

Anatomie d'un moteur de recherche

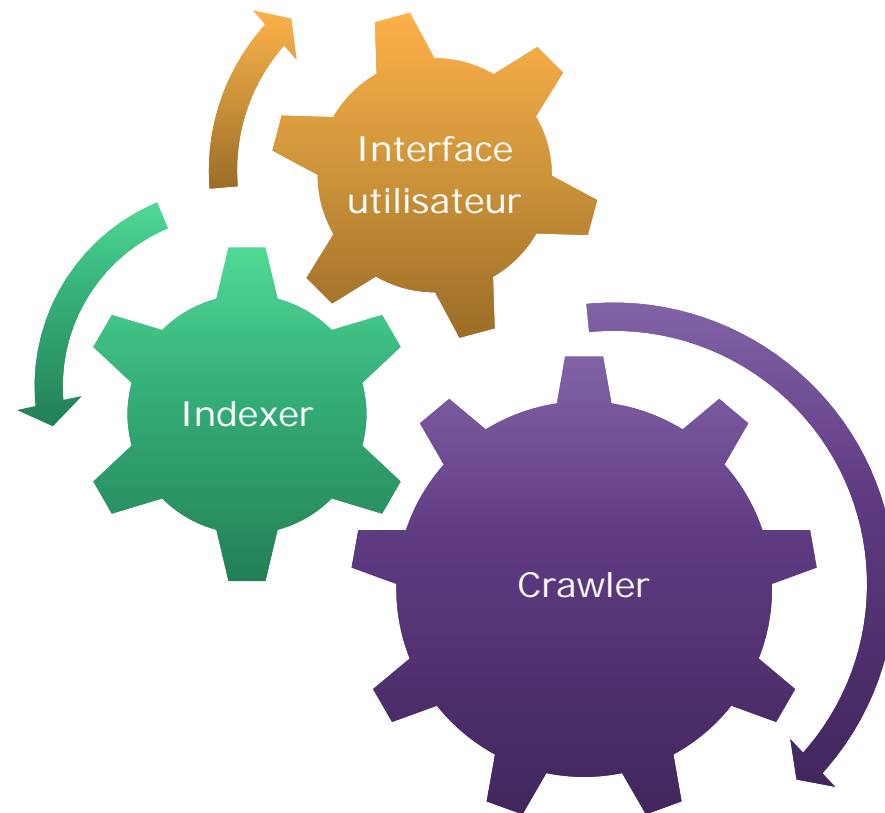
SEO Camp'us -4 et 5 février 2009

Philippe YONNET

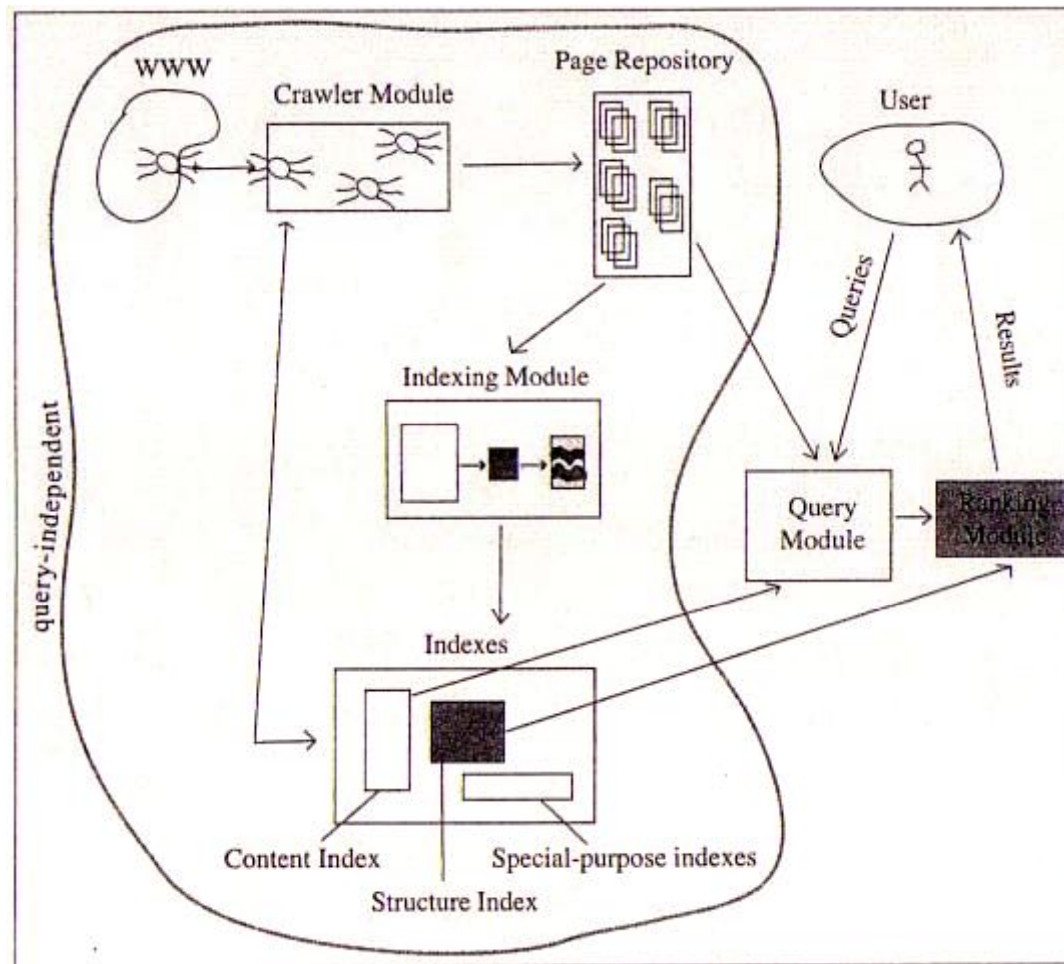
Directeur du pôle métiers – Aposition
Président de l'association SEOCamp



Les grands composants d'un moteur



L'architecture classique d'un moteur



Crawler le web

Combien de pages web ?

- Impossible à savoir précisément (voir plus loin)
- Google déclare connaître 1000 milliards d'urls différentes

Combien de pages utiles sur le web

- 120 / 130 milliards

Combien de pages dans les moteurs ?

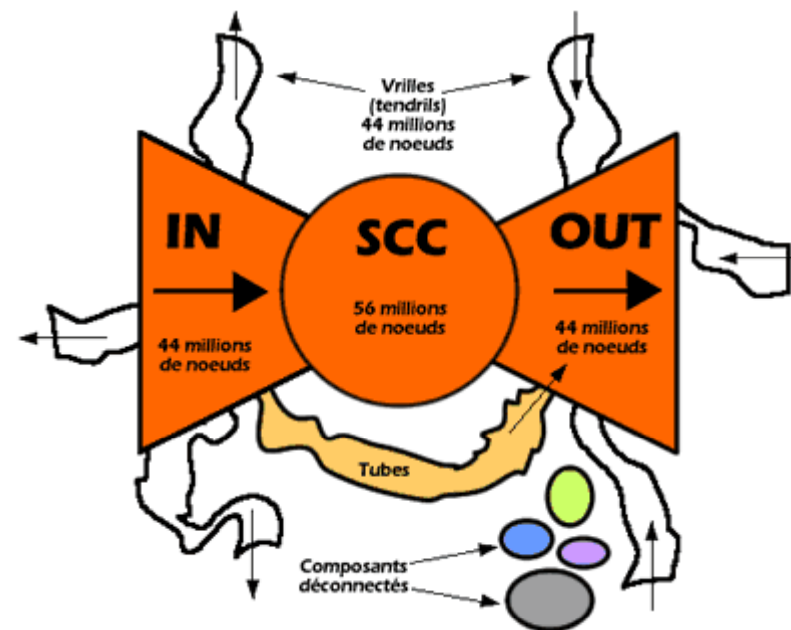
- Quelques dizaines de milliards
- Pas toutes placées dans le même type d'index
 - Index primaire / index secondaire
 - La révolution « Teragoogle » (alias Bigdaddy)

Crawler le web : un exercice difficile



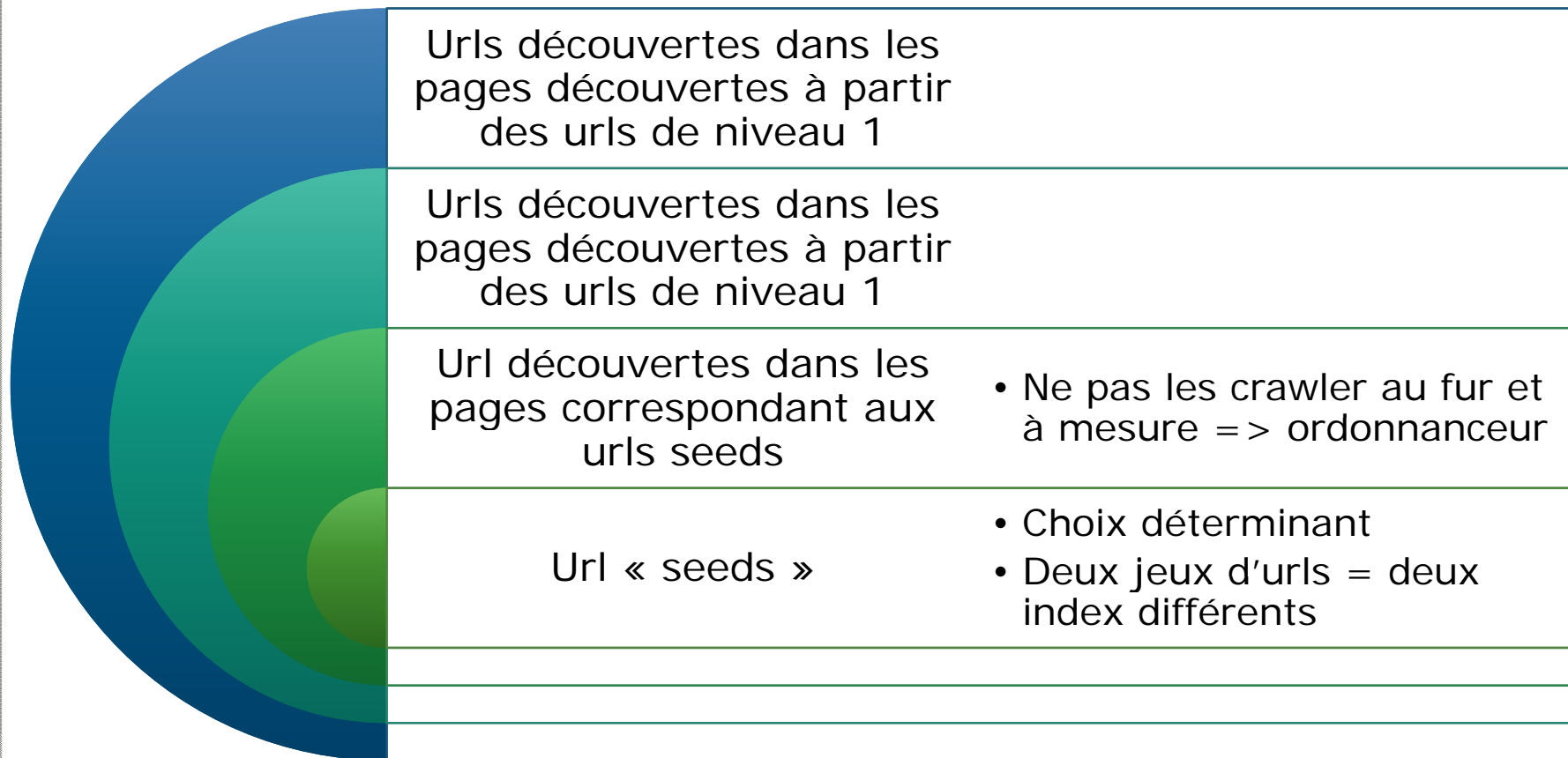
Kleinberg

La topologie du web est particulière

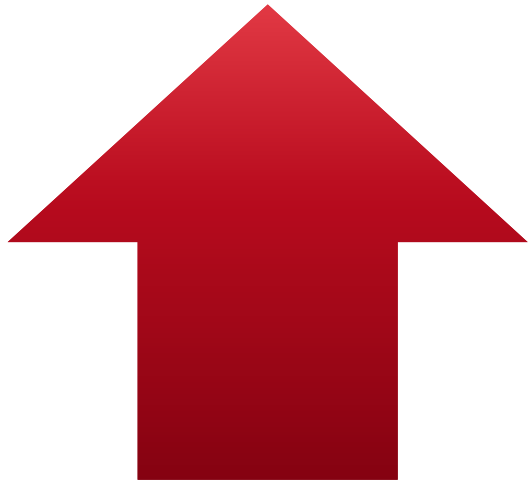


Jon Kleinberg (labo IBM Almaden) : « the web as a graph »

L'importance des urls seeds (urls "semence")



Evolution des techniques de crawl



Crawl fermé

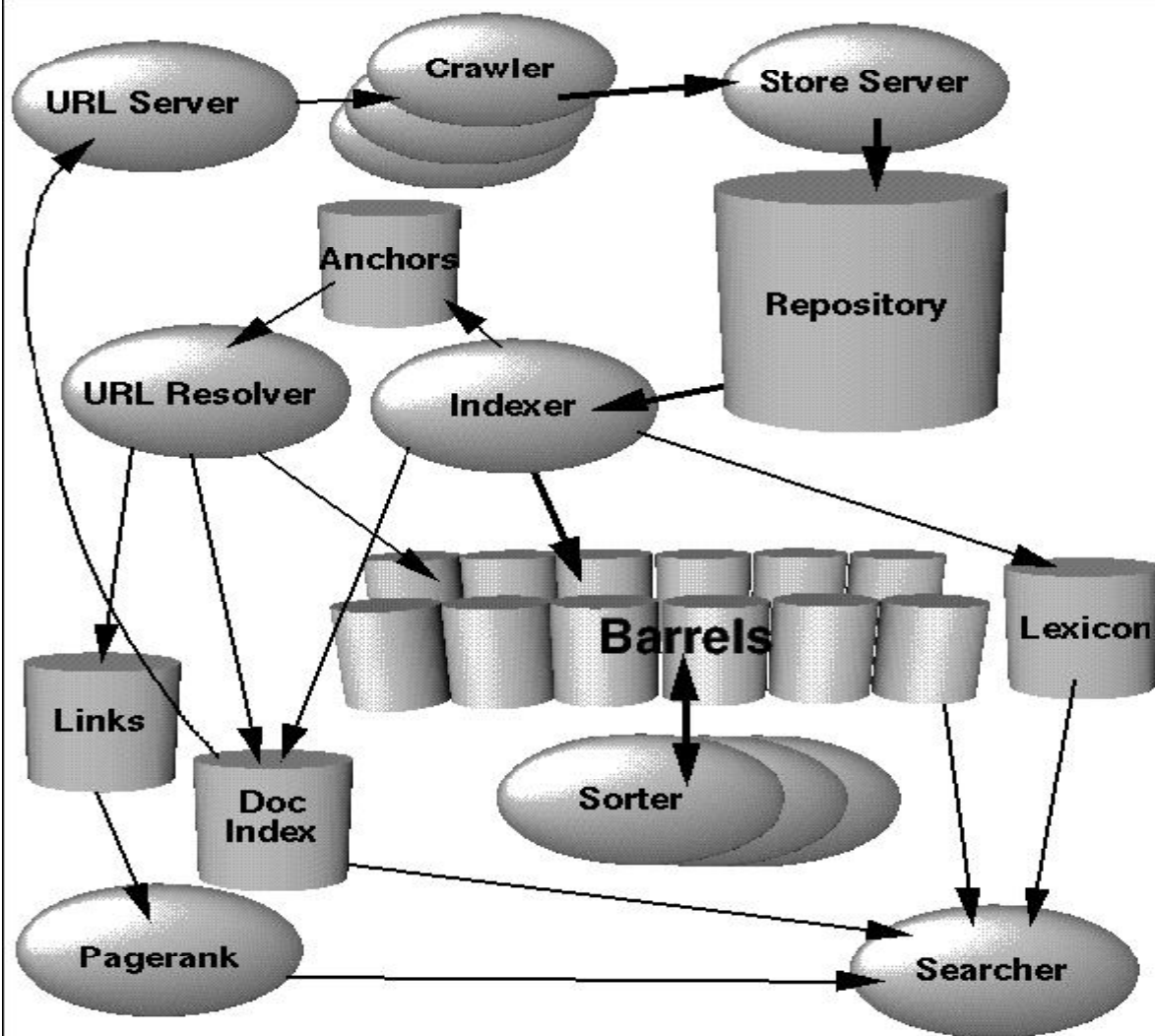
- Ancien système de Google
- On part du principe que l'on va crawler un nombre déterminé d'url (celles découvertes au crawl précédent + un niveau de nouvelles par exemple)
- Cycle de 28 jours
 - 1 : deep crawl
 - 2 : indexation + calcul du PR
 - 3 : mise à jour des datacenters



Crawl ouvert

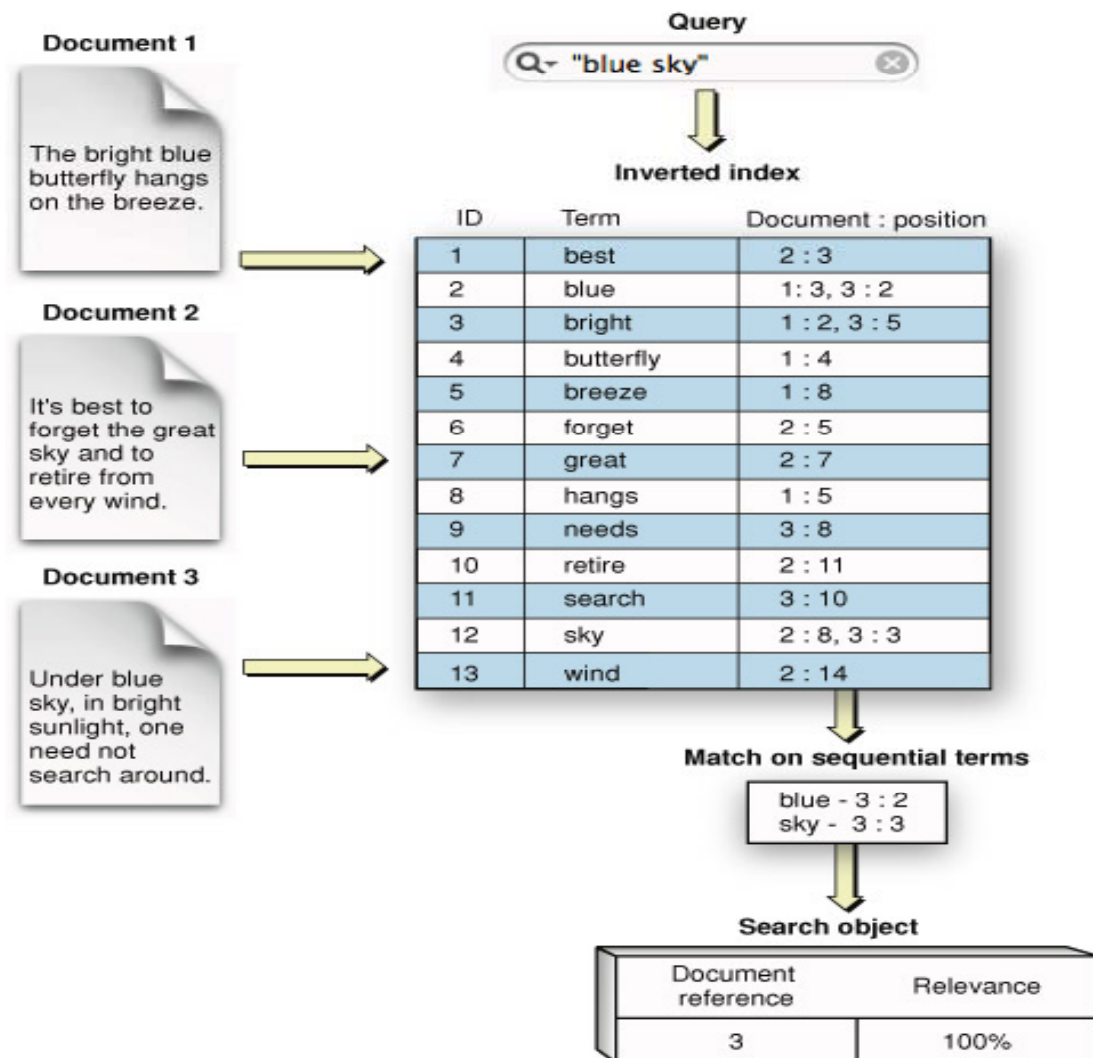
- Si on découvre des urls, on les crawle
- On recrawle les urls connus à différents intervalles de temps
- On ne cherche pas à crawler tout le web, et on considère que le crawl n'est jamais fini

Architecture de Google



L'architecture de Google selon l'article fondateur de Page et Brin

L'indexation : page + position



L'interface utilisateur

The Google logo is centered on the page. It consists of the word "Google" in its signature multi-colored font: 'G' is blue, 'o' is red, 'o' is yellow, 'g' is blue, 'l' is green, and 'e' is red. A small trademark symbol (TM) is located at the top right of the 'e'.

Web [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

Le snippet

Sign in

Web Images Video News Maps Desktop more »

Google

blake Field Search Advanced Search Preferences

Web Results 101 - 110 of about 1,820,000 for **blake Field**. (0.07 seconds)

Title
Blake completes Masters field TENNIS SPORT tvnz.co.nz ASB Classic & Heineken Open. Blake completes Masters field ... American James Blake on Sunday secured the eighth and last spot at the season-ending Masters ... tvnz.co.nz/view/page/488120/879943 - 61k - Cached - Similar pages
New Low-Complexity Bit-Parallel Finite Field Multipliers Using ... 21 Ian F. Blake , Shuhong Gao , Robert Lambert, Constructive problems for irreducible polynomials over finite fields , Proceedings of the third Canadian ... portal.acm.org/citation.cfm?coll=GUIDE&dl=GUIDE&id=305091 - Similar pages
Contributors and Sources Blake Ferris lives in New York where he experiments with exotic ... Force Field www.wondermagnet.com is a fun site with all kinds of silly magnet stuff, ... www.theodoregray.com/PeriodicTable/Contributors.html - 316k - Cached - Similar pages
Talisman Announces First Oil From Blake and Production Restart ... wholly owned subsidiary, Talisman Energy (UK) Limited, has resumed production in the Ross oil field and has commenced production in the Blake oil field over ... www.encyclopedia.com/doc/1G1-75993928.html - 32k - Cached - Similar pages
How Not to be a Racist by Blake A. Field By: Blake A. Field . Copyright 2003 Blake A. Field . Of course, I have always consistently denied that I am a racist. Repeatedly, I made the case that I love ... www.blakeswritings.com/Racist.html - Similar pages
FOR THE LOTTO by Blake A. Field by: Blake A. Field . Copyright 2001 Blake A. Field . Back when, back before the State of Indiana decided that gambling was OK ... www.blakeswritings.com/ForTheLotto.htm - Similar pages

Size

Cache

Similar

URL

Un exemple d'évolution de l'Interface L'expansion de requête à base de thésaurus

clio - Recherche Google - Mozilla Firefox

http://www.google.fr/search?hl=fr&q=clio&btnG=Rechercher&meta=

Rechercher [Recherche avancée](#) [Préférences](#)

Rechercher dans : Web Pages francophones Pages : France

Résultats 1 - 10 sur un total d'environ 29 800 000 pour **clio** (0,07 secondes)

Voyages culturels et historiques de Clío
Clío, la muse de l'Histoire, nous a prêté son nom pour vous offrir un vaste éventail de plaisirs culturels : voyages, conférences, visites et articles de ...
[www.clio.fr/](#) - 12k - [En cache](#) - [Pages similaires](#)

[des voyages](#) [Espace conférenciers](#)
[Les voyages de Clío](#) [articles](#)
[Conférences](#) [Espace Voyageurs](#)
[Toutes nos croisières](#) [Recevoir nos catalogues](#)

[Autres résultats, domaine clio.fr »](#)

CLIO - Actualité du conte et des conteurs
conte, conteur, toutes les actualités sur la littérature orale. Le Conservatoire contemporain de Littérature Orale a fêté en 2006 ses 25 ans d'existence.
[www.clio.org/](#) - 23k - [En cache](#) - [Pages similaires](#)

Clío
Revue scientifique spécialisée dans l'étude de l'histoire des femmes.
[clio.revues.org/](#) - 77k - [En cache](#) - [Pages similaires](#)

CLIO HFS en quelques mots Sommaire La revue
informations sur la Lettre de Clío. Lettre de Revues.org - Fédération de revues en sciences humaines et sociales - Syndication de contenu au format RSS 0.92 ...
[clio.revues.org/index800.html](#) - 36k - [En cache](#) - [Pages similaires](#)
de L. Capdevila - 2007 - [Les 2 versions](#)

Essayez ceci : [renault](#)

Renault France - constructeur automobile
Renault constructeur automobile français, vous présente sa gamme de véhicules, voitures neuves et d'occasion, ses financements et ses services automobile.
[www.renault.fr/](#)

Renault.com - Site officiel international du groupe Renault ...
Bienvenue sur le site **renault.com**. Retrouvez toute l'actualité du groupe **Renault**, ses gammes de véhicules particuliers et utilitaires, son histoire, ...
[www.renault.com/renault_com/ft/main/index.aspx](#)

Site officiel Renault Sport - Courses automobiles, voitures de ...
Courses automobiles : Clío Cup, Formule **Renault**, World Series by **Renault**, Mégane Trophy, Clío Super 1600, Clío Cup Rally. Voitures de sport : Clío **Renault** ...
[www.renault-sport.com/](#)

Terminé

démarrer 3 Microsoft ... 8 Microsoft ... Microsoft Exce... FORMATIONS formation [Réc... Sans titre - Paint clío - Recherch... Téléchargements FR 18:16

Recherche
sur clio
Renvoi sur
renault