Séance 2

Retour sur les exercices de la séance dernière : quelles conclusions tirer sur l'accès à l'information ?

A. Wikipedia

Lancée en 2001, 301 langues différentes, presque 50 millions d'articles en tout

Modifiable, corrigée (par des *Automates de correction* + *volontaires bénévoles*)

Un exemple de portail : Littérature

Quelques exemples de ce qu'on peut faire à partir d'un article wikipédia : modifier la page (de deux façons), trouver la liste des modifications, comparer deux versions, voir la page de discussion, voir l'état d'avancement et d'intéret de la page...

Quelques liens intéressants sur le fonctionnement de Wikipédia :

- Ce que Wikipédia n'est pas
- Principes fondateurs de Wikipédia
- Vérifiabilité sur Wikipédia

Autres projets liés : wikibooks, wikiversité, wikilivres, wikimedia commons

A écouter si vous êtes intéressés :

Emission de France Culture sur Wikipédia

B. Moteurs de recherche

Google, Qwant, DuckduckGo, Ecosia

/!\ Moteurs de recherche =/= Internet !
Certains logiciels peuvent contenir des moteurs de recherche (exemple, les explorateurs de fichiers récents, ou Spotlight sous Mac)

Moteurs de recherche **généralistes ou spécialisés** :

- Google scholar, CiteSeerX (articles scientifiques et académiques)
- Qwant junior (pour mineurs)
- Tinyeye (recherche par images), google images

Méta-moteur : Regroupe les résultats d'autres moteurs de recherche (ex: Lilo)

Affichage des *résultats* = Résultat d'un **algorithme de sélection** -> Classe les résultats
PageRank (Google) tri par popularité (Globalement basé sur le nombre de liens menant à cette page)
Calcul approximatif et très surveillé (influence les utilisateurs) -> *SEO* (Search Engine Optimization)

Internet de surface / internet profond

Internet de surface

Indexé ou indexable = accessible aux *robots d'indexation*
Facilement accessible via les moteurs de recherche, les liens hypertexte
Taille du web indexé par Google = 14.5 milliards de pages en 2015

Internet profond = deep web

/!\ deep web vs dark web
Web profond = environ *500 fois plus grand* que l'internet de surface

- sites internet dynamiques, accessibles uniquement en envoyant la bonne requête
- sites "orphelins" (pas de liens hypertexte vers ces pages)
- Tout contenu non textuel (de moins en moins vrai) = Video, Audio
- Tout site "protégé" contre l'indexation (presse en ligne protégée, videos à la demande, comptes en ligne, web mail)
- Bases de données diverses
- Site scriptés (javascript, flash, ajax)
- archives du web

Indexation

Faite par des « robots », (*bots*, spiders, crawlers ou agents)
Explorent automatiquement le Web en suivant récursivement les hyperliens entre les pages
Analyse : parcours d'une page, puis **indexation** (= donner des mots-clés, résumer)

On peut donner des **indications** aux robots d'indexation

- *Exclusion de certaines pages/zones*
  Ressource de format texte placée à la racine d'un site web
  Contient une *liste des ressources* du site qui ne sont pas censées être indexées
  Par convention, les robots consultent `robots.txt` avant d'indexer un site Web. Exemple :

  User-agent: * Disallow: /personnal-data

- *améliorer l'indexation*
  donner des infos complémentaires (pages à visiter, fréquence de màj, infos en plus)

Des fois, ça ne fonctionne pas très bien… (ex: Google Images "pizza aux anchois")

Requêtes

Comment faire une **bonne requête** ?

Exemple : recherche sur la source de l'Amour (fleuve) : si `source amour`, on risque de trouver autre chose !
Préciser : `source amour fleuve` ou `source amour geographie`

Astuce : rechercher en *anglais* si possible (stats anglais/français ?)

Syntaxe

Symbole	Signification	Exemple
+	Recherche de mots dans la même page	`jupiter + dieu`
-	Mots clés « exclus »	`jupiter -planete`
"…"	Recherche d'expression exacte	`"cuisse de Jupiter"`
	Troncature	`jupiter*`
site:	restriction à un site web	`science-fiction site:lemonde.fr`
filetype:	restriction à un format	`chaton filetype:pdf`