Cap.9. Baze de date web

CUPRINS

1. Baza de date web

Ce este o „baza de date Web" (Web database) ?

Baza de date Web: lista organizata de pagini web (titlul, antetul etc) = **indexare**.

Tipuri de indexare:

- **full text:** includerea tuturor cuvintelor dintr-o pagina web în BD pentru cautare, prin programe speciale = spiders sau robots. Utilizata de **Google**, **Altavista**
- **manuala**: o persoana decide ce cuvinte sunt importante si le selecteaza pentru indexare. Utilizata de **Yahoo Directories** sau **Magellan**

1. Baza de date web

Exista 3 clase de baze de date web:

- baze de date ce monitorizeaza **TOATE** categoriile de pagini WWW ;
- baze de date ce monitorizeaza **NUMAI paginile** WWW considerate **populare** (numar mare de vizitatori);
- baze de date ce monitorizeaza **NUMAI paginile** WWW **ce îndeplinesc anumite criterii** (legate de: calitatea informatiei furnizate, sau de tipul de informatie– ex. medical, stiintific, stiri etc).

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

2. Instrumente de interogare a BD web

a) Motoare de cautare

b) Directoare web (anuare online, repertoare tematice)

c) Biblioteci virtuale

d) Invisible (deep) Web

e) Motoare de meta-cautare (metasearch engine)

f) Utilitare de cautare de tip desktop

a) Motoare de cautare

Motor de cautare: o baza de date continând pagini Web ce pot fi regasite pe baza unor cuvinte cheie si care continua sa scaneze Internetul, cu ajutorul unor programe automate (spiders, robots) în cautare de pagini noi. Informatia rezultata :este indexata si stocata în baza de date.

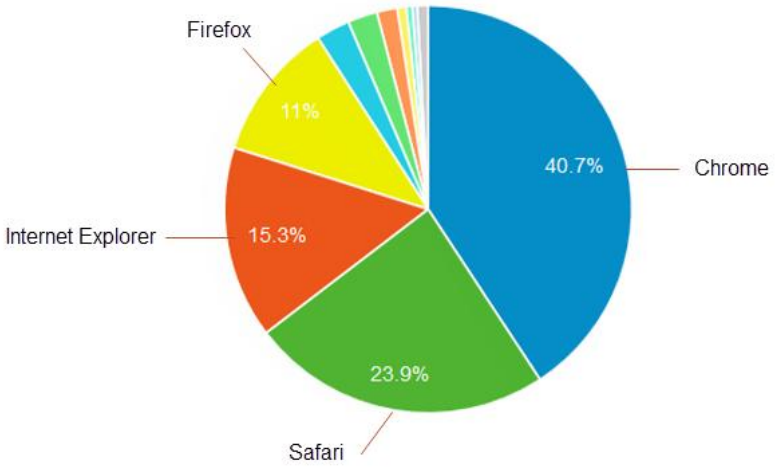
Structura unui motor de cautare :

- interfata de interogare,
- baza de date,
- algoritmul de cautare,
- structura de afisare

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

a) Motoare de cautare

Browsere WEB: cota de piata 2015



a) Motoare de cautare

Algoritmi de cautare:

- Algoritmi de căutare în listă
- Algoritmi de căutare arborescentă
- Algoritmi de căutare de tip SQL
- Algoritmi de căutare „în cunoştiinţă de cauză”
- Algoritmi de căutare cu constrângeri

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Istoric metode de cautare web= metode de interogare BD web

- pe bază de indecşi – crawling & indexing
- Lycos (1994)
- AltaVista (1995)

- pe baza ierarhiilor de termeni (servicii de tip catalog – topic directory):
- Yahoo! – Yet Another Hierarchical Officious Oracle! (1994)

- hibrid (indecşi + ierarhii de termeni)
- Excite

-pe baza legăturilor hipertext – hyperlink analysis
- Google (1996)

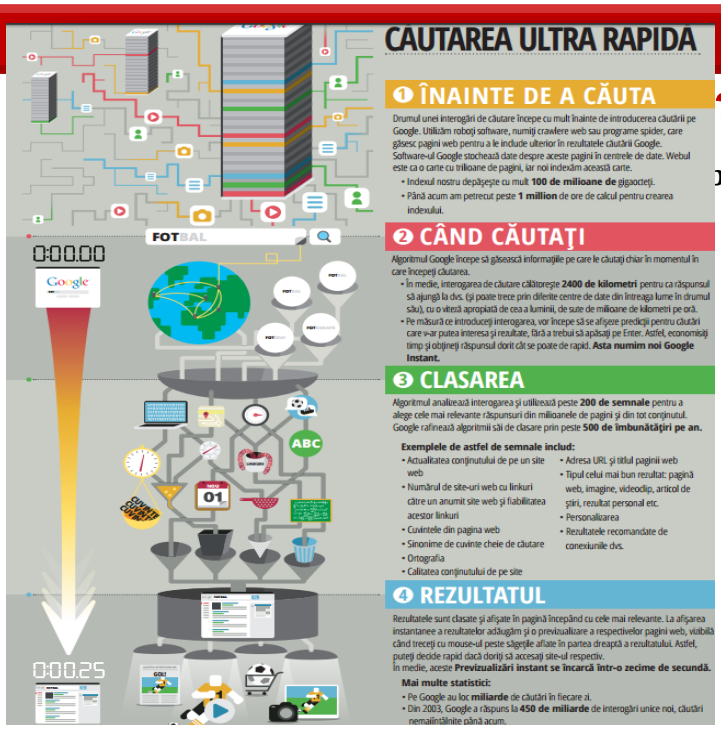
-interogare în paralel a mai multor motoare de căutare şi compilarea listelor de adrese ale paginilor găsite:
- Clusty, Dogpile, Kartoo, Mamma, SurfWax

a) Motoare de cautare

http://www.altavista.com	http://www.aol.com/netfind	http://www.askjeeves.com
http://www.directhit.com	http://www.alltheweb.com	http://www.excite.com
http://www.goto.com	http://www.google.com	http://www.hotbot.com
http://www.infoseek.com	http://www.go.com	http://www.inktomi.com
http://www.lycos.com	http://www.search.msn.com	http://www.yahoo.com
http://www.nlsearch.com	http://www.dmoz.org	http://www.snap.com
http://www.planetsearch.com	http://www.webcrawler.com	http://www.mamma.com

Din Romania

http://www.acasa.ro	http://www.alfa.ro	http://www.alias.ro
http://www.axanet.ro	http://www.betesda.com	http://www.bumerang.ro
http://www.cauta.ro	http://www.cefaci.ro	http://www.click.ro
http://www.croif.net	http://www.dot.ro	http://www.domino.kappa.ro
http://www.ebony.ro	http://www.ems.ro	http://www.edison.ro
http://www.go2net.ro	http://www.go2web.ro	http://www.gaseste.com



Cum functioneaza Google?

Strategia de regasire a informatiei in Google se bazeaza pe :
- ❑ accesarea web-ului cu **crawlere**
- ❑ şi **indexarea** a miliarde de documente de pe web.

Metoda de cautare Google include:
- ❑ **colectarea şi organizarea informaţiilor** de pe mii de calculatoare din miliarde de pagini web
- ❑ cautarea se face in **indexul** creat de Google nu in web
- ❑ **ordinea de furnizare a rezultatelor** este data de un **scor** obtinut prin clasificarea informatiilor prin intermediul a aprox. 200 de interogari :
  - ❑ **nr de aparitie a cuvintelor cautate in documentele web**: in titlu, in continut, in adresa URL, exista sinonime pentru cuvintele respective,
  - ❑ **calitatea site-ului** din care provin cuvintele cautate (se elimina spam-uri) si **popularitatea site-ului** (nr. de accesari ,PageRank)?
- ❑ **returnarea rezultatelor :** instantaneu

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Crawlere Google: GoogleBot

Crawler	User agent token	Full user agent string (as seen in website log files)
Googlebot (Google Web search)	`Googlebot`	`Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)` or (rarely used): `Googlebot/2.1 (+http://www.google.com/bot.html)`
Googlebot News	`Googlebot-News` (`Googlebot`)	`Googlebot-News`
Googlebot Images	`Googlebot-Image` (`Googlebot`)	`Googlebot-Image/1.0`
Googlebot Video	`Googlebot-Video` (`Googlebot`)	`Googlebot-Video/1.0`

Crawlere Google: GoogleBot

Google Mobile (feature phone)	Googlebot-Mobile	• SAMSUNG-SGH-E250/1.0 Profile/MIDP-2.0 Configuration/CLDC-1.1 UP.Browser/6.2.3.3.c.1.101 (GUI) MMP/2.0 (compatible; Googlebot-Mobile/2.1; +http://www.google.com/bot.html) • DoCoMo/2.0 N905i(c100;TB;W24H16) (compatible; Googlebot-Mobile/2.1; +http://www.google.com/bot.html)
Google Smartphone	Googlebot	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.96 Mobile Safari/537.36 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
Google Mobile AdSense	Mediapartners-Google or Mediapartners (Googlebot)	[various mobile device types] (compatible; Mediapartners-Google/2.1; +http://www.google.com/bot.html)

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Crawlere Google: GoogleBot

Google AdsBot landing page quality check	AdsBot-Google	AdsBot-Google (+http://www.google.com/adsbot.html)
Google app crawler (Used to fetch resources for mobile apps, obeys AdsBot-Google robots rules.)	AdsBot-Google-Mobile-Apps	AdsBot-Google-Mobile-Apps

Caracteristici Google

Răspunsuri imediate: afişează imediat răspunsuri şi informaţii pentru căutări cum ar fi vremea, rezultatele sportive şi date pe scurt.

Completare automată: anticipează ce aţi putea căuta. Înţelege chiar şi termenii care au mai multe sensuri.

Cărţi: găseşte rezultate din milioane de cărţi, inclusiv previzualizări şi text din biblioteci şi de la editori din întreaga lume.

Actualitate: afişează cele mai recente ştiri şi informaţii. Funcţia include preluarea de rezultate în timp util când căutaţi anumite date.

Google Instant: afişează rezultate imediate pe măsură ce introduceţi text.

Imagini: afişează imagini cu **miniaturi** corespunzând căutărilor, astfel încât puteţi decide dintr-o privire ce pagină doriţi să accesaţi.

Indexare: utilizează sisteme pentru colectare şi stocare documente pe internet.

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Caracteristici Google

Graful cunoştinţelor: oferă rezultate pornind de la o bază de date cu oameni, locuri şi lucruri din lumea reală şi de la conexiunile dintre acestea.

Mobil: include îmbunătăţiri concepute special pentru gadgeturile mobile: tablete şi smartphone-uri.

Ştiri: include rezultate din ziare online şi din bloguri din întreaga lume.

Înţelegerea interogării: înțelege în detaliu sensul cuvintelor introduse.

Căutare sigură: reduce numărul rezultatelor cu pagini web, imagini şi videoclipuri destinate adulților.

Metode de căutare: creează noi modalităţi de a căuta, inclusiv „căutarea după imagine” şi „căutarea vocală”.

Calitatea site-ului și a paginii: utilizează un set de semnale pentru a stabili cât de fiabilă, bine cotată şi demnă de încredere este o sursă. (Printre aceste semnale se numără PageRank, unul dintre primii algoritmi dezvoltaţi de Google, care analizează linkurile dintre pagini pentru a stabili relevanţa acestora.)

Caracteristici Google

Fragmente: afişează mici previzualizări cu informaţii, cum ar fi titlul paginii şi un scurt text descriptiv, despre fiecare rezultat al căutării.

Ortografie: identifică şi corectează erori de ortografie şi oferă alternative.

Sinonime: recunoaşte cuvintele cu sensuri similare.

Traducere : adaptează rezultatele în funcţie de limba şi de ţara unde vă aflaţi.

Căutare universală: combină conţinutul relevant, cum ar fi imaginile, ştirile, hărţile, videoclipurile şi conţinutul dvs. personal într-o pagină unificată cu rezultate ale căutării.

Context personalizat: oferă rezultate mai relevante în funcţie de regiunea geografică, Istoricul web şi alţi factori.

Videoclipuri: prezintă rezultate video cu miniaturi, astfel încât să puteţi decide rapid ce videoclip doriţi să urmăriţi.

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

b) Directoare web

Directoare web: colectie de pagini Web selectionate si organizate ierarhic în categorii de subiecte **de catre un editor uman**.

Serviciile de directoare acopera si indexeaza o portiune mult mai mica din paginile WEB existente, comparativ cu motoarele de cautare, dar poate furniza rezultate mult mai relevante pentru utilizator.

Serviciile de directoare NU interogheaza direct paginile WEB, ci cauta în interiorul bazei lor de date.

O serie de motoare de cautare = unelte hibride (motoare de cautare+servicii de directoare . Exemplu: Google™

Exemple Directoare web-Romania: http://www.dirpedia.ro/

http://www.iseo.ro
http://dir.rebelnetwork.ro
http://top-siteuri.ro
http://www.web-dir.net
http://www.director.yest.ro
http://link.radaseni.info.ro
http://www.seoportal.ro
http://www.federal.ro
http://www.director.mocka.ro
http://www.dmoz.org/World/Română/
http://zombalau.ro
http://www.diand.ro
http://director.ponturifierbinti.com
http://www.topdirector.ro
http://www.webconnect.ro
http://director.ziarulautentic.ro
http://www.1milioneuro.ro
http://www.wishbox.ro
http://www.despreagenti.ro
http://director.eul.ro
http://www.ghidulafacerii.ro
http://www.submit-url.ro
http://www.shopdirector.ro
http://www.promovarewebsite.net
http://www.dyronline.com
http://www.infodyr.com
http://www.arthzen.ro
http://www.joo.ro/director-web/
http://cue.ro
http://www.director.bia-design.ro
http://www.bia-director.ro
http://wadir.ro
http://www.director-web-romania.ro

http://cue.ro
http://www.director.bia-design.ro
http://www.bia-director.ro
http://wadir.ro
http://www.director-web-romania.ro
http://www.site-talk.ro
http://directorweb.verificare-firme.ro
http://director.pringalati.ro
http://www.wikiwebs.ro
http://www.agentia-pr.ro
http://www.seopromotion.ro
http://inscriesite.duv.ro
http://director.duv.ro
http://www.directoareweb.net
http://adsite.duv.ro
http://www.rodir.net
http://directorweb.cautatot.ro
http://www.links24.ro
http://www.amical.ro
http://www.webmaxi.ro/Director/
http://www.director-web.pro
http://www.arttouseit.net
http://www.linkpedia.ro
http://www.top90.ro
http://www.ddweblinks.eu
http://rodirectorweb.ro
http://www.ibl.ro
http://www.newbn.ro
http://www.world-directory.ro
http://www.director-web-seo.ro
http://www.top-best.ro
http://dir.ume.ro
http://www.quicklink.ro
http://www.director.coordonate.ro
http://www.will.ro
http://www.sitez.ro
http://www.linksdirector.ro

http://www.director.coordonate.ro
http://www.will.ro
http://www.sitez.ro
http://www.linksdirector.ro
http://dirweb.wink.ws
http://www.abcdinfo.ro
http://director-web-gratuit.ro
http://www.director.model-de.ro
http://www.noul.ro
http://tre.ro
http://www.director.crosmedia.ro
http://www.w5.ro/director/
http://www.ebirotic.ro
http://www.bestdirector.ro
http://directorwebseo.tk
http://www.adrese.ro
http://director.moreyou.ro
http://directorweb.moreyou.ro
http://www.top300.ro
http://www.ghidwww.ro
http://director-web.info-heaven.ro
http://director-web.bihor.ro
http://www.serviciiromania.ro
http://www.adedir.info
http://www.ghidw.com/directorlist
http://seotop.uv.ro
http://www.nettrade.ro
http://www.comercianti.ro
http://director-web.exn.ro
http://www.director.anunturiagricultura.ro
http://www.director.afix.ro
http://www.director.vanzare-masini.ro
http://www.wol.ro
http://www.webdesign-profesional.com/director/
http://www.colector.ro
http://1top.ro

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

c) Biblioteci virtuale

Biblioteci virtuale de „uz general”:

- INFOMINE (www.infomine.com)
- Internet Public Library (www.ipl.org)
- WWW Virtual Library (vlib.org)
- Academic Info (www.academicinfo.net)
- Internet Scout Project (scout.wisc.edu)
- BUBL Link -academic resources (bubl.ac.uk/link/)

Biblioteci virtuale de specialitate (pe domenii):

- Project Gutenberg (www.gutenberg.org) - beletristic: 53,000 free ebooks
- National Academies Press (www.nap.edu) - stiinta
- Free books for doctors (www.fb4d.com) - medicina
- The free management library (www.managementhelp.org) - management

d) Invisible (deep) Web

Invisible web =**colectii de informatii online stocate în baze de date accesibile pe Web**, dar care, din diferite motive, **nu sunt indexate de motoarele de cautare** traditionale. (se mai numesc specially search engines)

Exemple:

- ProFusion (www.profusion.com)
- Invisible-web (www.invisible-web.net)
- Complete Planet (www.completeplanet.com)
- Resource Discovery Network (www.rdn.ac.uk)
- Direct Search (http://www.freepint.com/gary/direct.htm)
  http://www.resourceshelf.com /

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

e) Motoare de meta-cutare (metasearch engine)

Motor de meta-cautare : instrument de cutare care trimite cerea simultan catre mai multe motoare de cautare clasice, servicii de directoare web si uneori spre colectii de tip invisible web.

Motoarele de meta-cautare :

- **nu au propriile BD web**, le folosesc pe cele ale serviciilor apelate.
- **nu permit înscrierea manuala** a unei pagini (site) în baza de date.

Exemple:

- Metasearch (www.metasearch.com)
- ez2Find (www.ez2find.com)
- Vivisimo (www.vivisimo.com)
- MetaCrawler (www.metacrawler.com)
- InfoGrid (www.infogrid.com)
- Infonetware (www.infonetware.com)
- iBoogie (www.iboogie.tv)

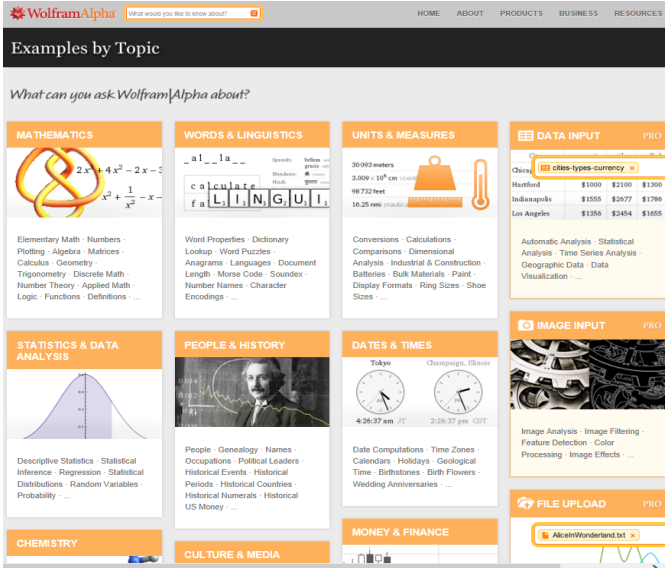
f) Instrumente cautare tip desktop

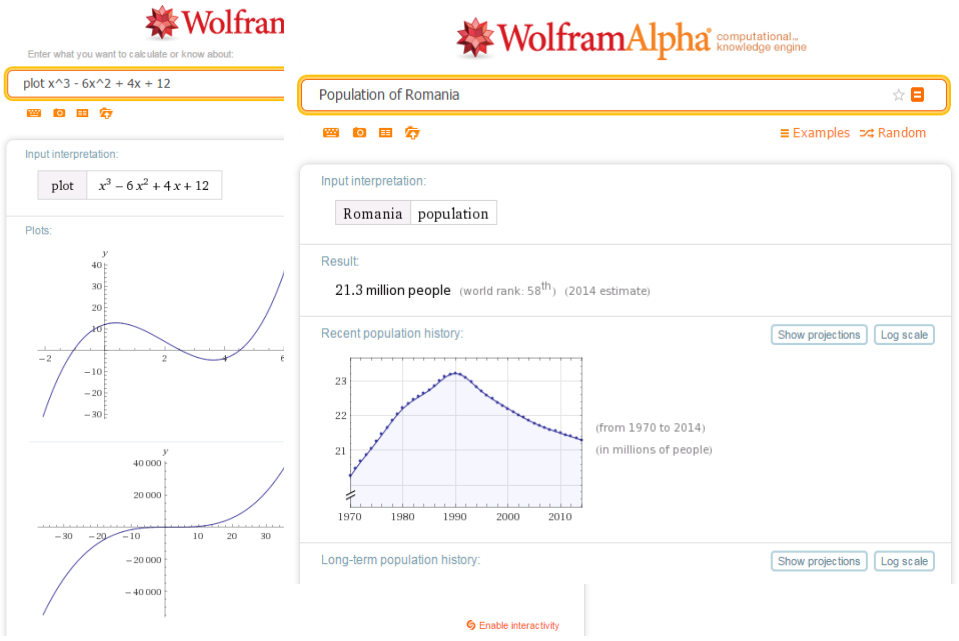
Client side search software : programe instalate local pe calculator, functioneaza similar cu motoarele de cautare

- Google Desktop (desktop.google.com)
- Copernic (www.copernic.com)
- Arrow Search (www.rt-software.co.uk/arrow_search/)
- WebFerret (www.ferretsoft.com/download.htm)
- ProtoSearch (www.npccenterprises.com/products/protosearch2.shtml)

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

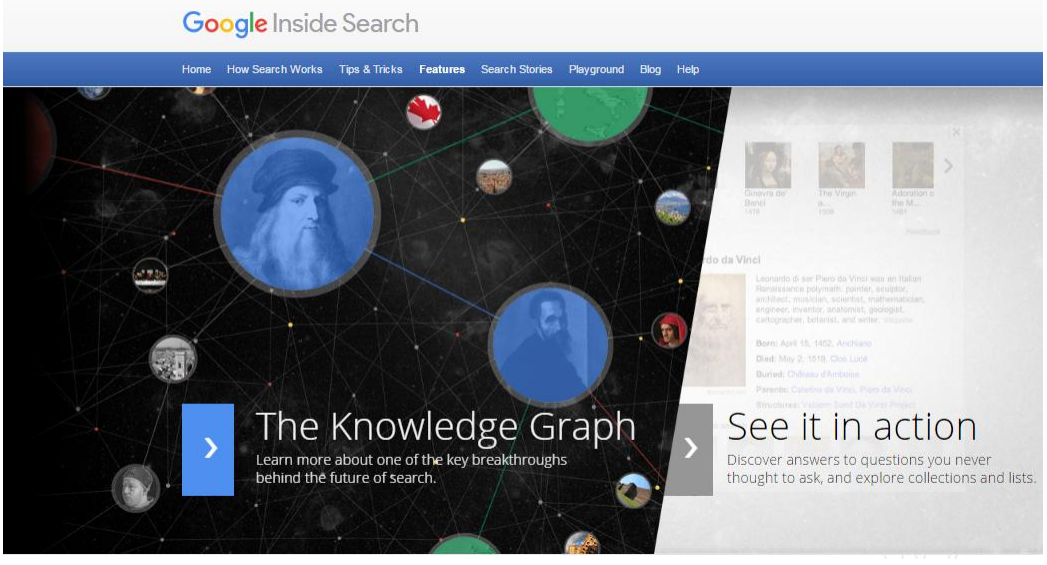
3. Cautarea avansata a informatiei: Wofram Alpha



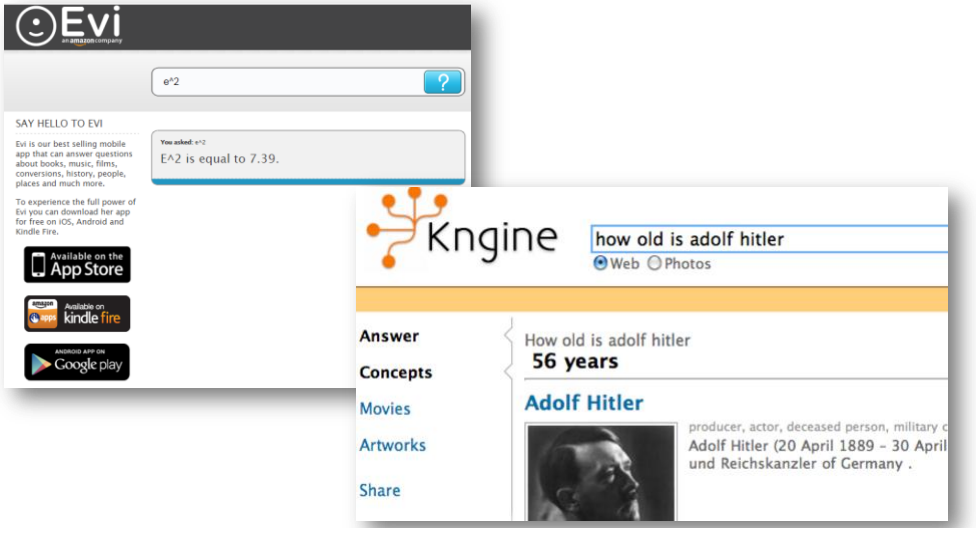


Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Cautarea avansata a informatiei: Google Knowledge Graph

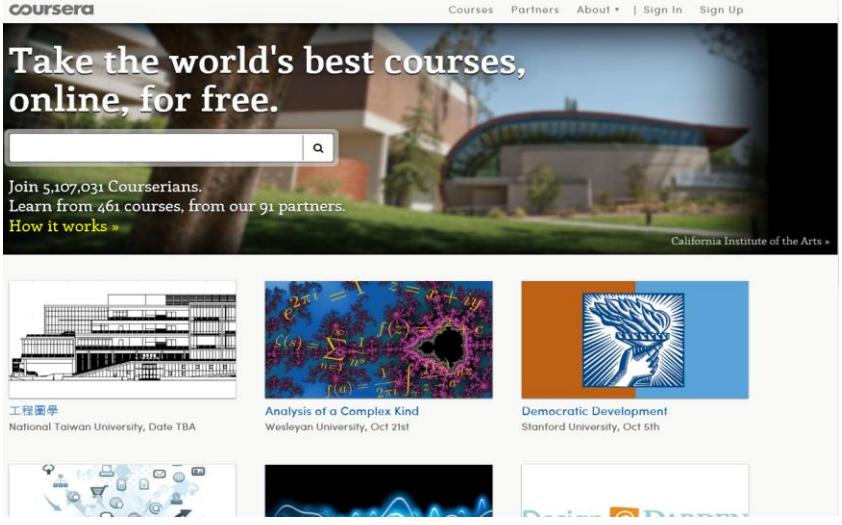


Cautarea avansata a informatiei prin apps: Evi, Kngine



Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Cautarea avansata a informatiei: Coursera



Cautarea avansata a informatiei: MIT