



Cap.9. Baze de date web

CUPRINS

1. Baze de date web. Indexare
2. Instrumente de cautare a informatiei pe web
3. Cautarea avansata a informatiei

1



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).



2. Instrumente de interogare a BD web

- a) Motoare de cautare
- b) Directoare web (anulare 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.

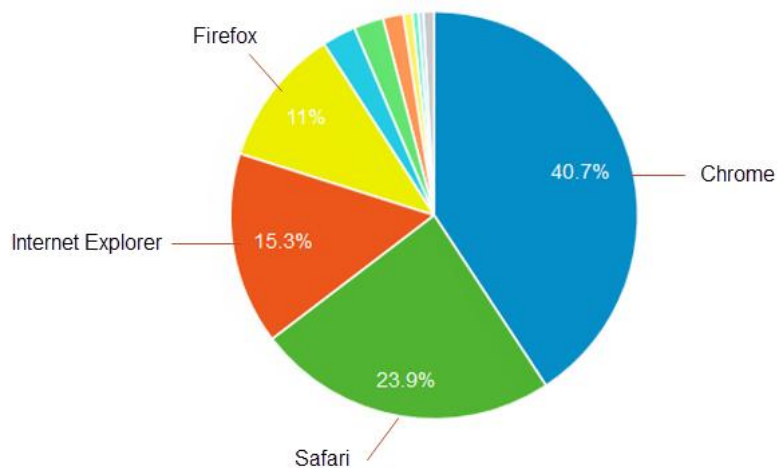
Structura unui motor de cautare :

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



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 cunoștință de cauză”
- Algoritmi de căutare cu constrângeri



Istoric metode de cautare web= metode de interogare BD web

- **pe bază de indecși – 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** (indecși + ierarhii de termeni)
 - Excite
- **pe baza legăturilor hipertext** – hyperlink analysis
 - Google (1996)
- **interogare în paralel** a mai multor motoare de căutare și 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.dominio.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

CAUTAREA ULTRA RAPIDA

- ÎNAINTE DE A CĂUTA**

Drumul unei interogări de căutare începe cu mii de trairite de introducere a căutării pe Google. Utilizăm robozi software, numiți crawlere web sau programe spider, care găsesc pagini web pentru a le include ulterior în rezultatele căutării Google. Software-ul Google stochează date despre aceste pagini în centrele de date. Webul este ca o carte cu trilioane de pagini, iar noi înlocuim această carte.

 - Indexul nostru depășește cu mult 100 de milioane de gigaocteți.
 - Până acum am petrecut peste 1 milion de ore de calcul pentru crearea indexului.
- CÂND CĂUTAȚI**

Algoritmii Google încep să găsească informațiile pe care le căutați chiar în momentul în care începeți căutarea.

 - În medie, interogarea de căutare călătorește 2400 de kilometri pentru ca răspunsul să ajungă la dvs. și poate trece prin diferite centre de date din întreaga lume în drumul său, cu o viteză egalată doar de cea a lumii, de sute de milioane de kilometri pe oră.
 - Pe măsură ce introduceți interogarea, vor începe să se afișeze predicții pentru căutări care v-ar putea interesa și rezultate, fără a trebui să apăsați pe Enter. Astfel, economisiți timp și obțineți răspunsul dorit cât se poate de rapid. **Asta numim noi Google Instant.**
- CLASAREA**

Algoritmii analizează interogarea și utilizează peste 200 de semnale pentru a alege cele mai relevante răspunsuri din milioanele de pagini și din tot conținutul. Google rafinează algoritmi săi de clasare prin peste 500 de îmbunătățiri pe an.

Exemplele de astfel de semnale includ:

 - Actualitatea conținutului de pe un site web
 - Numărul de site-uri web cu linkuri către un anumit site web și fiabilitatea acestor linkuri
 - Cuvintele din pagina web
 - Sinonime de cuvinte cheie de căutare
 - Ortografia
 - Calitatea conținutului de pe site
 - Adresa URL și titlul paginii web
 - Tipul celui mai bun rezultat: pagină web, imagine, videoclip, articol de ziari, rezultat personal etc.
 - Personalizarea
 - Rezultatele recomandate de conexiunile dvs.
- REZULTATUL**

Rezultatele sunt clasate și afișate în pagină începând cu cele mai relevante. La afișarea instantanee a rezultatelor adăugăm și o previzualizare a respectivelor pagini web, vizibilă când treci cu mouse-ul peste siglele albastre în partea din față a rezultatului. Astfel, puteți decide rapid dacă doriți să accesați site-ul respectiv. În medie, aceste **Previzualizări instant se încarcă într-o zecime de secundă.**

Mai multe statistici:

 - Pe Google au loc **miliarde** de căutări în fiecare zi.
 - Din 2003, Google a răspuns la **450 de miliarde** de interogări unice noi, căutări nemăsurabile până acum.



Cum functioneaza Google?

Strategia de regasire a informatiei in Google se bazeaza pe :

- accesarea web-ului cu **crawlere**
- și **indexarea** a miliarde de documente de pe web.

Metoda de cautare Google include:

- colectarea și organizarea informațiilor** 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) și **popularitatea site-ului** (nr. de accesari ,PageRank)?
- returnarea rezultatelor** : instantaneu



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	<ul style="list-style-type: none">• 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)



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: afișează imediat răspunsuri și informații pentru căutări cum ar fi vremea, rezultatele sportive și date pe scurt.

Completare automată: anticipează ce ați putea căuta. Înțelege chiar și termenii care au mai multe sensuri.

Cărți: găsește rezultate din milioane de cărți, inclusiv previzualizări și text din biblioteci și de la editori din întreaga lume.

Actualitate: afișează cele mai recente știri și informații. Funcția include preluarea de rezultate în timp util când căutați anumite date.

Google Instant: afișează rezultate imediate pe măsură ce introduceți text.

Imagini: afișează imagini cu **miniaturi** corespunzând căutărilor, astfel încât puteți decide dintr-o privire ce pagină doriți să accesați.

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



Caracteristici Google

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

Mobil: include îmbunătățiri concepute special pentru gadgeturile mobile: tablete și smartphone-uri.

Știri: include rezultate din ziare online și din bloguri din întreaga lume.

Înțelegerea interogării: înțelege în detaliu sensul cuvintelor introduse.

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

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

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



Caracteristici Google

Fragmente: afișează mici previzualizări cu informații, cum ar fi titlul paginii și un scurt text descriptiv, despre fiecare rezultat al căutării.

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

Sinonime: recunoaște cuvintele cu sensuri similare.

Traducere : adaptează rezultatele în funcție de limba și de țara unde vă aflați.

Căutare universală: combină conținutul relevant, cum ar fi imaginile, știrile, hărțile, videoclipurile și conținutul dvs. personal într-o pagină unificată cu rezultate ale căutării.

Context personalizat: oferă rezultate mai relevante în funcție de regiunea geografică, Istoricul web și alți factori.

Videoclipuri: prezintă rezultate video cu miniaturi, astfel încât să puteți decide rapid ce videoclip doriți să urmăriți.



b) Directoare web

Directoare web: colecție de pagini Web selectionate și organizate ierarhic în categorii de subiecte de către un editor uman.

Serviciile de directoare acoperă și indexează o porțiune mult mai mică din paginile WEB existente, comparativ cu motoarele de căutare, dar poate furniza rezultate mult mai relevante pentru utilizator.

Serviciile de directoare NU interoghează direct paginile WEB, ci caută în interiorul bazei lor de date.

O serie de motoare de căutare = unelte hibride (motoare de căutare+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.seoport.ro>
<http://www.federal.ro>
<http://www.director.mokka.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.despreagentii.ro>
<http://director.eu1.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://rodiretorweb.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.coordinate.ro>
<http://www.will.ro>
<http://www.sitez.ro>
<http://www.linksdirector.ro>

<http://www.director.coordinate.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://ltop.ro>



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/>



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)



3. Cautarea avansata a informatiei: Wofram Alpha

The screenshot shows the WolframAlpha website interface. At the top, there is a search bar with the text "What about you like to know about?". Below the search bar, there are navigation links: HOME, ABOUT, PRODUCTS, BUSINESS, and RESOURCES. The main content area is titled "Examples by Topic" and "What can you ask WolframAlpha about?". It features a grid of topic-based search examples:

- MATHEMATICS**: Includes a search for $2x^2 - 4x^2 - 2x - 1$ and $x^2 + \frac{1}{x^2} - x$. Sub-topics include Elementary Math, Numbers, Plotting, Algebra, Matrices, Calculus, Geometry, Trigonometry, Discrete Math, Number Theory, Applied Math, Logic, Functions, and Definitions.
- WORDS & LINGUISTICS**: Includes a search for "al...la...". Sub-topics include Word Properties, Dictionary, Lookup, Word Puzzles, Anagrams, Languages, Document Length, Morse Code, Soundex, Number Names, Character Encodings, etc.
- UNITS & MEASURES**: Includes a search for "3099 meters". Sub-topics include Conversions, Calculations, Comparisons, Dimensional Analysis, Industrial & Construction, Batteries, Bulk Materials, Paint, Display Formats, Ring Sizes, Shoe Sizes, etc.
- DATA INPUT**: Includes a search for "cities types-currency x". Sub-topics include Automatic Analysis, Statistical Analysis, Time Series Analysis, Geographic Data, Data Visualization, etc.
- STATISTICS & DATA ANALYSIS**: Includes a search for a normal distribution curve. Sub-topics include Descriptive Statistics, Statistical Inference, Regression, Statistical Distributions, Random Variables, Probability, etc.
- PEOPLE & HISTORY**: Includes a search for "Tokyo". Sub-topics include People, Genealogy, Names, Occupations, Political Leaders, Historical Events, Historical Periods, Historical Countries, Historical Numerals, Historical US Money, etc.
- DATES & TIMES**: Includes a search for "4:26:37 am JT". Sub-topics include Date Computations, Time Zones, Calendars, Holidays, Geological Time, Birthstones, Birth Flowers, Wedding Anniversaries, etc.
- CHEMISTRY**: Includes a search for "AlcainWunderland.txt".
- CULTURE & MEDIA**: Includes a search for "AlcainWunderland.txt".
- IMAGE INPUT**: Includes a search for "Image Analysis". Sub-topics include Image Filtering, Feature Detection, Color Processing, Image Effects, etc.
- FILE UPLOAD**: Includes a search for "AlcainWunderland.txt".



Enter what you want to calculate or know about:

plot $x^3 - 6x^2 + 4x + 12$

Input interpretation: plot $x^3 - 6x^2 + 4x + 12$

Plots:

Population of Romania

Input interpretation: Romania population

Result: 21.3 million people (world rank: 58th) (2014 estimate)

Recent population history: (from 1970 to 2014) (in millions of people)

Long-term population history:



Cautarea avansata a informatiei: Google Knowledge Graph

Google Inside Search

Home How Search Works Tips & Tricks Features Search Stories Playground Blog Help

The Knowledge Graph

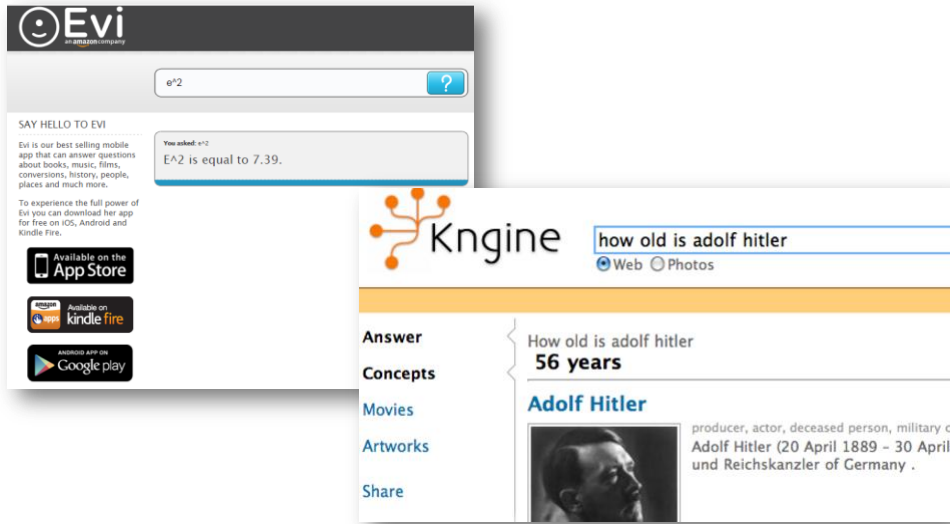
Learn more about one of the key breakthroughs behind the future of search.

See it in action

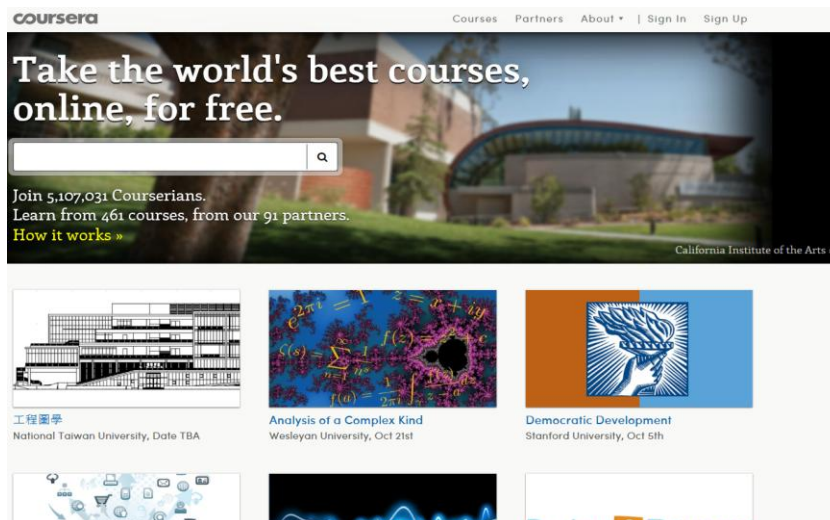
Discover answers to questions you never thought to ask, and explore collections and lists.



Cautarea avansata a informatiei prin apps: Evi, Kngine



Cautarea avansata a informatiei: Coursera





Cautarea avansata a informatiei: MIT

MIT OPENCOURSEWARE 15 YEARS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Subscribe to the OCW Newsletter

Home | Contact Us

Find Courses | About | Donate | Featured Sites | Search | Advanced Search

Cellular Solids: Structure, Properties and Applications

» New lecture videos

Support OCW

This course material is free and of prime quality™

Francesco Independent Learner Italy

DONATE NOW

FEATURED COURSES

» Find Courses

- EDITORS PICK**
Space Propulsion
- EDUCATOR**
Ethics in Your Life: Being, Thinking, Doing (or Not?)
- NEW**
Advanced Topics in Hispanic Literature and Film: The Films of Luis Buñuel
- VIDEO**
String Theory and Holographic Duality

OCW makes the materials used in the teaching of MIT's subjects available on the Web.

Get Started