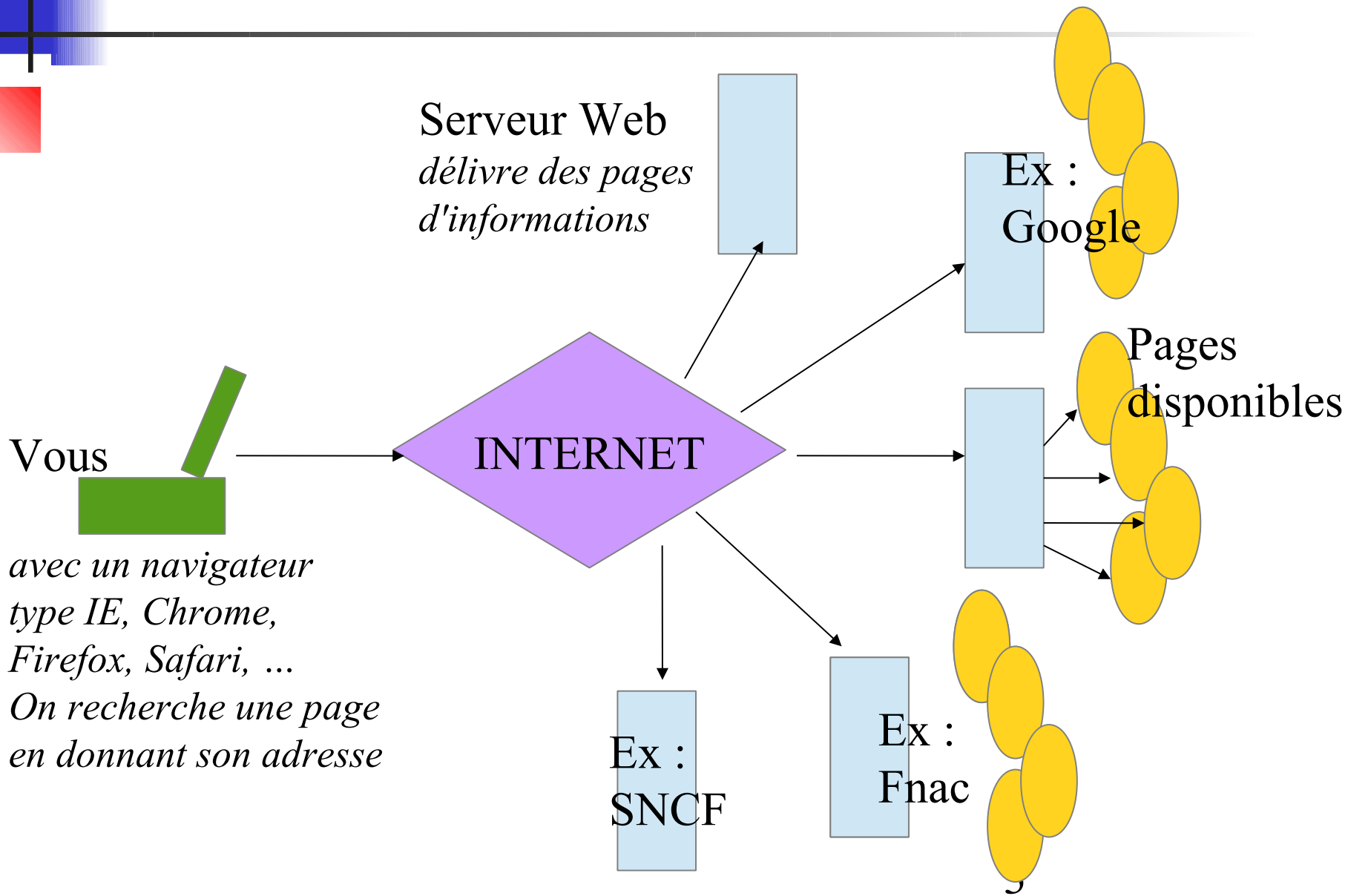
Recherche d'informations sur Internet

Les moteurs de recherche

par Alain Laponche (ORB)
octobre 2011
réactualisé pour Clio en janvier 2019

Principe des moteurs de recherche

Rappel sur le web



L'enjeu

- Au début, des gens étaient payés pour explorer manuellement Internet, lire les pages web et effectuer un classement par thème dans un annuaire (Yahoo! à l'origine)

- Mais aujourd'hui, il existe plusieurs milliards de page (plusieurs millions nouvelles chaque jour !)
- Comment traiter cette masse d'informations et trouver les plus pertinentes sur un sujet donné ?

Les moteurs de recherche

- **De façon automatique, proposer l'adresse de pages web répondant à un critère de recherche**

- **"Un moteur est un site web permettant l'accès à des ressources web (des pages, des vidéos, ...) à partir de requêtes constituées de mots-clés"**

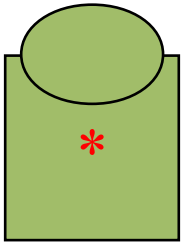
Le principe des moteurs

- 1 . Un robot (spider, crawler, ...) balaye Internet automatiquement, note l'adresse de chaque page web rencontrée, et pour chaque page, alimente un index avec tous les termes présents sur cette page
  (ex : 'brest' à la page http: //www.*bidule*.fr/*chose*)
  Plusieurs mois sont nécessaires pour un balayage (qui reste toujours incomplet)

- 2 . Quand vous posez une question, cet index est interrogé et fournit instantanément la liste des réponses
  (ex : 88 millions d'adresses en réponse à 'brest')

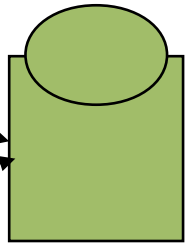
Principe

Pages déjà parcourues



Robot

Indexeur



Moteur de recherche

Base de données
(Mot / liste d'adresse)

Ordonnancement *

Question / Réponse

Ce que l'on ne peut pas rechercher ...

- Les pages Intranet (les pages internes aux entreprises)
- Les pages non appelées depuis une autre page
- Les textes dans une image ou dans des formats exotiques
- Les pages à accès payant : archives des journaux, sites 'coquins', ...
- Les pages constituées à partir de l'interrogation d'une base de données : horaires de train, disponibilité de chambres, ...
- Toute page qui nécessite au préalable de répondre à une question 'ouverte' (ex : votre prénom ?)
- Ou protégée par un mot de passe
- Au total, quelques % des informations présentes sur Internet sont accessibles à un moteur !

Autres limites des moteurs

- Les robots sont impuissants devant les différentes formes d'écriture d'une même notion (navire et bateau, animal et animaux, ...) ;
  et ils ne hiérarchisent pas les notions
  (une goélette n'est pas un navire)
- Ils traitent de la même façon un site institutionnel et une page personnelle ; ils ne reconnaissent pas les sites de référence !
- La croissance du web conduit à des milliers de réponse dans laquelle les bonnes sont perdues !

- *L'algorithme de Google propose une parade aux deux derniers points*

Le problème du classement

- Rechercher l'adresse des pages Web contenant le mot « mediator »
- Au 01/01/19, il y en avait 41,7 millions
- Comment les classer ?
  (sachant qu'un lecteur ne lira généralement que les 10 premières réponses)
  - En fonction du prix payé par le rédacteur ?
  - En fonction de leur contenu ?
  - En fonction de leur notoriété ?
  - *Google ferait intervenir plus de 200 critères !*

En fonction de leur contenu

- Le nombre de mots placés avant la première apparition du mot recherché
- La fréquence d'apparition de ce mot sur la page (ex : 3 fois / 500 mots, mieux que 4 / 50000)
- La présence du mot recherché dans le titre
- Le mot a été utilisé comme 'tag' (mot clé mis en place par l'auteur, mais non visible par le lecteur)
- La langue utilisée dans la page (/ à la question)
- L'ancienneté de la page
- *On peut retenir plusieurs de ces critères en les pondérant*

Et si plusieurs mots recherchés

- Rechercher : 'chemin de fer'
- Rappel : Il faut la présence de tous les mots. Puis :
- On *regarde l'ordre des mots*. Ainsi une page avec « Le cheval a perdu son fer dans ce chemin de travers » sera sélectionnée, mais classée après :
- « Ce chemin de croix est très long ... c'est vraiment le fer de lance ... »
- Mais *on regarde aussi la distance entre les termes recherchés*. Et la page contenant la phrase suivante sera encore mieux placée : « Ce chemin emprunte le pont de fer »
- Et bien entendu, toutes les pages contenant «chemin de fer» seront en tête

En fonction de leur notoriété

- Soit 5 pages répondant à la recherche
  - Page 1 (contient un lien vers la page 2)
  - Page 2 (aucun lien)
  - Page 3 (liens vers les pages 1 et 2)
  - Page 4 (aucun lien)
  - Page 5 (lien vers la page 3)
- Quel ordre de présentation ?
  - Page 2, puis 3 et 1, puis 5 et 4
    donc en fonction du nombre de liens qui pointent vers chaque page

Tricher avec les robots !

- L'enjeu est capital pour les entreprises : leurs pages doivent absolument être dans les premières présentées !
- Donc tentation de tromper les robots des moteurs de recherche :
  - mots dissimulés dans le fond d'écran (en particulier, mot souvent recherché, mais sans rapport avec le sujet traité dans la page !)
  - répétition exagérée du même mot,
  - Générateur d'appels automatiques,
  - notoriété renforcée par des sites bidons

Les problèmes pour l'utilisateur

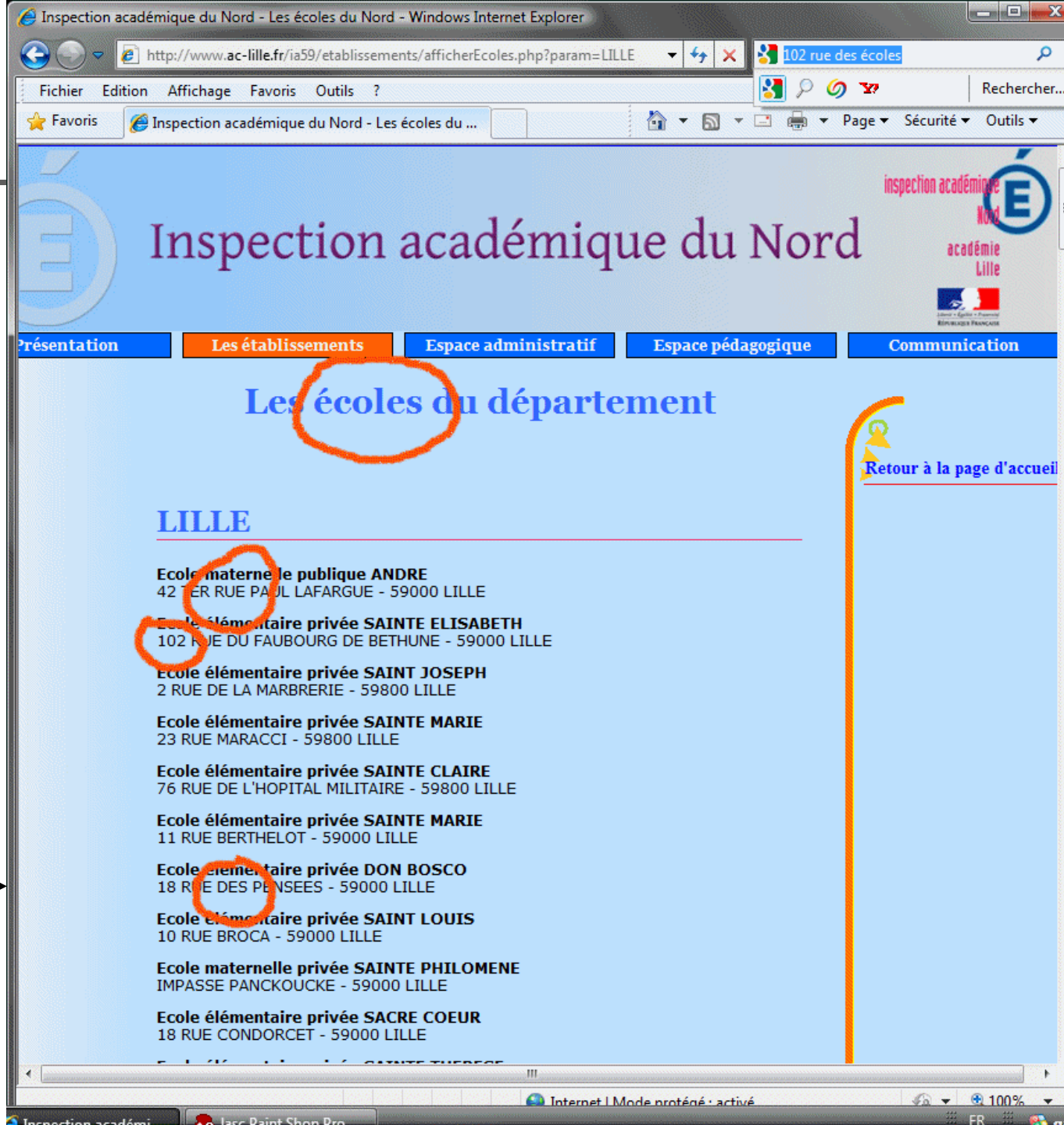
- Les synonymes (navire et bateau)
- Les polysémies (bar, lieu, sinus, ...)
- Les différentes écritures d'un même mot (pluriel, conjugaison, ...)

  (ex : oeil et yeux !)
- Les réponses importantes 'oubliées'
- Les réponses 'non pertinentes' (le bruit)

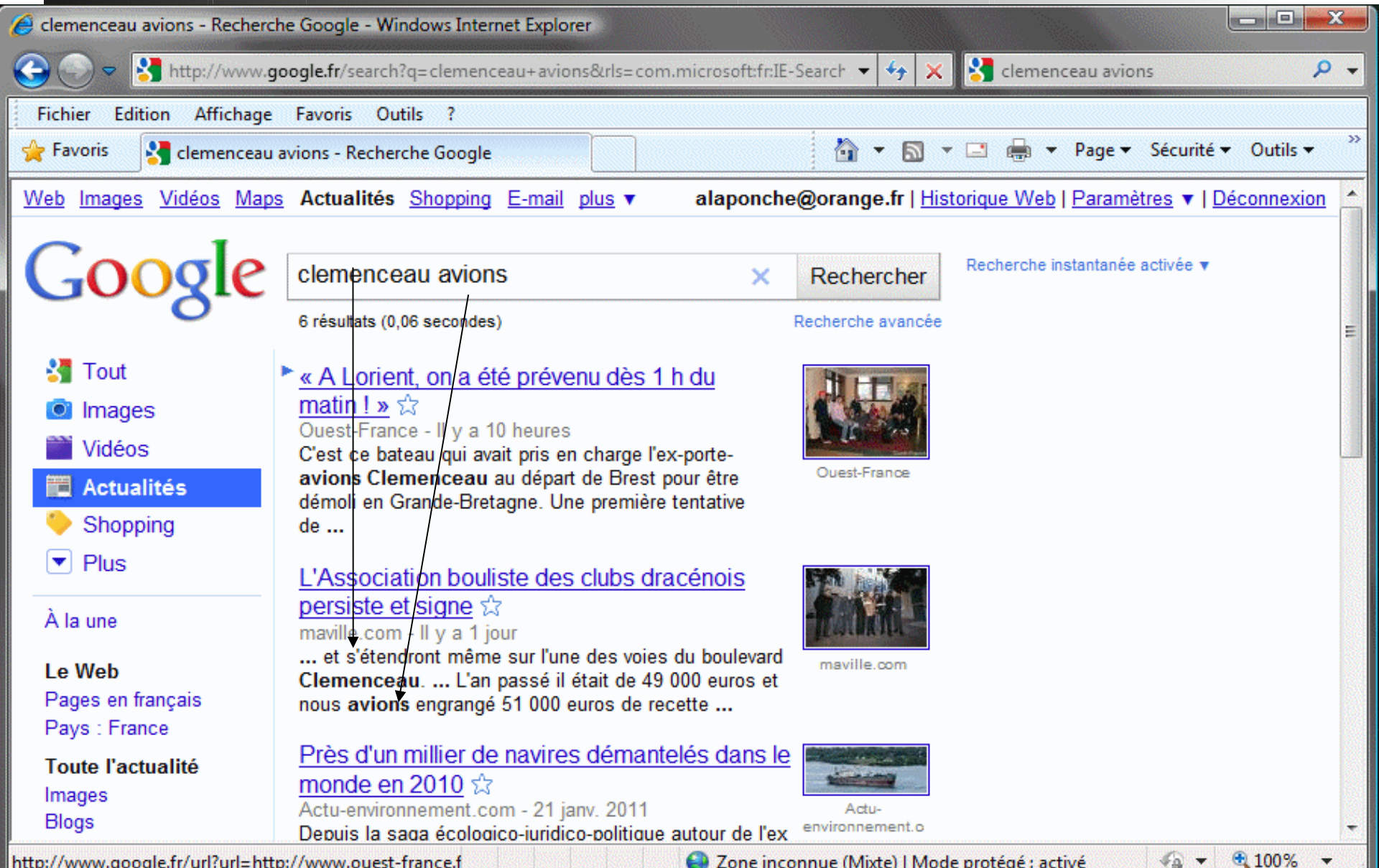
  surtout en provenance de sites commerciaux et de particuliers. Tient au principe même de recherche...

Sélection des pages

- Recherche avec '102 rue des écoles'
- Parmi les réponses →



Autre exemple de réponse bruitée



Les avantages des moteurs

- Richesse infinie (le web est sillonné jour et nuit, et la probabilité d'une absence totale de réponse est très faible)
- Accès **rapide et simple**

Les inconvénients des moteurs

- Indexation sans comprendre le sens des mots, sans repérer les fautes d'orthographe, sans distinguer les homonymies
  --> « bruit » ou « silence » important
- Pertinence des résultats très liée à la formulation de la question
  (*on n'est pas forcément un bon documentaliste*)
- Pas de structuration logique comme dans un annuaire

Infos sur les moteurs

- Le précurseur : Altavista
- Le leader actuel : Google
- En France, Orange a cherché à développer son propre moteur

- Toutes les informations sur les moteurs de recherche (actualités, statistiques) : http://www.abondance.com/

Quelques moteurs

Les principaux moteurs

- Google
- Bing (Microsoft)
- Exalead (Dassault Systèmes)
- Yahoo
- Qwant (français)

Google

- www.google.fr
- Le leader des moteurs de recherche
- Gogol = 1 suivi de cent zéros = infini
- Création en 1997 par Sergueï Brin et Larry Page
- Version française en 2000 ; cette année, il indexait déjà 1 milliard de pages web !
- Société californienne

Quelques chiffres 2015

- 30 000 milliards de pages web indexées (20 milliards visités / jour)
- Réponse à 3,3 milliards de requête par jour
- 94% du marché européen des moteurs de recherche (Bing 2,5%, Yahoo 1 %)
- 53 800 employés
- Le plus gros réseau informatique mondial
- La première marque au monde !

Les raisons du succès de Google

- La pertinence des réponses
  Grace à son algorithme de tri : à l'inverse des moteurs précédents, cette pertinence s'améliore au fur et à mesure que le web grossit !
- La vitesse d'obtention des réponses, liée :
  - au nombre de ses ordinateurs
  - à la sobriété de ses pages (pas d'image)
- Le volume de sa base de données
  estimé à 8 milliards de pages
- L'absence (apparente) de publicité
- La diversité des services associés

PageRank, le secret de Google

- Note de 0 à 10 attribuée par Google pour chaque page trouvée sur le web
- Plus une page est citée par d'autres pages, plus son 'pagerank' (sa **notoriété**) sera élévé
  (cela signifie que d'autres administrateurs de sites reconnaissent la pertinence de cette page)
- Encore faut-il que les pages qui font un lien vers une page donnée, aient elles-même une bonne notoriété !
  Ex : pour un restaurant, une citation par le site de Michelin a plus de poids qu'une citation par mon site personnel
- *Les réponses de Google sont essentiellement classées selon le pagerank des pages*

Le financement Google

- La plupart de ses produits sont disponibles gratuitement
  Et il y a donc très peu de vente de produits
- Notons toutefois :
  la vente d'espace de stockage en ligne

- Sa principale ressource : la publicité
  Même si très peu visible sur ses pages
- 1 - La vente de mots-clés aux enchères
- 2 - La vente de “profils”

Système AdWords

- **Vente de mots-clés aux enchères**
- Si un des mots vendus est tapé par un utilisateur, celui-ci voit apparaitre une liste de “liens sponsorisés”
- Ces liens correspondent aux annonceurs qui ont enchéri pour ce terme
  (plus l'enchère est élevée, plus le lien apparait)
- Si l'utilisateur clique sur un tel lien, l'annonceur est alors facturé.
  On parle de facturation au clic

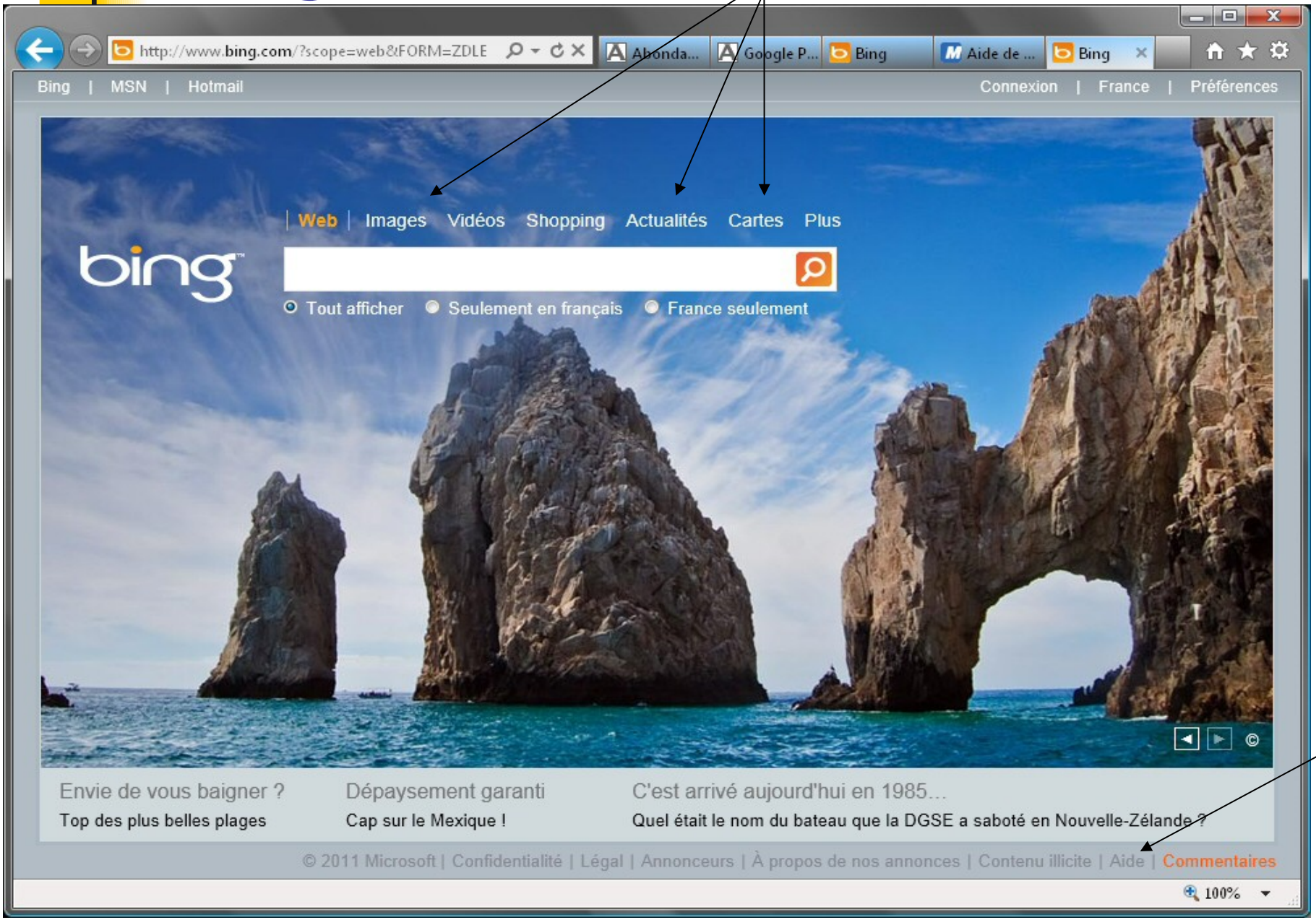
Système AdSense

- **Vente de "profils"** concernant les utilisateurs de Google
- Une entreprise peut acheter ces profils et ainsi, afficher sur son site web des encarts publicitaires correspondants aux centres d'intérêt des personnes qui consultent son site
- Ces profils sont élaborés à partir des recherches effectuées par ces personnes, du contenu des messages stockés sur Gmail, ...

Bing

- http://www.bing.com/
- La contre-attaque de Microsoft
- Version française lancée en mars 2011

Bing


http://www.bing.com/?scope=web&FORM=ZDLE
Abonda...
Google P...
Bing
Aide de ...
Bing
Bing | MSN | Hotmail
Connexion | France | Préférences
Web | Images Vidéos Shopping Actualités Cartes Plus
bing
Tout afficher
Seulement en français
France seulement
Envie de vous baigner ?
Top des plus belles plages
Dépaysement garanti
Cap sur le Mexique !
C'est arrivé aujourd'hui en 1985...
Quel était le nom du bateau que la DGSE a saboté en Nouvelle-Zélande ?
© 2011 Microsoft | Confidentialité | Légal | Annonceurs | À propos de nos annonces | Contenu illicite | Aide | Commentaires
100%

Yahoo France

Exalead

- Français
- Racheté par Dassault Systèmes en 2010
- Très pointu ; par exemple recherche sur une base phonétique
- Plutôt orienté entreprises (il est inclus dans des sites commerciaux, type sncf)

Qwant

- Français, lancé en 2013
- “Le moteur qui respecte votre vie privée”
- Acord en 2016 avec la Fondation Mozilla, conduisant à un Firefox optimisé Qwant
- Moteur par défaut dans certaines collectivités ou ministères (La Défense)
- Fin 2018 : 80 % du trafic vient de France

Les moteurs propres à un site

- Ces moteurs ne balayent qu'un site, celui sur lequel ils sont installés
- Objectif : trouver les pages traitant d'un sujet, présentes sur un site
- Le formulaire de recherche est en général accessible depuis la page d'accueil du site
- Exemple : BMO, Le Télégramme

Les moteurs pour PC

- Paradoxal : il faut un dixième de seconde pour trouver une page web sur un sujet donné existante en Australie
  Mais plusieurs minutes pour trouver un fichier traitant du même sujet sur son propre ordinateur
- D'où depuis Vista, des logiciels font sur votre PC le travail des moteurs (analyse permanente des contenus de vos fichiers)
- Exemple avec le formulaire de Windows

Le formulaire de saisie

Où taper ses critères de recherche ?

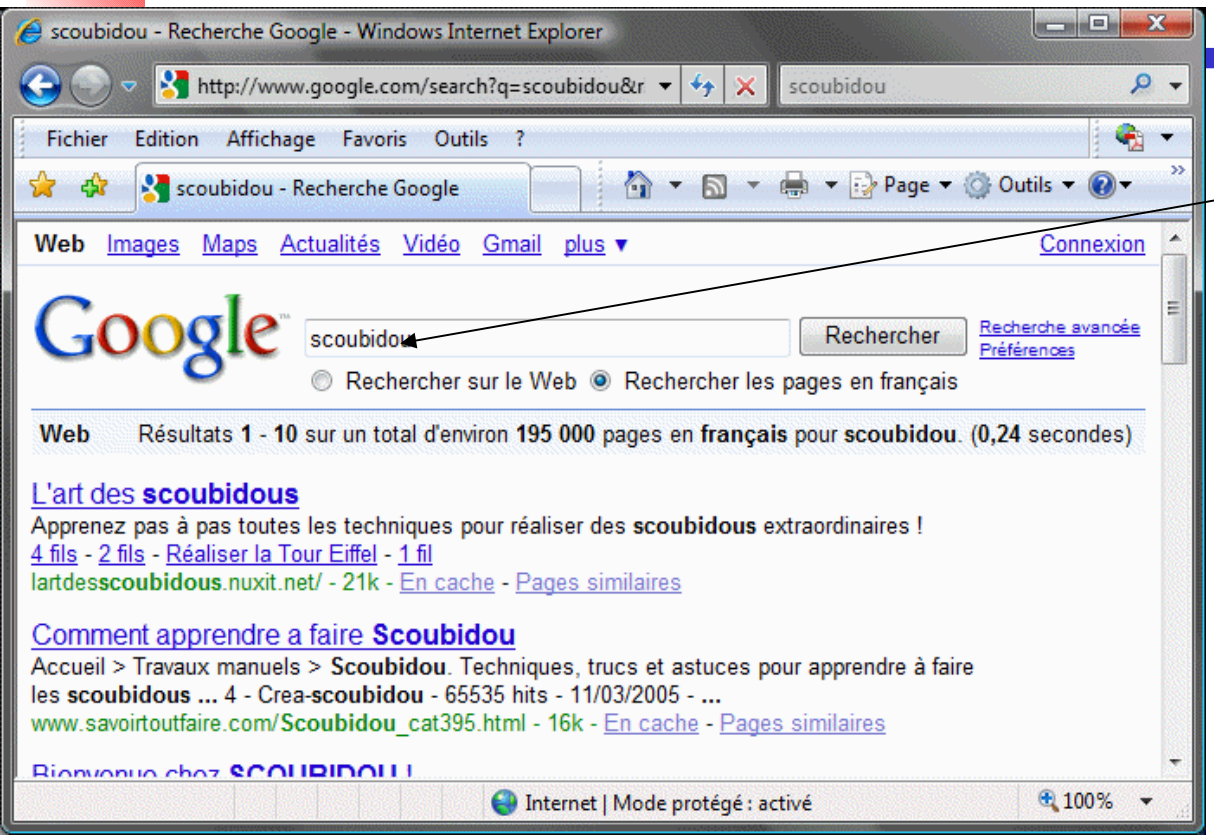
Où taper ses critères de recherche ?

- Il faut positionner son curseur dans un **formulaire** et y saisir les termes recherchés
- Les formulaires de saisie sont présents :
  1 – sur la page d'accueil du moteur
  (encore faut-il se rendre dessus ! ;
  par exemple l'avoir rangé dans ses favoris)
  2 – sur chaque page de résultats
  (bien sûr, pas valable pour la première interrogation !)
  3 - dans la ligne de commande de votre navigateur
  (en partie haute de la fenêtre du navigateur)

1- L'accès à la page d'accueil

- soit saisir l'adresse
  Exemple : www.google.fr
- soit sélectionner cette adresse dans un 'favori' préalablement enregistré
- soit lancer une recherche quelconque (ici « ddd ») dans la fenêtre de saisie du navigateur, puis une fois obtenue une réponse, cliquer sur le logo du moteur

2 – Le formulaire dans une page de résultats



- Les pages de réponses comportent toujours un formulaire qui permet de poser une nouvelle question

Je peux donc remplacer “scoubidou” par autre chose !

3 - Le formulaire intégré au navigateur

- Certains navigateurs proposent un champ de recherche à l'extrémité droite de la ligne d'adresse
- Mais la plupart offre désormais la possibilité de taper dans le champ 'adresse web', soit une adresse, soit des critères de recherche
  (si la saisie n'a pas la forme xxx.xxxxxx.fr, elle sera destinée au moteur de recherche)
- Tous les navigateurs proposent un moteur de recherche par défaut, que l'on peut choisir parmi une liste

Paramétrage du moteur

- On peut en général choisir :
  - la langue de dialogue avec le moteur
  - la langue ou l'origine des pages sélectionnées
  - et divers autres options
- Pour cela, trouver le lien “Paramètres” “Outils” ou “Réglages”

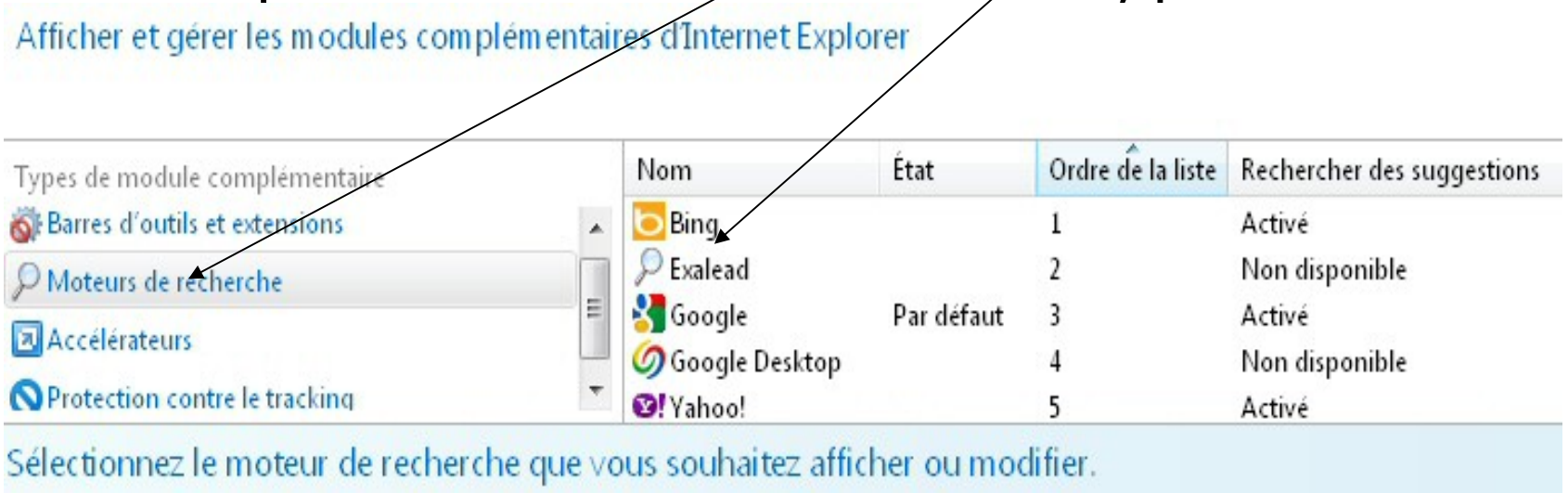
- Choix valables seulement pour l'ordinateur en question (ou sinon, il faut disposer d'un compte Google et s'y connecter)

Choisir son moteur

- Tous les navigateurs proposent par défaut un moteur de recherche
- Il suffit de taper les termes recherchés dans la première ligne du navigateur
- Mais si on veut en utiliser un autre
  - Soit il faut accéder au site (ex :www.google.fr)
  - Soit changer le moteur par défaut

Choix du moteur de recherche avec IE9

- Supposons Google par défaut ; or on voudrait Bing
- Pour en changer :
  - Cliquer sur le picto « Outils » (en haut à droite)
  - Retenir « Gérer les modules complémentaires »
  - Dans « Types de module », cliquer sur « Moteurs »
  - Dans la liste, cliquer sur sur le nom Bing
  - Cliquer sur le bouton « Par défaut », puis sur « Fermer »



Choisir la version de moteur

- Pour Google, il existe une version par pays
  - adaptée aux régles linguistiques
  - spécialisée sur les pages de ce pays
  - utilisée pour tester des fonctionnalités
- Outre le formulaire simple, il peut y avoir un interface dit « avancé »
  Ex : https://www.google.com/advanced_search

Les méta-moteurs

- Un méta-moteur recueille et réordonne les résultats fournis par plusieurs moteurs de recherche classiques. L'ordre final est le résultat d'un compromis entre tous les résultats proposés par ces moteurs
- Lui-même n'a donc pas sa propre base de donnée ... et il « pille » les vrais moteurs
- Intérêts : une seule requête pour faire l'équivalent de plusieurs recherches ; une seule syntaxe d'interrogation à connaître
- Inconvénients : il ne donne pas accès aux formulations sophistiquées que peuvent proposer certains moteurs

La formulation d'une requête

varie d'un moteur à l'autre, mais en général ...

Ecriture des termes recherchés

- Indifférence aux majuscules / minuscules (ex *ibm et Ibm*)
- Sensibilité réduite aux accents, cédilles, ...
- Les mots non significatifs d'une langue sont ignorés (en français : *le, mes, car, un, ...*)
- Tout ce qui n'est pas lettre ou chiffre est équivalent à un espace (ex : *Mantes-la-Jolie / Mer* = mantes jolie mer)
- L'ordre des mots joue sur l'ordre des réponses

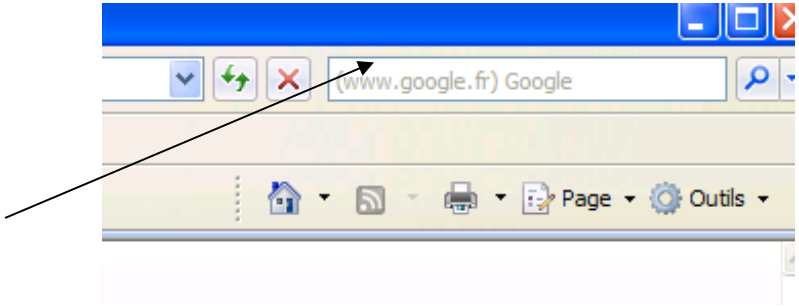
- En conséquence :
  '*Quelle est la date de naissance de Charles Trenet ?*'
  est équivalent à '*date naissance charles trenet* '
  *(*mais l'idéal est '*charles trenet date naissance ')*

Si plusieurs termes concernés

- Rechercher la présence simultanée de plusieurs termes
  Les saisir séparés par un espace (ex : peinture picasso)
- Exclure un terme
  Exemple : musee peinture -picasso
  (ici, le moins écarte les pages avec le mot picasso)
  Il ne faut pas d'espace entre le moins et le terme écarté
- Rechercher une expression exacte
  Exemple : "rue de Lyon" dubois
  Grâce aux guillemets, les pages parlant de la rue Dubois située à Lyon ne sont pas retenues. Mais une page parlant de Mr Dubois, habitant « rue de Lyon » à Brest sera retenue.
- Que recherche-t-on avec ceci ?

Exemples de sélection

- Supposons une page web avec :
  « Suite au passage du cyclone Hugo, la rue Victor Hänsen a beaucoup souffert »
  On recherche ...
- Victor Hänsen
- victor hansen
- rue hansen
- rue victor hugo
- "rue victor hugo"
- rue victor hugo -passage



Exemple de recherche itérative

- **laponche** : hotel et bernard scientifique (parmi 120 000 réponses)
- **laponche -hotel -bernard** : alain (moi) + michelle, stephanie, eugène et françois (la bouverie) + maxime, Gael, Coline, Jerome et Stéphanie (sur Facebook)
- **laponche -hotel -bernard -facebook -alain -bouverie** : beatrice, laetitia, benoit
- **laponche -hotel -bernard -facebook -alain -bouverie -beatrice -laetitia -benoit** : pierre, jean, jeanne, olivier (il reste encore 6 370 pages)
- Puis étude prénom par prénom : Ex « olivier laponche »

Conseils

- Utiliser des mots précis (plutôt que des génériques). Ex : 'rosier' plutôt que 'rose'
- Attention aux polysémies (mot à double sens). Pour lever une ambiguïté, ajouter un terme générique. Ex : 'bar poisson'
- Penser aux synonymes (ex : 'loup' et 'bar', 'navire' et 'bateau') et aux différentes formes d'écriture d'un même mot (pluriel, conjugaison, ...)
- Attention aux fautes d'orthographe
  (les votres, mais aussi celles des rédacteurs ; ex infartus)

Stratégie de recherche

- Commencer toujours par une question simple (un ou deux termes), et en fonction des premières réponses, adapter la requête
- Ne consulter que les 2 premières pages de réponse : si aucune satisfaisante, revoir la question
  - en cherchant des termes synonymes ou plus précis
  - en écrivant les termes dans l'ordre de présence la plus probable dans les pages visées
    Ex: *'porte avion foch' est mieux que 'foch navire'*
- Etudier l'aide associée à chaque moteur pour connaitre leurs opérateurs de recherche (OU, recherche sur un site, ...)

Quelques exemples de recherche

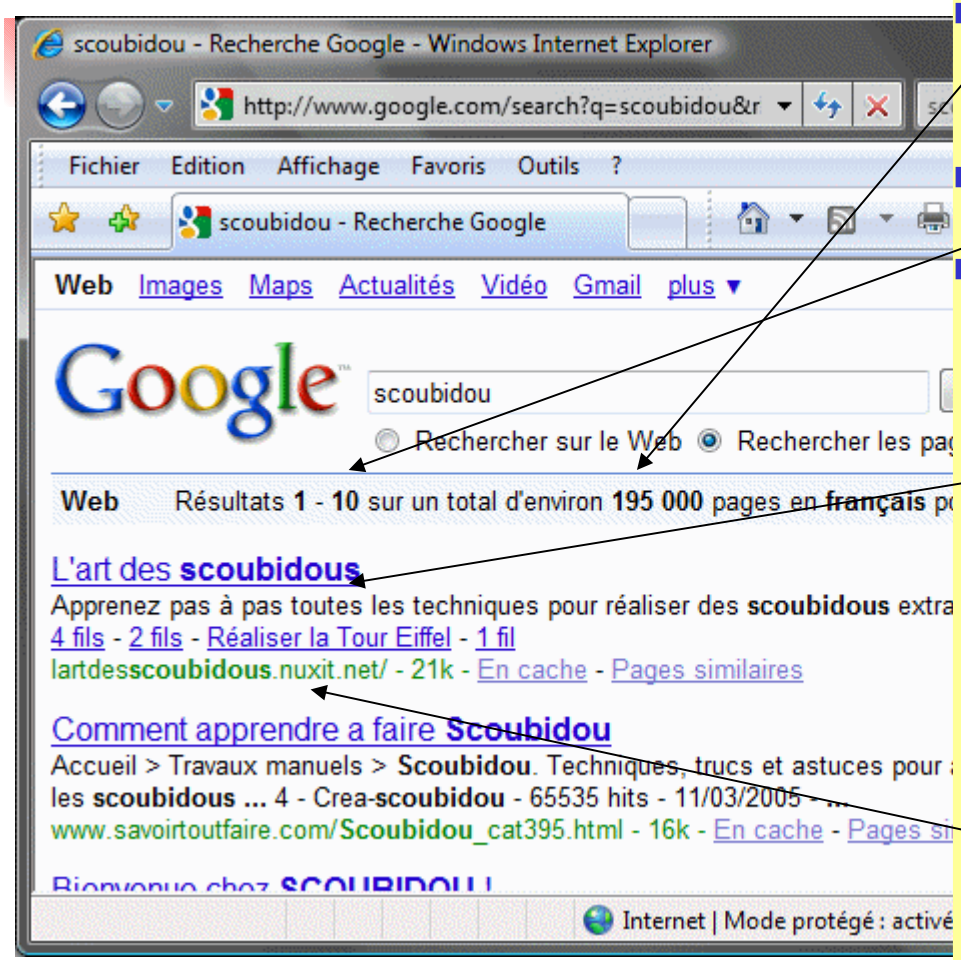
- La hauteur du Mont Blanc
- La couleur du drapeau argentin
- Auteur du texte « Les feuilles mortes »
- L'usage du médicament Réactine
- La longueur exacte d'un marathon
- Les événements importants survenus en 1234
- Cours actuel de l'action Alcatel
- Heure du prochain avion partant de Guipavas
- La distance par la route entre Brest et Perpignan
- 31108 / 14 =

L'exploitation des réponses

Comprendre les réponses !

- Si je tape SNCF, les réponses ne proviennent pas d'un site de la SNCF, mais bien de la base de données du moteur
- Les réponses sont constituées par une liste d'adresse de pages web comportant les mots recherchés. Il faut cliquer sur une réponse pour accéder à la page correspondante et donc au site demandé

Présentation type (ici Google)



- Il y a 195 000 pages en français comportant le mot « scoubidou »
- Les 10 premières sont affichées ici
- Chaque page de réponse comporte :
  - Le titre de la page : *cliquer sur ce lien pour obtenir la page complète*
  - Un extrait du contenu de cette page sur quelques lignes
  - L'adresse web de la page
  - Sa taille (ici 21 k octets)

Des aides et filtres

- Permettent d'améliorer les réponses
- Des "filtres" sur la langue, la date de publication, le pays … en réduisent le nombre
- Des "recherches associées" proposent de reformuler la requête

Les différents types d'informations que l'on peut rechercher

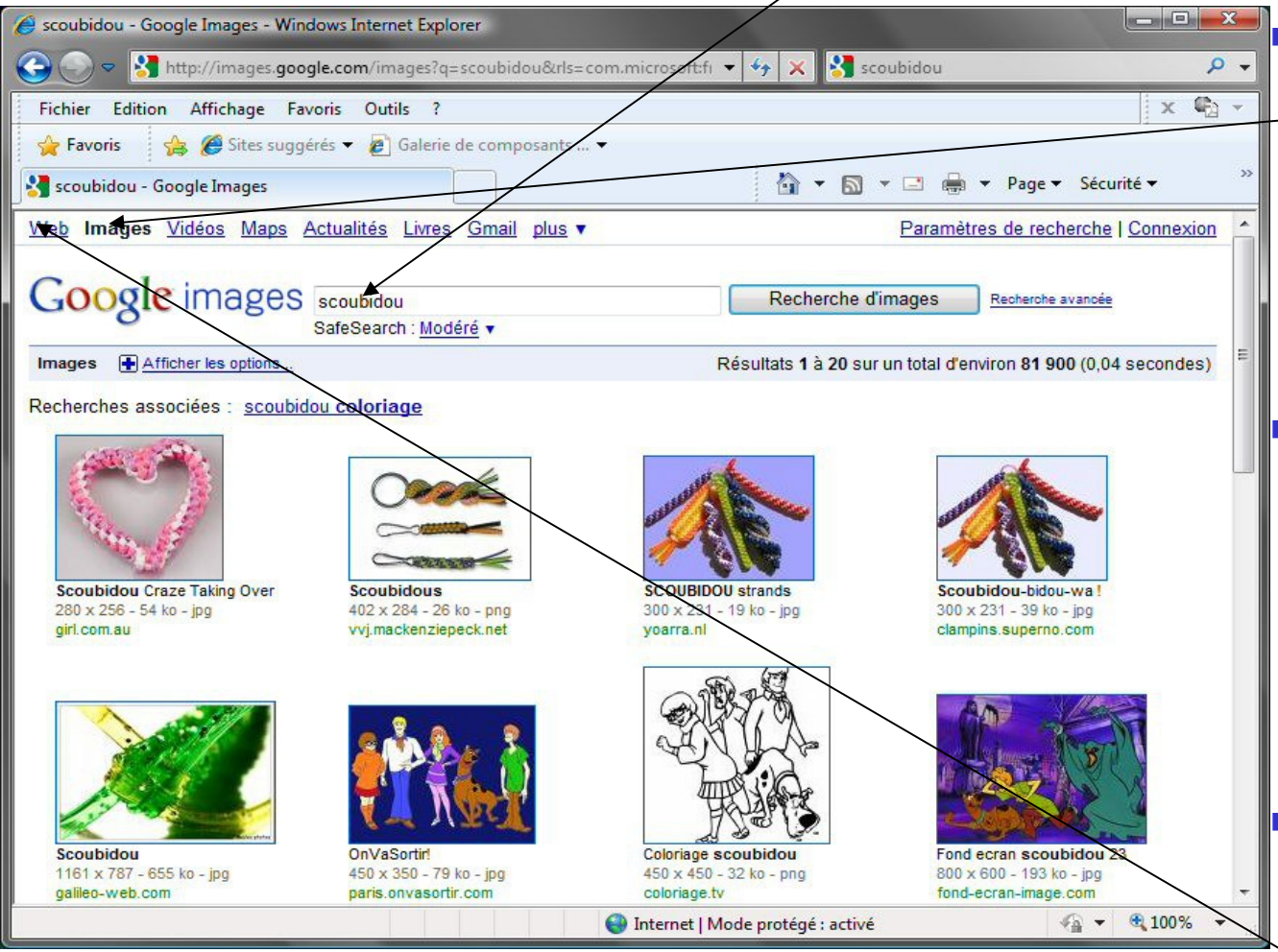
Ce que l'on peut rechercher

- Des pages web répondant aux critères tapés
  Mais aussi :
  - des images et des vidéos
  - la localisation de lieux sur des cartes
  - des informations d'actualités
- Et parfois
  - des forums, des réseaux sociaux, ...
  - des livres (Google)

La recherche d'images

- Même principe avec robot et base de données
- Mais ici l'indexation porte essentiellement sur les textes entourant l'image (légende, titre de page, mots voisins de l'image, ...)
- Il peut y avoir utilisation des données associées à la prise de vue (date, ouverture, ...)
- Et pour les moteurs les plus développés, un traitement sur l'image (couleur, portrait, reconnaissance de visage, ...)

Exemple : Images de 'scoubidou'



- Il suffit de cliquer sur le lien « Images » et 81 900 photos ou graphes sont proposés
- Fonctionnement identique pour « Vidéos », « Actualités » et « Livres »
- Pour revenir sur une recherche textuelle, cliquer sur « Web »

Accès aux actualités

- Les pages des médias (journaux, télé, ...) font l'objet d'une analyse permanente (alors que les pages traditionnelles ne le sont qu'entre une fois par jour à une fois par an)
- Le contenu de certains réseaux sociaux (Twitter et Facebook) est également analysé
- Accès depuis le lien 'Actualités'

Les actualités