume sans cesse croissant

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

ENSIAS 2010

5

## .. en perpétuelle croissance



ENSIAS 2010

Source: IDC, 2008

6

## .. produite par tout le monde

---



## .. L'information (numérique) est disponible partout

ENSIAS 2010

7

## ... 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*)

ENSIAS 2010

8

## ... 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

ENSIAS 2010

## Contours de la RI

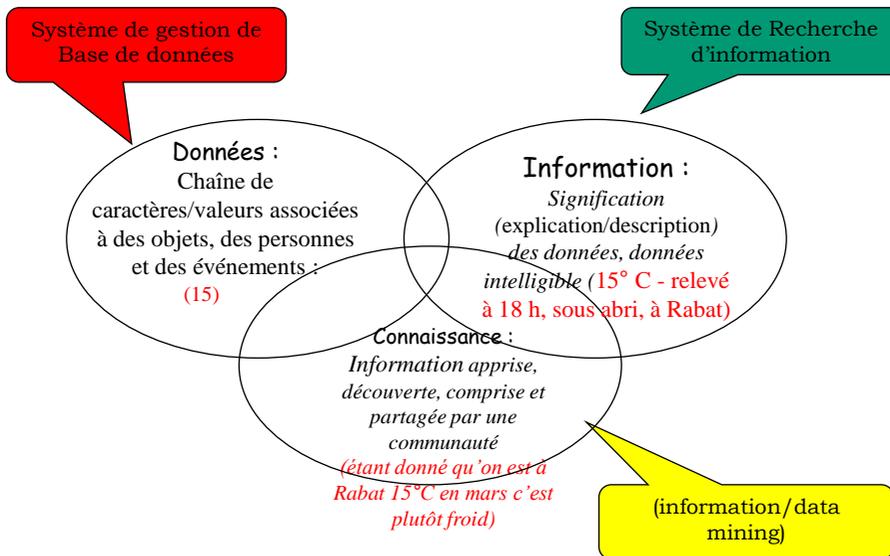
---

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

ENSIAS 2010

10

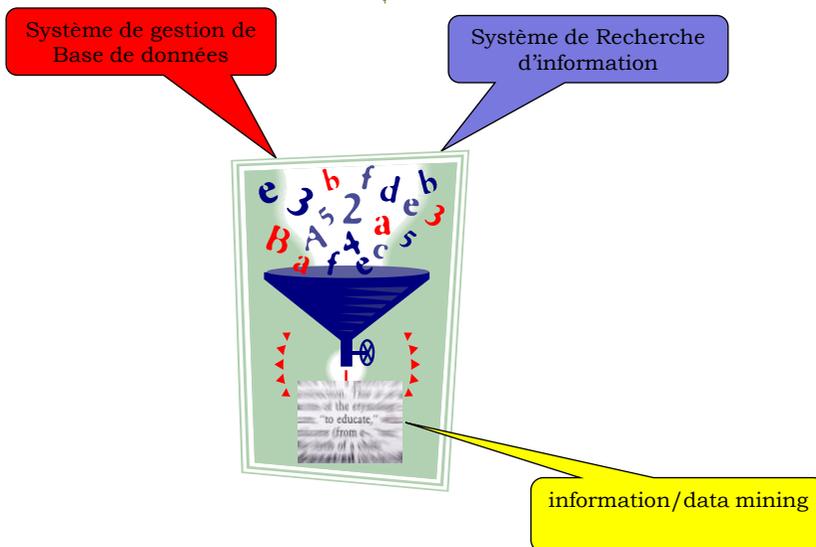
## Contours: Données-Information-Connaissance



ENSIAS 2010

11

## Veille a besoin des trois domaines



ENSIAS 2010

12

## 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 ? »

ENSIAS 2010

13

The screenshot shows the WolframAlpha interface in a Mozilla Firefox browser. The search query is 'averroes'. The results are interpreted as 'Averroès (philosophe)'. The page displays basic information in a table, a timeline, and a list of 'A few things to try'.

Basic information:	
full name	Abu al-Walid Muhammad
date of birth	1126 AD (884 years ago)
place of birth	Cordoba, Spain
date of death	1198 AD (age: 72 years) (812 years ago)
place of death	Marrakech, Marrakech-Tensift-Al Haouz, Morocco

Timeline: Averroès

Computed by: [Wolfram Mathematica](#) | [Source information »](#) | [Download as: PDF](#) | [Live Mathematica](#)

PageRank: | Alexa Rank: 4,536 | PR: 13

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

---

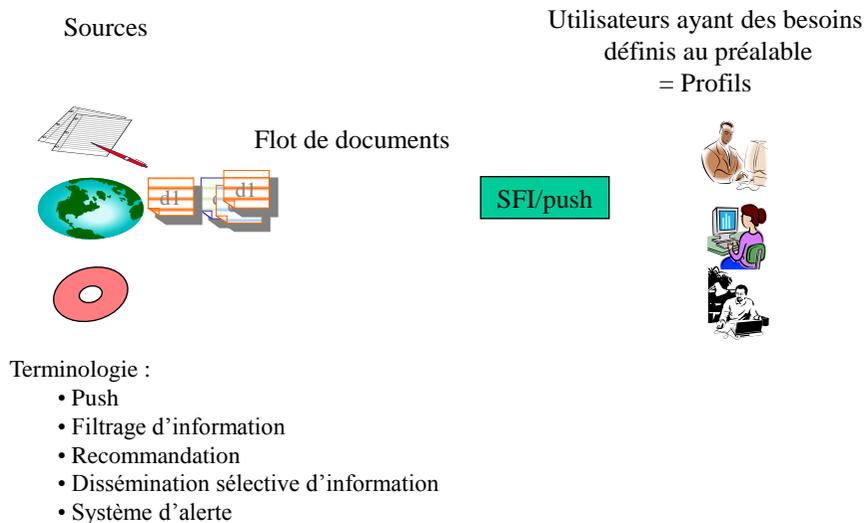
- Filtrage d'information (*filtering/recommendation*)
  - Dissémination sélective d'information

ENSIAS 2010

15

## Filtrage d'information

---



ENSIAS 2010

16



## 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

---

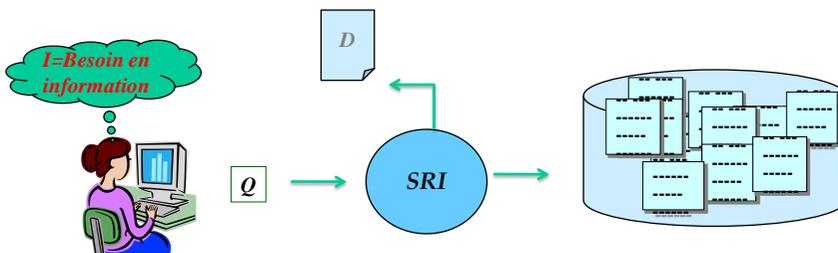
- Comprendre ce qu'est la Recherche d'information (RI)
  - Intérêt et contours
  - Fonctionnement «interne» d'un système de RI
- Cas d'étude : Recherche d'information sur le Web
  - Bonnes pratiques en tant que producteur d'information
  - Bonnes pratiques en tant que consommateur de l'information

ENSIAS 2010

19

## Comprendre le fonctionnement d'un SRI

---

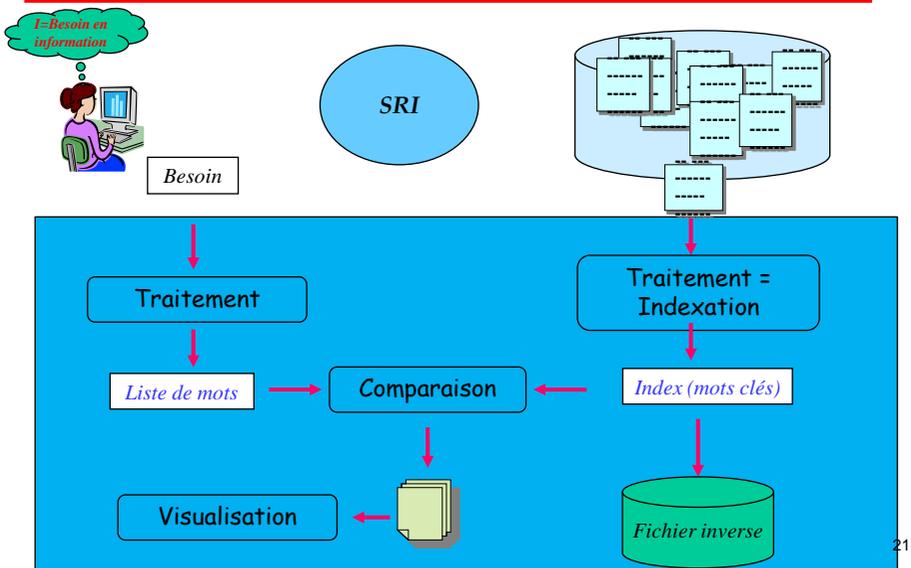


- Sélectionner dans une collection
  - les **informations**
  - ... **pertinentes** répondant aux
  - ... **besoins en information des utilisateurs**

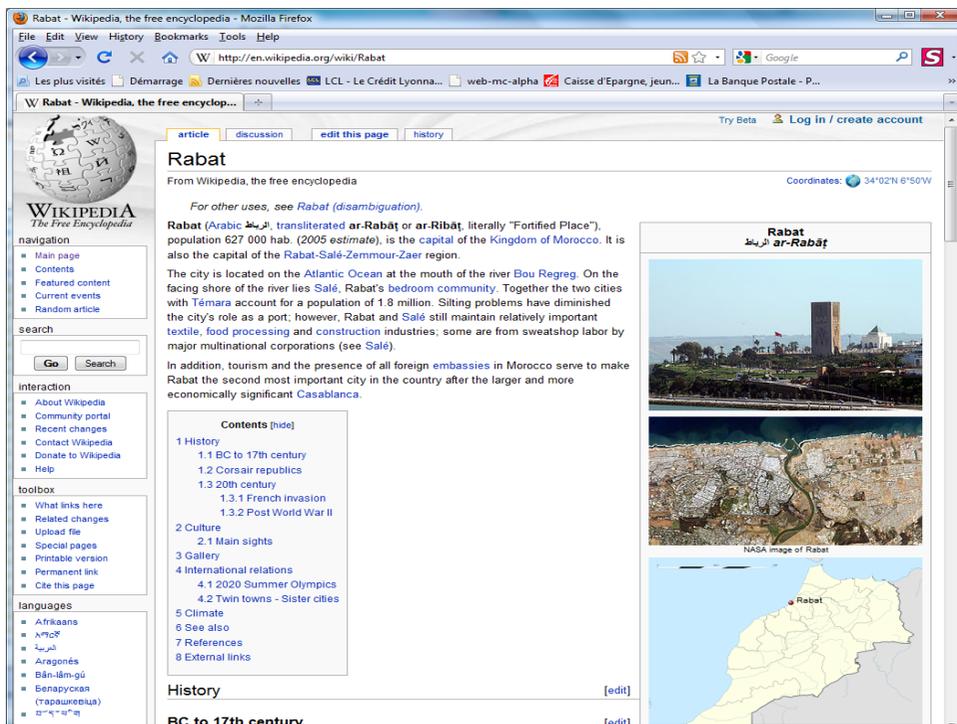
ENSIAS 2010

20

# Fonctionnement d'1 SRI : Ouvrir la boîte noire



ENSIAS 2010



```

Source of http://en.wikipedia.org/wiki/Rabat - Mozilla Firefox
File Edit View Help
<!DOCTYPE html PUBLIC "-//W3C/DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en" dir="ltr">
<head>
<title>Rabat - Wikipedia, the free encyclopedia</title>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
<meta http-equiv="Content-Style-Type" content="text/css" />
<meta name="generator" content="MediaWiki 1.16alpha-vmf" />
<link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?title=Rabat&action=edit" />
<link rel="edit" title="Edit this page" href="/w/index.php?title=Rabat&action=edit" />
<link rel="apple-touch-icon" href="http://en.wikipedia.org/apple-touch-icon.png" />
<link rel="shortcut icon" href="/favicon.ico" />
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia (en)" />
<link rel="copyright" href="http://creativecommons.org/licenses/by-sa/3.0/" />
<link rel="alternate" type="application/rss+xml" title="Wikipedia RSS Feed" href="/w/index.php?title=Special:RecentChanges&feed=rss" />
<link rel="alternate" type="application/atom+xml" title="Wikipedia Atom Feed" href="/w/index.php?title=Special:RecentChanges&feed=atom" />
<link rel="stylesheet" href="http://bits.wikimedia.org/skins-1.5/common/shared.css?257z23" type="text/css" media="screen" />
<link rel="stylesheet" href="http://bits.wikimedia.org/skins-1.5/common/commonPrint.css?257z23" type="text/css" media="print" />
<link rel="stylesheet" href="http://bits.wikimedia.org/skins-1.5/monobook/main.css?257z23" type="text/css" media="screen" />
<link rel="stylesheet" href="http://bits.wikimedia.org/skins-1.5/chick/main.css?257z23" type="text/css" media="handheld" />
<!--[if lt IE 5.5000]><link rel="stylesheet" href="http://bits.wikimedia.org/skins-1.5/monobook/IE50Fixes.css?257z23" type="text/css" medi
<!--[if IE 5.5000]><link rel="stylesheet" href="http://bits.wikimedia.org/skins-1.5/monobook/IE55Fixes.css?257z23" type="text/css" media="
<!--[if IE 6]><link rel="stylesheet" href="http://bits.wikimedia.org/skins-1.5/monobook/IE60Fixes.css?257z23" type="text/css" media="scree
<!--[if IE 7]><link rel="stylesheet" href="http://bits.wikimedia.org/skins-1.5/monobook/IE70Fixes.css?257z23" type="text/css" media="scree
<link rel="stylesheet" href="/w/index.php?title=MediaWiki:Common.css&usemsscach=yes&ctype=text&2Fc&maxage=2678400&actio
<link rel="stylesheet" href="/w/index.php?title=MediaWiki:Print.css&usemsscach=yes&ctype=text&2Fc&maxage=2678400&actio
<link rel="stylesheet" href="/w/index.php?title=MediaWiki:Handheld.css&usemsscach=yes&ctype=text&2Fc&maxage=2678400&ac
<link rel="stylesheet" href="/w/index.php?title=MediaWiki:Monobook.css&usemsscach=yes&ctype=text&2Fc&maxage=2678400&ac
<link rel="stylesheet" href="/w/index.php?title=-&action=raw&maxage=2678400&qen=css" type="text/css" media="all" />
<script type="text/javascript">
var skin="monobook",
stylepath="http://bits.wikimedia.org/skins-1.5",
wgUrlProtocols="http\\:\\\\|https\\:\\\\|ftp\\:\\\\|irc\\:\\\\|gopher\\:\\\\|telnet\\:\\\\|nntp\\:\\\\|worldwind\\:\\\\|ma
wgArticlePath="/wiki/$1",
wgScriptPath="/w",
wgScriptExtension=".php",
wgScript="/w/index.php",
wgVariantArticlePath=false,
wgActionPaths={},
wgServer="http://en.wikipedia.org",
wgCanonicalNamespace="",
wgCanonicalSpecialPageName=false,
wgNamespaceNumber=0,
wgPageName="Rabat",

```

## 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

*Titre: 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	Doc #	Freq
ambitious	1	1	1	1	1
be	1	1	2	2	1
brutus	2	2	3	1	1
capitol	1	1	5	1	1
caesar	2	3	6	1	1
did	1	1		2	2
enact	1	1		1	1
hath	1	1		1	1
i	1	2		2	1
i'	1	1		1	2
it	1	1		1	1
julius	1	1		2	1
killed	1	2		1	1
let	1	1		1	2
me	1	1		2	1
noble	1	1		1	1
so	1	1		2	1
the	2	2		2	1
told	1	1		1	1
you	1	1		2	1
was	2	2		2	1
with	1	1		2	1
				1	1
				2	1
				2	1

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.

d3:  
I did enact Julius  
Caesar I was killed  
i' the Capitol.  
Brutus killed me.

d4:  
I did enact Julius  
Caesar I was killed  
i' the Capitol.  
Brutus killed me.

d5:  
I did enact Julius  
Caesar I was killed  
i' the Capitol.  
Brutus killed me.

d6:  
I did enact Julius  
Caesar I was killed  
i' the Capitol.  
Brutus killed me.

ENSIAS 2010

## Répondre à une demande (requête)



caesar

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

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.

d3:  
I did enact Julius  
Caesar I was killed  
i' the Capitol.  
Brutus killed me.

d4:  
I did enact Julius  
Caesar I was killed  
i' the Capitol.  
Brutus killed me.

d5:  
I did enact Julius  
Caesar I was killed  
i' the Capitol.  
Brutus killed me.

d6:  
I did enact Julius  
Caesar I was killed  
i' the Capitol.  
Brutus killed me.

ENSIAS 2010

## 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 document (*tf*), sa fréquence dans la collection (*idf*), sa position dans le texte(*p*), taille du document (*dl*) ...

$$Score(D) = \text{fonction}(tf, idf, dl)$$

ENSIAS 2010<sup>27</sup>

## 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

ENSIAS 2010

28

# 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<sup>12</sup> octets)
  - Google : Dizaine de milliards de Pages
- Millions d'utilisateurs, efficacité des accès

ENSIAS 2010

29

## Exemple de résultats d'un moteur de recherche

The image shows a screenshot of a Google search results page for the query "ecole informatique". The browser window title is "ecole informatique - Recherche Google - Mozilla Firefox". The search bar contains "ecole informatique" and the search button is labeled "Rechercher". The results show a list of links to various computer science schools and training centers. Two green boxes with arrows point to specific results: one labeled "Normale" points to the first result "Formation Informatique", and another labeled "Pub" points to the advertisement "Ecole d'ingénieur ESISAR".

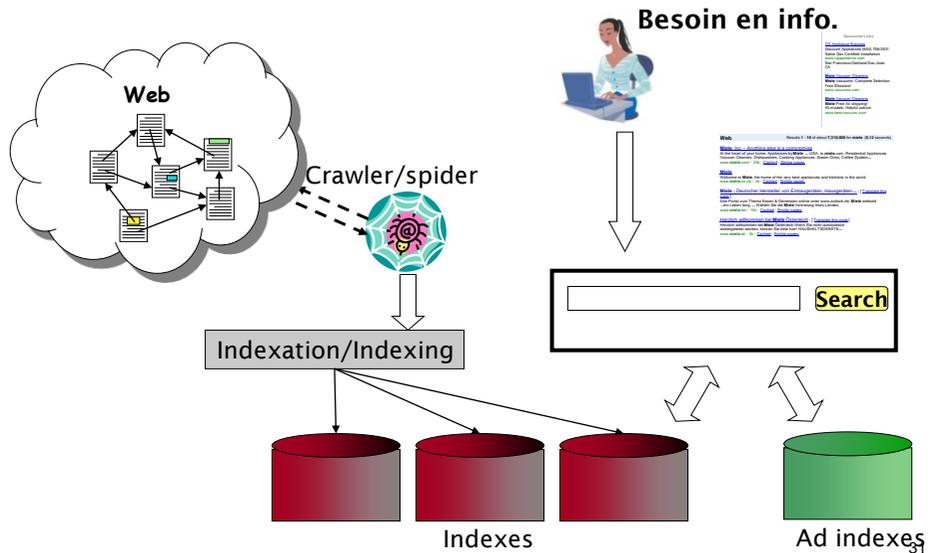
Annotations:

- "Normale" points to the first search result: [Formation Informatique](#) (informatics.demos.fr)
- "Pub" points to the advertisement: [Ecole d'ingénieur ESISAR](#)

ENSIAS 2010

30

## Approche de base de la recherche sur le Web



ENSIAS 2010

## 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)

ENSIAS 2010

32

## 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://en.wikipedia.org/wiki/Search_engine_optimization)
    - [http://fr.wikipedia.org/wiki/Optimisation\\_pour\\_les\\_moteurs\\_de\\_recherche](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

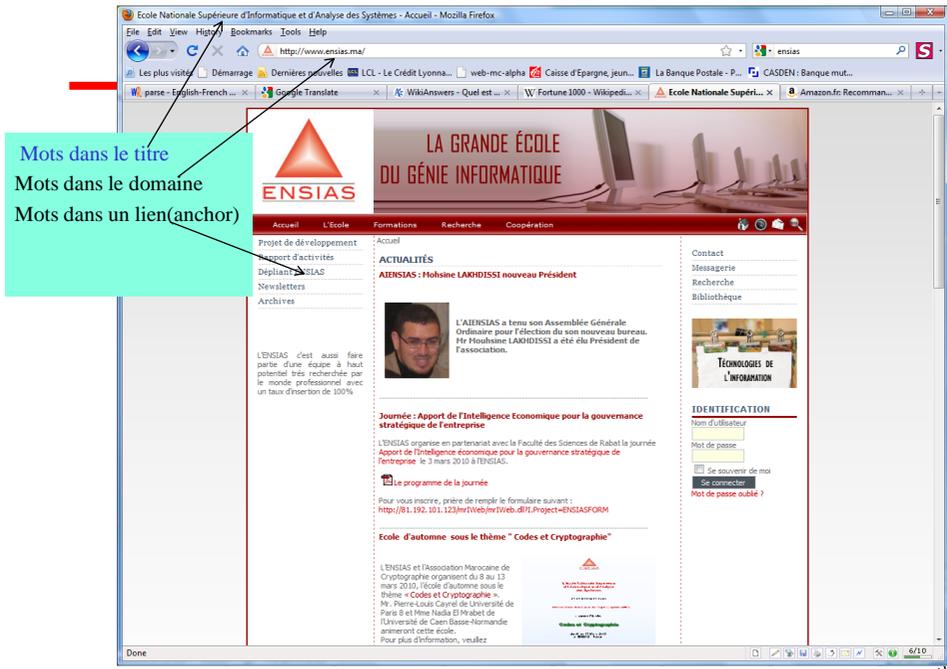
---

- 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 »

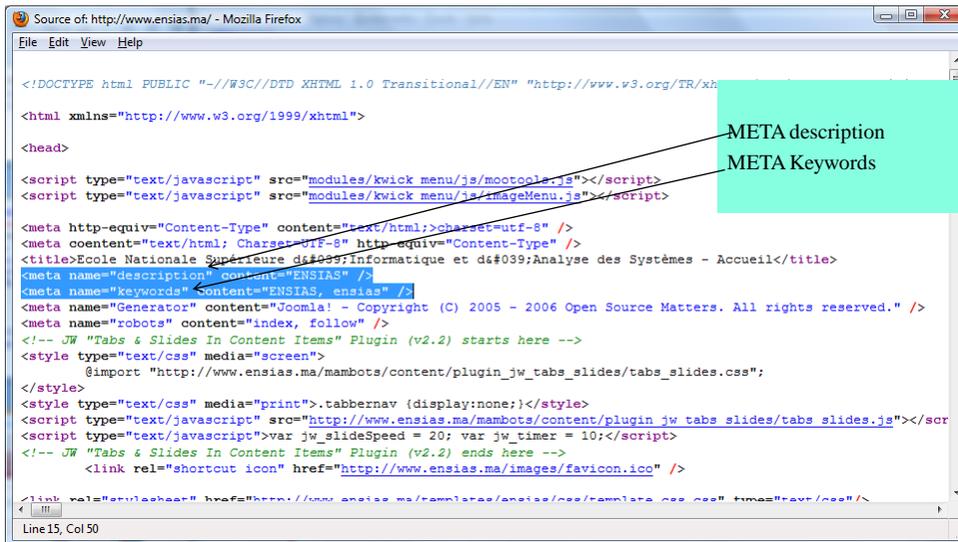
## Facteurs utilisés par les moteurs pour trier des pages (1 / 5)

---

- 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...



ENSIAS 2010

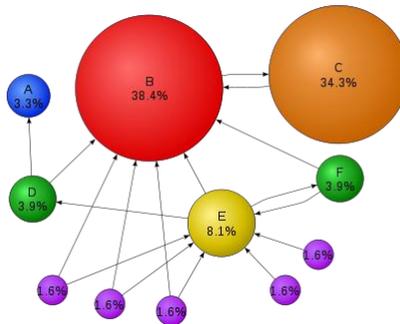


ENSIAS 2010

## 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*)
    - $PR(p) = f(\text{nombre de sites pointant } p, \text{ popularité de ces sites, } \dots)$



© *wikipédia*

ENSIAS 2010

39



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

ENSIAS 2010

40

## 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

Comment souhaitez-vous générer des idées de mots clés ?

Expressions ou termes descriptifs  
(exemple : thé vert)

Contenu de site Web  
(exemple : www.exemple.fr/produit?id=74893)

Entrez un mot clé ou une expression par ligne :  
veille économique

Utiliser des synonymes

[Filtrer mes résultats](#)

Sélectionnez les colonnes à afficher :  
Afficher/masquer les colonnes

Mots clés	Concurrence entre annonceurs	Volume de recherche locale : janvier	Volume de recherche mensuel global	Type de ciblage
<b>Mots clés en rapport avec le(s) terme(s) entré(s) - trié par pertinence</b>				
veille économique	<input type="checkbox"/>	2 400	2 900	<a href="#">Ajouter</a>
veille économique définition	<input type="checkbox"/>	Données insuffisantes	58	<a href="#">Ajouter</a>
veille intelligence économique	<input type="checkbox"/>	Données insuffisantes	1 300	<a href="#">Ajouter</a>
<a href="#">Tout ajouter - 3 »</a>				
Télécharger tous les mots clés : <a href="#">texte</a> , <a href="#">csv (pour Excel)</a> , <a href="#">csv</a>				
<b>Mots clés supplémentaires à envisager - trié par pertinence</b>				
veille économique	<input type="checkbox"/>	1 600	1 900	<a href="#">Ajouter</a>
veille stratégique	<input type="checkbox"/>	2 900	4 400	<a href="#">Ajouter</a>
veille stratégique	<input type="checkbox"/>	8 100	8 100	<a href="#">Ajouter</a>
veille concurrentielle	<input type="checkbox"/>	8 100	9 900	<a href="#">Ajouter</a>
intelligence économique	<input type="checkbox"/>	27 100	27 100	<a href="#">Ajouter</a>
veille technologique	<input type="checkbox"/>	12 100	14 800	<a href="#">Ajouter</a>
veille	<input type="checkbox"/>	Données	2 400	<a href="#">Ajouter</a>

## Bonnes pratiques : consommateur de l'info

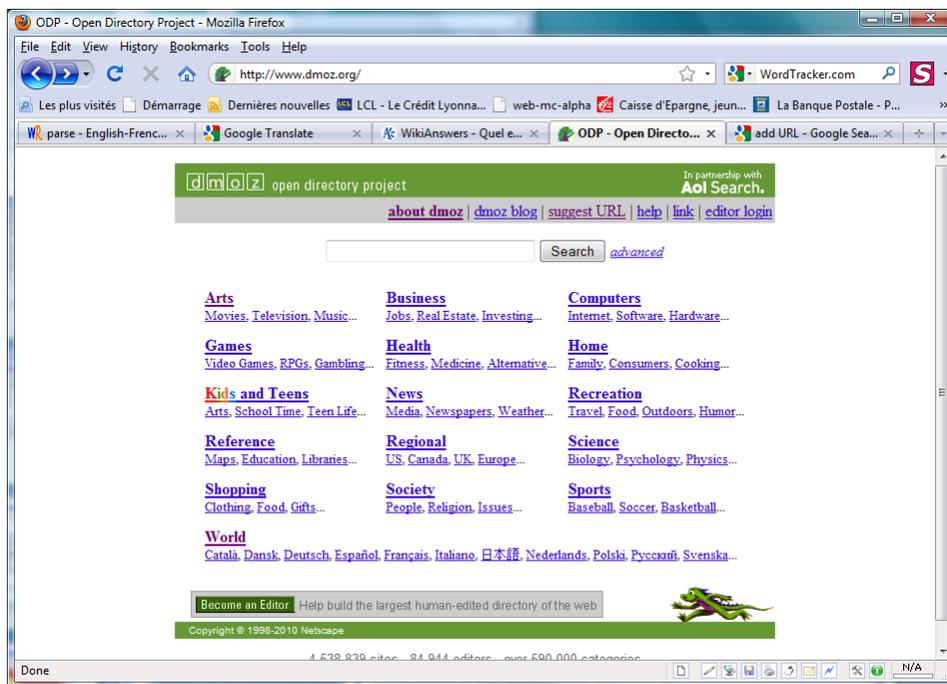
- Quel est le meilleur moteur de recherche ?
  - Il ya plus de 300 moteurs de recherche disponibles pour une utilisation libre. Chacun possède ses propres atouts
- Choisir son moteur selon ses besoins et sa tâche
- Utiliser les différentes possibilités du moteur (en particulier en terme d'interrogation)

## Choisir son moteur

- Mode d'accès
  - Mots (clés) : Google, Bing, Exalead, ...
  - Annuaire : 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!!)

ENSIAS 2010

47

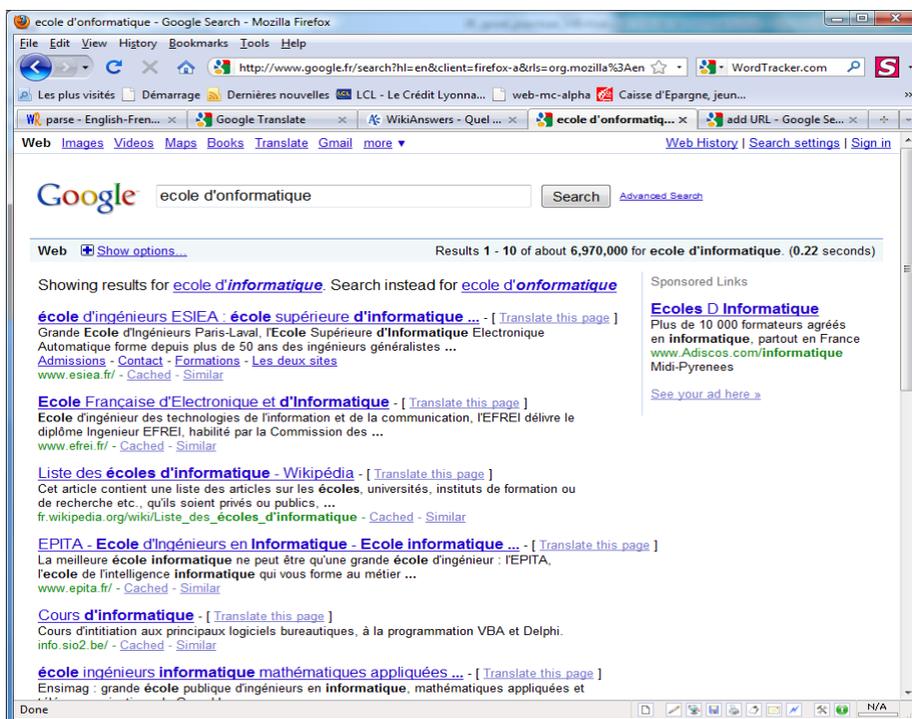


# Choisir son moteur

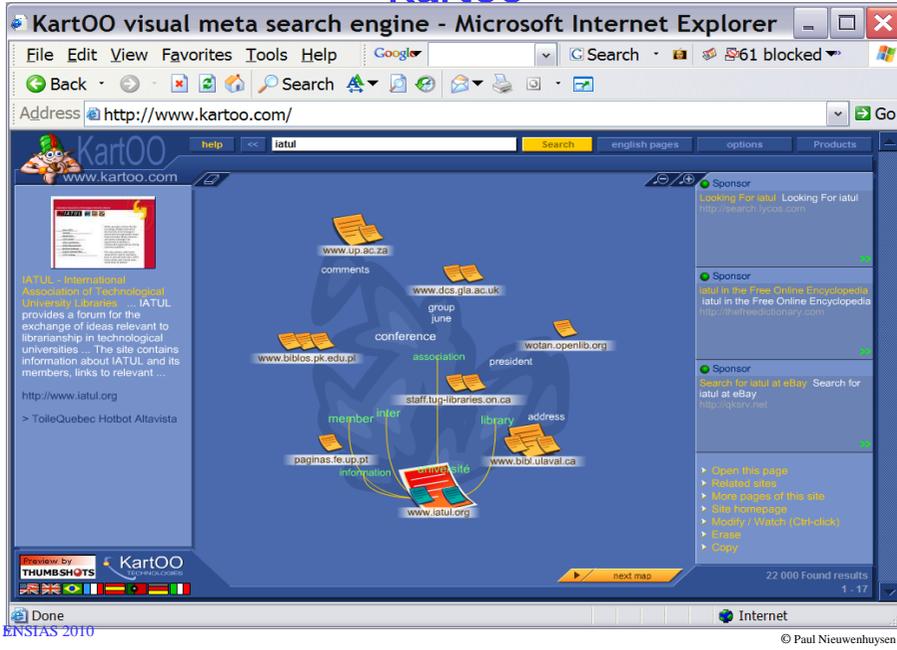
- Mode d'accès
  - Mots (clés) : Google, Bing, Exalead, ...
  - Annuaire : Yahoo, DMOZ, About, ...
- Mode de visualisation/présentation
  - Liste linéaire "à la google"
  - Classification (Clusty/ search-cube)
- Type de moteurs
  - Généralistes vs. Verticaux (type de médias/thématiques)
- Autres facteurs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

ENSIAS 2010

49



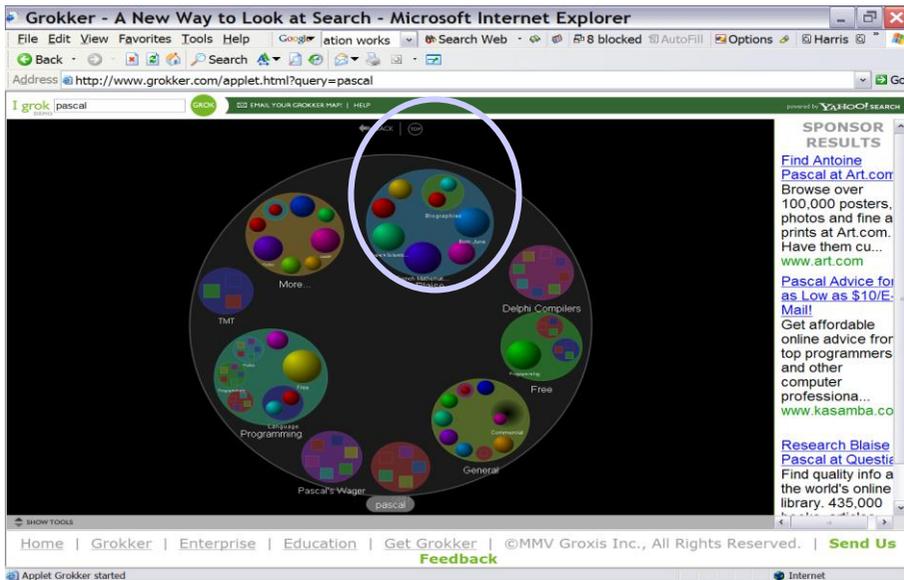
# Kartoo



INSTITAS 2010

© Paul Nieuwenhuysen

# Grokker



52

© Paul Nieuwenhuysen

## Choisir son moteur

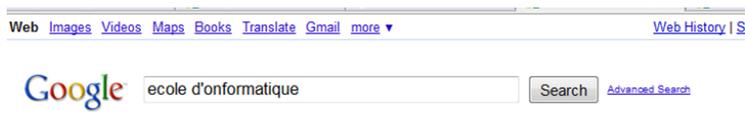
---

- 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](http://www.searchFinance.com))
  - Technologie ([www.knowledgestorm.com](http://www.knowledgestorm.com))
  - Droit (Legal Search Engines)
  - Science : **Scirus** , CiteSeerX ([citeseerx.ist.psu.edu](http://citeseerx.ist.psu.edu)) pour l'informatique et IT
  - *Medical Search Engines (Medhunt* ([www.hon.ch/MedHunt](http://www.hon.ch/MedHunt)))
  - *Travel Search Engines*
  - *Chemfinder* ([www.chemfinder.com](http://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

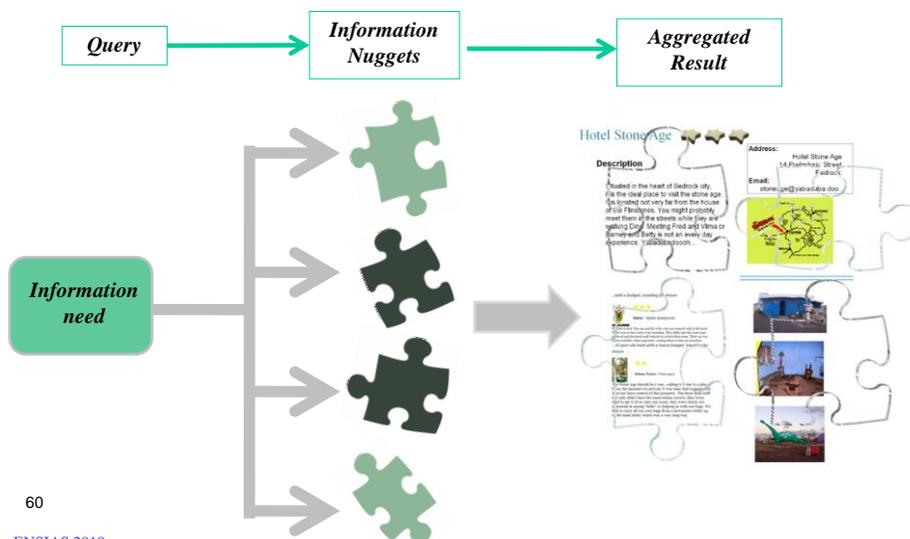
## Conclusion : Nos travaux actuels

- 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

ENSIAS 2010

59

## Recherche agrégée



60

ENSIAS 2010

---

# Merci

## Open-source

---

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