

Deel 5: Netwerken

***E**en digitaal systeem staat zelden alleen. Net zoals ieder van ons rond zich een sociaal netwerk van vrienden, kennissen, collega's, bekenden verzamelt, zo werkt een computer zelden in zijn "eentje".*

Computers kunnen in een netwerk samenwerken en uitgroeien tot één supercomputer door hun rekenkracht te combineren zoals bij het renderen van een 3D-film gebeurt. Ze kunnen ook hun opslagruimte combineren zoals bij big data of zoekmachines het geval is.

Het meest bekende netwerk is het internet, waarvan het web (http), e-mail (mailto), chat... de bekendste toepassingen vormen. Het internet of things moet in de nabije toekomst alle digitale apparaten via het internet informatie laten uitwisselen.

In dit deel leer je niet meteen hoe je netwerken opstelt, maar leer je iets dieper en vooral grondiger graven in het internet. Je leert sneller en beter informatie verzamelen en je leert hoe zoekmachines functioneren.

Waarom gaat zoeken in Google zoveel maal sneller dan een simpel bestand terugvinden op je harde schijf? Dat zou een fijne examenvraag zijn waarbij je de informatie uit het hoofdstuk over bestandssystemen en uit dit deel moet vergelijken.

5.1 De werking van het internet

5.1.1 Clients en servers: de klanten en hun obers

Het internet kan je niet zo maar vastpakken. **Waaruit bestaat het** precies, welke onderdelen maken deel uit van het “internet”? In de eerste plaats bestaat het uit een **bonte verzameling hardware**: routers, kabels, servers, masten voor mobiele telefonie, satellieten in een baan om de aarde, radio's, mobiele telefoons, computers.... Ze vormen van een wereldwijd netwerk van netwerken. Het internet is flexibel: als je onderdelen verwijdert of toevoegt, ondervindt het naar nauwelijks hinder van of aanpassingsmoeilijkheden. Een router verwijderen of een kabel uittrekken legt niet het ganse internet plat. Het lijkt in dat geval meer op een speldenprik in je vingertopje.

Bepaalde onderdelen zijn relatief statisch en vormen de ruggengraat van het internet. Zo'n onderdelen kan je niet zonder problemen overboord gooien. Je computer, mobiele telefoon... vormen eindpunten in het internet. Men spreekt in dit geval vaak van “**clients**” die door “**servers**” worden bediend, zoals een restaurantbezoeker door een ober wordt bediend. Een klant (client) bestelt eten en de ober levert wat later de gevraagde schotel. Je browser op je computer vormt een client die een webpagina opvraagt en de webserver stuurt even later de gevraagde webpagina naar de browser van de “bezoeker”.

Als je een pakje bestelt bij een postorderbedrijf, dan legt je pakje een hele weg af via het bedrijf dat het artikel produceert, de transportmaatschappij, het postkantoor, de postbode... vooraleer het pakje bij de “client” is. Al die tussenschakels vindt je ook terug op het internet, men noemt het “**nodes**”. Tussen een server en een client liggen op die manier één of meerdere nodes die het verzenden en uitwisselen van “data” **pakketjes** mogelijk maken. **De server(s), nodes en clients zijn met elkaar verbonden**. Die verbindingen kunnen fysiek (tastbaar) zijn in de vorm van “kabels”, maar ook draadloos (via satellieten, radioverbindingen, GSM-masten...)

5.1.2 Goede afspraken maken goede vrienden: protocollen

Om heel het verkeer vlot te laten verlopen zijn er verkeersregels nodig, afspraken tussen alle gebruikers, servers, nodes, clients. Enkel duidelijke afspraken maken een vlot verkeer mogelijk. Die regels en afspraken zijn vastgelegd in een reeks **protocollen**. Zoals het verkeersreglement uitlegt dat je pas mag doorrijden met je auto als het licht op groen springt en je moet rekening houden met voorrangregels, zo leggen **communicatieprotocollen** vast op welke manier een vlotte uitwisseling van informatie tussen servers, nodes en clients mogelijk wordt. Protocollen bepalen niet enkel regels, maar ook een reeks te volgen stappen om bepaalde taken uit te voeren. Alle toestellen die op het internet zijn aangesloten, moeten die protocollen volgen. Als iedereen zo maar zijn zin doet, alle toestellen hun eigen regels volgen en hun eigen taaltje spreken, dan is een vlot verkeer niet mogelijk. Een protocol bepaalt dus ook de code (taal) waarin toestellen met elkaar communiceren, net zoals iedere weggebruiker de pictogrammen op de verkeersborden moet begrijpen om te weten welke regels hij in elke verkeerssituatie dient te volgen.

Huizen hebben een adres. Dit zorgt ervoor dat de postbode de brieven makkelijk bij de juiste personen kan afleveren. Op het internet hebben alle clients, servers en nodes eveneens een adres. Indien dit niet het geval zou zijn, zou het uitwisselen van informatie tussen een client en een server behoorlijk rampzalig worden. Een server

zou niet meer weten aan welk toestel hij de gevraagde informatie moet sturen. Het protocol dat zorgt voor de verdeling van die adressen, heet het **Internet Protocol** of kortweg **IP**, wat je zeker kent van de veelgebruikte term IP-adres. Het protocol dat zorgt voor de verzending van informatie tussen diverse adressen het het **Transmission Control Protocol of TCP**. Beide protocollen worden vaak in één adem uitgesproken als TCP/IP. Sommige servers (DNS-servers, of **Domain Name Servers**) nemen de rol van adressenlijst en richtingaanwijzer op zich. Ze sturen elke vraag die binnenkomt door naar het gepaste adres. Hoe meer toestellen op het internet worden aangesloten, hoe meer adressen nodig zijn. Net dat vormt een probleem. Elk adres bestaat uit een reeks cijfers en cijfers worden door digitale systemen binair gecodeerd. De vierde versie van IP gebruikte hiervoor een systeem van 32 bits. Dat voldeed geruime tijd en kon voldoende verschillende adressen leveren voor elk toestel. Door het toenemende aantal op het internet aangesloten toestellen (mobiele toestellen, webserver, internet of things...) raakte het aantal beschikbare adressen uitgeput. Het Internet Engineering Task Force (stel het je niet voor als een reeks Amerikaanse superhelden) besliste reeds in 1991 om IP te vernieuwen naar een **adresseringssysteem** van 128 bits: dit vernieuwde IP-protocol kreeg de welluidende naam IPv6 (versie 6).

Bovenop de TCP/IP-laag stoeit men met nog een reeks andere protocollen. Ongetwijfeld kan je **HTTP**, het **hypertext transfer protocol** dat een reeks regeltjes bevat voor het uitwisselen van webpagina's en data tussen webserver en browsers. Het world wide web (www) maakt gebruik van dit protocol. Het wereldwijde web bestaat uit een gigantische reeks websites en webpagina's. Websites vraag je op door een **domeinnaam** of **URL** (uniform resource locator) in te voeren in je browser. Een URL bestaat uit een reeks onderdelen:

<i>Protocol</i>	<i>domein</i>	<i>domeinnaam</i>	<i>extensie</i>	<i>Map, bestand (eventueel)</i>	<i>Map, bestand (eventueel)</i>
http://	www. , users. , ...	blabla	.com, .be ...	/mapnaam /index.html	/bestandsnaam.html /nogEenMap

Op één server (met één enkel IP-adres) kunnen zich meerdere websites bevinden. Op één IP-adres kunnen zich dus meerdere domeinen bevinden, net zoals in één appartementsgebouw meerdere gezinnen (personen met "namen") kunnen wonen.

Een webontwikkelaar die op zijn computer een website bouwt, verstuurt alle gemaakte webpagina's naar de webserver. Hij doet dit met een stuk software dat snel bestanden kan uitwisselen met een webserver via het **FTP-protocol (File Transfer Protocol)**. Uiteraard kan niet iedereen zo maar een website aanpassen. De webontwikkelaar heeft hiervoor een gebruikersnaam en wachtwoord nodig, net zoals bij het MAILTO-protocol (je weet wel: e-mail).

5.1.3 Packet switching: uitwisselen van informatie tussen computers

Wanneer je http://www.schoolvoorbeeld.be/Multimedia_28.html invoert in je browser, stuurt je computer een aanvraag (request) naar je Internet Service Provider of ISP (Telenet, Belgacom...). De ISP stuurt die aanvraag door naar bijvoorbeeld een DNS (domain name server) die nakijkt bij welk IP-adres de domeinnaam thuishoort. Als hij een overeenkomst vindt (er bestaan ook domeinnamen die nog geen "plaatsje" hebben op een IP-adres, je zou in dit geval van dakloze domeinnamen kunnen spreken), stuurt hij de aanvraag door naar de webserver die zich op dat IP-adres bevindt. De webserver stuurt het gevraagde bestand (en eventueel gekoppelde bestanden zoals afbeeldingen, stijldocumenten enz.) terug naar de computer (browser) die erom gevraagd heeft. Die pakketjes volgen niet noodzakelijk dezelfde route als de "request". Een response (antwoord) kan een totaal

andere weg volgen, afhankelijk van welke verbindingen (telefoonlijnen of glasvezelkabels) er op dat moment vrij zijn (de weg met de minste file). Soms wordt een pakketje zelfs opgesplitst in meerdere brokjes om bij de eindbestemming weer te worden samengevoegd. Het lijkt wel alsof de ontvanger alle puzzelstukjes weer moet samenvoegen. Die techniek van **packet switching** werd reeds in de jaren 1960 ontwikkeld. Tom Igoe legt het in zijn boek “Making things Talk” als volgt uit:

“Vergelijk het met het opsturen van een fiets naar een ander adres. De fiets is veel te groot om in één pakje te verzenden via de post. Daarom haal je hem eerst uit elkaar zodat elk onderdeel in een kleine doos past. Op een netwerk (...) wordt ieder bericht in kleine stukjes gebroken van min of meer dezelfde grootte. Elk stuk krijgt een pakketnummer. Daarna schrijf je het adres van de bestemming en de afzender op elke doos. Daarna verzend je het. De koeriersdienst verspreidt je pakjes misschien over meerdere vrachtwagens als dat zo beter past. Op het internet gebeurt dit in de transportlaag. Die laag is verantwoordelijk voor het verzenden van alle pakjes naar hun bestemming. (...) Elke router stuurt de pakjes één voor één naar de routers waarmee hij verbonden is. Als hij met meerdere routers verbonden is, kiest hij de minst drukke. Elk pakje kan een andere weg naar zijn bestemming volgen. De ontvanger leest de header-gegevens en voegt alle pakjes weer samen.”

5.2 Van hypertext tot het web

Toen ik zelf nog op de schoolbanken zat, raakte ik snel afgeleid door het simpelweg lezen van een woord in een oefening. Ik herinner me dat ik in het vak Nederlands ooit een zin over een astronaut moest verdelen in zinsdelen. In mijn hoofd zweefde ik ondertussen al ergens halfweg tussen de aarde en de maan. Dat is exact wat we in ons hoofd doen: we leggen analogieën en volgens sommige onderzoekers is dat ook de manier waarop onze hersenen informatie opslaan.

5.2.1 Wat is hypertext en hypermedia?

Veel leraars zouden in dat geval zeggen dat hun leerlingen te snel zijn afgeleid. Het leggen van verbanden tussen inhoud, teksten, fragmenten van teksten... is wat iedereen helemaal gewend is door de komst van het web. Op webpagina's vind je niet simpelweg “tekst”, maar “**hypertext**”, **tekst die verbanden legt met andere teksten of media**.

In veel gevallen legt de webontwikkelaar de links zelf, maar het kan ook geautomatiseerd gebeuren zoals bijvoorbeeld in een zoekmachine zoals Google of in de online encyclopedie Wikipedia (zie afbeelding hieronder).

Enkele begrippen [bewerken]

Hypertext [bewerken]

Een hypertext is een welbepaalde (maar meestal niet onveranderlijke) verzameling hypertext. Bekende voorbeelden:

- Wikipedia
- de Internet Movie Database
- Amazon, eBay, of een willekeurige andere webapplicatie of website
- Windows Help-handleidingen van Microsoft Windows-applicaties
- GNU info-handleidingen van GNU-applicaties

Hypertextsysteem [bewerken]

Een hypertextsysteem is een softwaresysteem voor het maken en gebruiken van hypertext. Bekende voorbeelden:

- Mediawiki (waar Wikipedia in wordt gemaakt)
- Windows Help met bijvoorbeeld de Microsoft Windows Help Workshop
- GNU info met bijvoorbeeld texinfo
- softwaredocumentatie-hypertextgeneratoren zoals doxygen, Javadoc en NDoc

Hypertextstandaard [bewerken]

Een hypertextstandaard is een verzameling technische afspraken voor het representeren en gebruiken van hypertext. De met grote afstand meest bekende standaard op dit gebied is het World Wide Web, dat bestaat uit drie onderdelen:

- HyperText Markup Language (HTML), een opmaaktaal voor hypertext;
- de Uniform Resource Locator (URL), een documentadresmethode voor het Internet;
- het HyperText Transfer Protocol (HTTP), een protocol voor het opvragen en oversturen van documenten op het Internet.

Geohiedenis - ...

Ooit zei iemand me:

“Heel die multimedia en dat internet, het is eigenlijk niet meer dan

'klikken' en nog eens 'klikken'."

Gedeeltelijk klopt zo'n uitspraak ook wel. Hypertekst en bij uitbreiding "hypermedia" vormt de basis van multimedia. Tekst, afbeeldingen, film, geluid... **wordt aan elkaar gelinkt al dan niet via voor de gebruiker zichtbare links.**

Oneerbiedig zouden we onszelf de "zap" generatie kunnen noemen. *Eén belangrijk punt van kritiek* moeten we meenemen in het multimedieverhaal: **te veel koppelingen leiden te erg af van de kern van de zaak.** Leg enkel zichtbare koppelingen als die bijdragen tot een betere multimedia-ervaring.

Voorbeeld

Een voorbeeld van de HTML-code van een webpagina: alle media (afbeeldingen, films, audio, stijlbestanden, scripts...) zijn gelinkt. **Ze zitten niet in het bestand, maar zijn extern bewaard.**

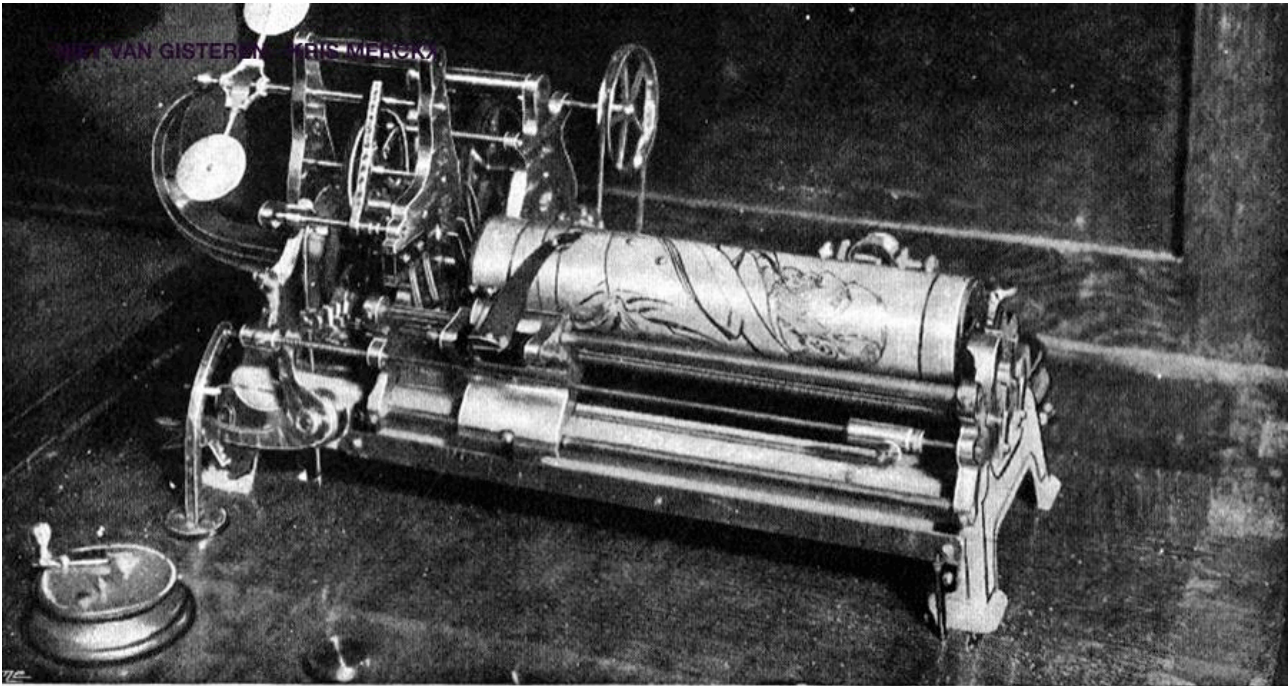
```
<!doctype html>
<html lang="nl-be">
  <head>
    <title>Wobbel</title>
    <meta name="author" content="Kris Merckx"/>
    <meta charset='utf-8'>
    <script src="http://ajax.googleapis.com/ajax/libs/jquery/2.1.1/jquery.min.js"></script>
    <link href="http://fonts.googleapis.com/css?family=Open+Sans+Condensed:300,700,300italic" rel='stylesheet' type='text/css'>
    <link rel="stylesheet" href="http://ajax.googleapis.com/ajax/libs/jqueryui/1.11.1/themes/smoothness/jquery-ui.css" />
    <script src="http://ajax.googleapis.com/ajax/libs/jqueryui/1.11.1/jquery-ui.min.js"></script>
    <link rel="stylesheet" href="http://ajax.googleapis.com/ajax/libs/jqueryui/1.11.1/themes/smoothness/jquery-ui.css" />
    <link rel="stylesheet" href="wobbel/basis_links.css" />
    <script src="wobbel/basis_links.js"></script>
  </head>
  <body>
    <h1>Een link</h1>
    <a href="http://www.blabla.be" title="een link">Blabla</a>
    <h1>Een afbeelding</h1>
    
    <h1>Een video</h1>
    <video width="320" height="240" controls="">
      <source src="http://www.fotografiejoel.be/film.mp4" type="video/mp4">
      Uw browser ondersteunt de weergave van mp4-film niet.
    </video>
    <h1>Een audiofragment</h1>
    <audio width="320" height="100" controls="">
      <source src="http://www.hetklikt.be/songs/files/beatit2.mp3" type="audio/mp3">
      Uw browser ondersteunt de weergave van mp3 niet.
    </audio>
```

5.2.2 Teksten mondiaal verspreiden

Paul Otlet, de Belgische uitvinder van het analoge internet

<RANDINFO>

De telegraaf was het SMS- en mailsysteem van de 19^e eeuw. Zowel qua succes als qua impact op de samenleving kende de telegraaf vergelijkbare successen. Met als belangrijkste verschilpunt dat niet iedereen een telegraaf in huis had en "mobiele" telegrafie bestond nog niet. Voor het eerst werd het dankzij de telegraaf (die al 'optische' voorlopers kende in de 17^e en 18^e eeuw) mogelijk om flitsend snel berichten uit te wisselen over de ganse wereld. Visionaire denkers zagen meteen de toekomstperspectieven die het uitwisselen van informatie via elektrische signalen mogelijk maakte: ooit zou je niet alleen tekst, maar ook beeld en geluid via een "telegraaf" kunnen verzenden.



The transmitter in operation at the *New York Herald* office, showing Croker being telegraphed to the *Chicago Times Herald*, 1000 miles away.

Otlet, één van die Belgen die de hemel ingeprezen wordt in de US en de UK. Hij wordt niet zonder reden de vader van het “analoge” internet genoemd. Alex Wright, professor aan de universiteit van Oxford wijdde in 2014 nog een boek aan Otlet.

De moeder van Paul Marie Ghislain Otlet (1868 – 1944) stierf op zijn derde en tot zijn twaalfde kreeg hij thuisonderwijs waar hij opgroeide als een eenzame boekenwurm. De bibliotheek zou levenslang zijn favoriete omgeving zijn. Niet voor niets ontwierp hij de Universele Decimale Classificatie (UDC) voor het klasseren van boeken. Samen met nobelprijswinnaar Henri La Fontaine (1854 – 1943) vatte hij in 1895 het plan op een bibliografie te schrijven van alle wereldwijd gepubliceerde kennis die hij vervolgens ook publiek beschikbaar wilde maken. Ze beperkten zich niet alleen tot boeken, maar namen ook artikels, magazines, afbeeldingen, pamfletten... media die door bibliotheken in die tijd genegeerd werden, in hun lijsten op. Op die manier bouwden ze een databank op met meer dan 12 miljoen steekkaarten. Van de overheid kregen ze toelating om de databank onder te brengen in overheidsgebouwen te Brussel. Dankzij de overheidssteun konden ze ook personeel in dienst nemen. Via brief of een telegraafverbinding konden mensen van waar ook ter wereld informatie opvragen. Deze analoge zoekmachine kreeg meer dan 1500 vragen per jaar. Deze analoge zoekmachine kreeg de naam “**Mundaneum**”. In zijn boeken “*Traité de documentation* (1934) en “*Monde: Essai d’universalisme*” uit 1935 schreef hij over de mogelijkheid van een mechanisch collectief geheugen dat alle informatie zou bevatten en toegankelijk moest zijn via telecommunicatie. Hij droomde van een netwerk (“réseau”) van “computers” (hij noemde het “elektrische telescopen”) waarin mensen informatie konden zoeken en bladeren door miljoenen gelinkte documenten, afbeeldingen, geluiden en filmfragmenten. In 1934 zei de overheid haar medewerking en steun op. De doodsteek voor dit visionaire Mundaneum kwam door de Duitse invasie. Duitsers gebruikten de originele ruimte als tentoonstellingsruimte voor kunst van het “Derde Rijk” en vernietigden duizenden steekkaarten. Paul Otlet stierf als een gebroken man.

Paul Otlet stond niet alleen met zijn ideeën. De tijd was er rijp voor. De Amerikaanse wetenschapper **Vannevar Bush** (1890- 1974) liep met vergelijkbare ideeën rond. In zijn essay “*As We May Think*” beschreef hij de **Memex (Memory Extender)**, een toestel dat op basis van microfilms alle belangrijke gegevens zou opslaan. In zijn concept moest de informatie zo worden georganiseerd dat informatie snel zou kunnen worden opgeroepen en geraadpleegd. Het zou leiden tot nieuwe encyclopediesystemen met een “vlechtwerk van

associatieve sporen” die je met de Memex zou kunnen uitlezen. Ted Nelson (1937) startte in 1960 met een gelijkaardig concept op basis van computertechnologie dat hij **Project Xanadu** doopte, maar veel verder dat een plan is het nooit gekomen.

</EINDE RANDINFO>

5.2.3 De komst van het digitale internet

De voorganger van het wereldwijde web dat we nu kennen was ARPAnet of het “Advanced Research Projects Agency Networks”. ARPAnet was opgezet door het Amerikaanse leger dat in volle Koude Oorlog op zoek was naar een manier om te kunnen blijven communiceren in geval van een nucleaire oorlog. Met dit doel voor ogen werd de TCP/IP-communicatiestandaard ontwikkeld die bepaald hoe informatie binnen een computernetwerk moet worden uitgewisseld. Internet is in feite niet meer dan een verzameling computers die via telefoonkabels informatie uitwisselen. Gaandeweg werden steeds meer 'netwerken' van computers met elkaar verbonden. Hierdoor is het internet een netwerk van duizenden netwerken geworden. Om die computers en netwerken met elkaar te kunnen verbinden, moesten ze ook softwarematig met elkaar kunnen spreken. Hiervoor werd het TCP/IP-protocol ontwikkeld. Elk netwerk en elke computer binnen dat netwerk krijgt een IP-adres (een internet protocol-'adres' bestaande uit een aantal cijfers). Netwerken of computers met een vast (statisch) IP-adres zijn voor alle computers van het internet te bereiken. Zij dienen als server. Binnen een netwerk krijgen computers vaak een dynamisch (= op regelmatige tijdstippen wijzigend) IP-adres. Zulke toestellen dienen als client.

5.2.4 Weven op het web

Door de opkomst van het internet sinds de zestiger jaren van de vorige eeuw, ontstonden allerlei mogelijkheden voor het uitwisselen van informatie zoals e-mail, IRC (internet relay chat), nieuwsgroepen, bulletin boards waar vooral professionele computergebruikers en computernerds en hobby'isten gebruik van maakten. De doorbraak van een wereldwijd informatiesysteem kwam tussen 1989 en 1991. Tim Berners Lee was actief aan het CERN in Geneve. Op zoek naar een manier om eenvoudig informatie te kunnen delen met andere academici en wetenschappelijke instuteren bedacht hij het wereldwijde web. **De hypertexttaal HTML (hypertext markup language) moest een standaard worden voor documenten.** Een muisklik op een link in een document, leidt de gebruiker naar een ander HTML-document. Voor het uitwisselen van HTML-documenten tussen computers werd het **HTTP-protocol (hypertext transfer protocol)** uitgewerkt. Tim ontwierp ook de eerste webbrowser, een programma waarmee je webpagina's online kan bekijken. Hij doopte zijn systeem **World Wide Web, kortweg www.** Al snel sprongen ook gewone mensen en bedrijven op de trein door hun eigen websites (een verzameling webpagina's) te publiceren.

5.3 Computers begrijpen jou niet

Als je een tekst invoert op je computer of een tekst publiceert op internet, weet je zelf waarover die tekst gaat. Computers kennen de inhoud niet, of beter gezegd, ze begrijpen absoluut niet waarover je het hebt. Ze snappen de betekenis niet. Voor een computer of stuk software is een tekst niet meer dan een reeks karakters die op hun beurt nog eens worden vertaald in bits (nullen en enen, je weet wel).

Wanneer een docent je op een examen een vraag stelt, dan kan je daar in het beste geval vlot op antwoorden. Je begrijpt de vraag, maakt in je hoofd de juiste analogie en vergelijkingen en je kan de inhoud simultaan vertalen in je eigen woorden. Je leert het niet uit je hoofd en je rammelt het niet woord voor woord betekenisloos af. Dat zou je natuurlijk kunnen doen, maar een docent verwacht dat je intrinsiek verbanden kan leggen, de betekenis snapt.

Wanneer je in de zoekmachine Google de zoekterm “haar” invoert, dan moet je niet meteen verwachten dat Google het dichtstbijzijnde kapsalon op het scherm tovert. “Haar” heeft meerdere betekenissen (vb. een bezittelijk voornaamwoord). Meer nog, je voert zoektermen in omdat je beseft dat Google nog minder raad zou weten met de vraag “Wat is het dichtstbijzijnde kapsalon?” Google herkent geen natuurlijke taal, zoals zo goed als geen enkel softwareprogramma eigenlijk. *Daarom moet je ook niet lachen met Google Translate als hij geschifte vertalingen oplevert. Het is eerder verbazingwekkend te noemen wat “Google Translate” wel kan.*

Google en andere zoekmachines snappen totaal niet waarover de inhoud van die miljarden webpagina's op het wereldwijde web gaat.

5.3.1 Hoe doorzoeken software en zoekmachines teksten?

In de meeste programma's kan je met de toetsenbordcombinatie CTRL+F een bepaald woord zoeken. Wanneer je op “zoeken” klikt of op de ENTER-toets, controleert de software de tekst woord per woord (woorden vindt hij door van spatie naar spatie te springen). Wanneer hij een woord vindt dat overeenkomt met jouw zoekterm, licht de software het woord op en vertelt je over het algemeen ook hoeveel overeenkomsten hij heeft gevonden.

*Een tagcloud tovert de meest voorkomende woorden in een tekst om in een grafische weergave. Hoe vaker een bepaald woord wordt gevonden, hoe groter het wordt afgebeeld.*⁴⁸



⁴⁸ Afbeelding van <http://www.mickcowdanger.dds.nl/easywareSite/wp-content/uploads/2012/01/tagcloud.png>, geraadpleegd op 18 september 2014.

Zo dom zijn computers nu ook weer niet. Dankzij “**reguliere expressies**” kan software ook zoeken naar *bepaalde repititieve patronen* in een tekst. Je zou bijvoorbeeld kunnen vragen om alle internetlinks in een document terug te vinden. Dat vraagt al wat ingewikkelder algoritmes om dit te herkennen. De software moet zowel *www.honden.be* als *users.telenet.be/kapsalonmarijke* en *poesjemauw.interessantesites.nl* terugvinden.

Reguliere expressies zijn zelf voor doorwinterde programmeurs moeilijk om te programmeren.

Een voorbeeld van een reguliere expressie om links terug te vinden:

```
/(https?:\V)?([\da-z\.-]+)\.([a-z\]{2,6})([\Vw \.-]*)*\/?$/
```

Oeps, dat ga je zeker niet op 1-2-3 leren. Besturingssystemen bouwen soms een gebruiksvriendelijke schil over de zoekfunctie, maar hierdoor gaan een aantal mogelijkheden wel verloren.

Als je bijvoorbeeld alle afbeeldingen op een Windowscomputer wil zoeken, dan kan je in de zoekfunctie dit simpelweg aanduiden. Wil je echter enkel JPG-afbeeldingen, dan helpt een instructie als:

```
*.jpg
```

De zoekmachine Google maakt massaal gebruik van reguliere expressies om de teksten op het wereldwijde web te doorzoeken. Je kan ze ook voor een stuk zelf gebruiken via gebruiksvriendelijke instructies of commando's⁴⁹:

De instructie

```
allinurl:hond + ext:jpg
```

levert afbeeldingen die in hun bestandsnaam of url het woord “hond” bevatten, ook als er bijvoorbeeld “honden” of “hondenkapsalon” staat.

Google SafeSearch is uit

Web Afbeeldingen Video's Nieuws Maps Boeken

Ongeveer 43 resultaten

Elk land
Land: België

Elke taal
Pagina's geschreven in het Nederlands

Elke periode
Afgelopen uur
Afgelopen 24 uur
Afgelopen week
Afgelopen maand
Afgelopen jaar

Alle resultaten
Woord voor woord

[Afbeeldingen van allinurl:hond + ext:jpg](#)

[File:Guat-Hond-border.jpg - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/File:Guat-Hond-border.jpg
This is a candidate to be copied to Wikimedia Commons. Freely licensed or public domain media are more accessible to other Wikimedia projects if placed on ...

[File:DM-hond, lopen.jpg - Wikimedia Commons](#)
commons.wikimedia.org/wiki/File:DM-hond,_lopen.jpg
Date, 28 December 2008. Source, Transferred from nl.wikipedia (Original text : eigen foto). Author, R. Hille at nl.wikipedia. Permission (Reusing this file).

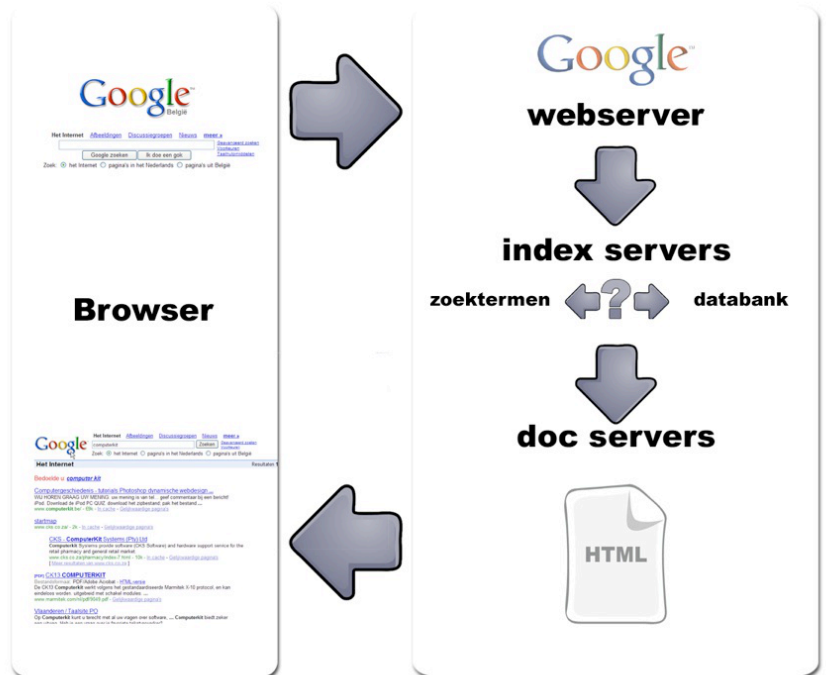
[File:Hond.jpg - Wikimedia Commons](#)
commons.wikimedia.org/wiki/File:Hond.jpg
7 Mar 2014 ... DescriptionHond.jpg. Jonge hond, foto is van Faragon zelf en valt onder de

Reguliere expressies worden ook gebruikt in webpagina's om de invoer van een gebruiker te controleren, bijvoorbeeld om te controleren of het door jouw ingevoerde e-mailadres wel een e-mailadres is en niet een hoop onzinnige tekens.

⁴⁹ Meer voorbeelden op http://www.googleguide.com/advanced_operators_reference.html.

5.3.2 Hoe werken zoekmachines?

Al snel was er op het groeiende web nood aan een zoekmechanisme om in die overvloed van pagina's en informatie een weg te vinden. Zoekmachines zoals Google spelen in op die vraag. Als we het www zouden vergelijken met een boek of encyclopedie dan is Google niet de inhoudstafel, maar eerder de index. Het indexeren van miljarden steeds wijzigende webpagina's gebeurt natuurlijk niet door mensenhanden. Dit zou veel te veel tijd in beslag nemen en zelfs voor een "gratis" dienst als Google onbetaalbaar worden. Het indexeren gebeurt met behulp van een zoekrobot, de **Googlebot**. Uiteraard is dit geen echte "robot", maar een stuk software. De Googlebot vraagt een webpagina op bij een server. De server stuurt de gevraagde pagina terug. Met **reguliere expressies** zoekt de software naar hyperlinks. Wanneer hij zo'n link vindt, weet de software dat hij opnieuw een volgende webpagina heeft gevonden die hij op dezelfde manier kan doorzoeken. De zoekmachine zoekt dus naar hyperlinks en probeert te ontdekken waarover de inhoud van de pagina gaat, bijvoorbeeld door te controleren of bepaalde woorden vaak voorkomen. *Zie ook 5.3.1 voor meer informatie over zoekmachines.* Zoals je merkt, lijkt de werking van de Googlebot erg op die van een browser, alleen is de Googlebot veel sneller dan een menselijke surfer: hij kan duizenden pagina's tegelijk opvragen. Om een webserver niet te overbelasten, spreidt de Googlebot zijn simultane vraag over meerdere servers en websites. Vindt de Googlebot op de opgevraagde pagina's nog links naar andere pagina's, kan hij ook die opvragen.



Opgelet: begrijp dit niet verkeerd! **Het indexeren gebeurt vooraf.** Google schuimt niet het ganse world wide web af op het moment dat jij een zoekopdracht invoert. De zoekmachine heeft reeds een soort inhoudstafel of index van het web gemaakt en vult die voortdurend aan. Wanneer jij een zoekopdracht intikt, kijkt Google even in zijn databank en vist daar uit welke webpagina's het meest overeenkomen met jouw vraag.

Eenvoudig voorgesteld zou een index van een zoekmachine er als volgt kunnen uitzien:

Zoektermen	Omschrijving van pagina	URL
hond, Tervuren	Hondenkapsalon Nancy in Tervuren, gespecialiseerd in verzorging van rijkeluihonden.	Www.hondenkapsalonnancy.be
hond, Tervuren	Hondenkennel WoefWoef in Duisburg (Tervuren): voor iedereen die een rasechte straathond in huis wil halen.	Www.woefwoef.net
...

Uiteraard is dit een erg simplistische voorstelling van hoe zo'n index is opgebouwd. In werkelijkheid is de precieze opbouw van de index en de zoekalgoritmes van Google strikt geheim, maar zelfs dan moet je beseffen dat hun databank wat meer bevat dan één enkele langgerekte spreadsheettabel. Maar de tabel hierboven kan

wel een klein beetje verduidelijken (hoop ik toch) hoe een indexer voor zorgt dat je ultrasnel zoekresultaten gepresenteerd krijgt.

1. De gebruiker voert de zoekterm “hond” in.
2. Hij krijgt de adressen www.hondenkapsalonnancy.be en www.woefwoef.net.

5.3.3 De noodzaak van “semantiek”

Om tekst en andere media beter toegankelijk te maken en vooral de inhoud ervan te ontsluiten voor software en zoekmachine, zijn er tal van technieken bedacht die het mogelijk maken voor de maker (auteur) of uitgever om het document beter doorzoekbaar te maken.

Bedoeling is om duidelijk aan te geven waarover elk onderdeel van een tekst gaat. Het meest duidelijke voorbeeld vinden we terug in webpagina's.

In het voorbeeld op de afbeelding heeft de webontwikkelaar in de broncode van de webpagina aangeduid waarover de inhoud van de webpagina gaat. Zoekmachines lezen die informatie uit en kunnen op basis van die gegevens jouw site een betere of minder goede plaats geven in de zoekresultaten.

In een webpagina noemt men zulke voor de gebruiker verborgen informatie “metadata”.

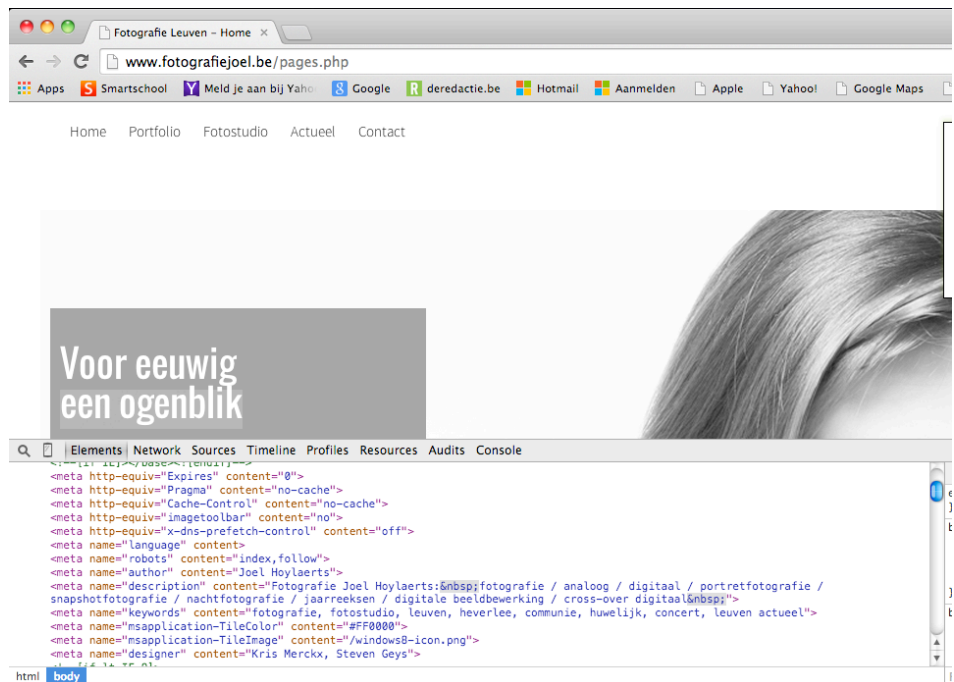
Wat zou de functie zijn van de onderstaande metadata? Bekijk ook de afbeelding met de zoekresultaten in Google.

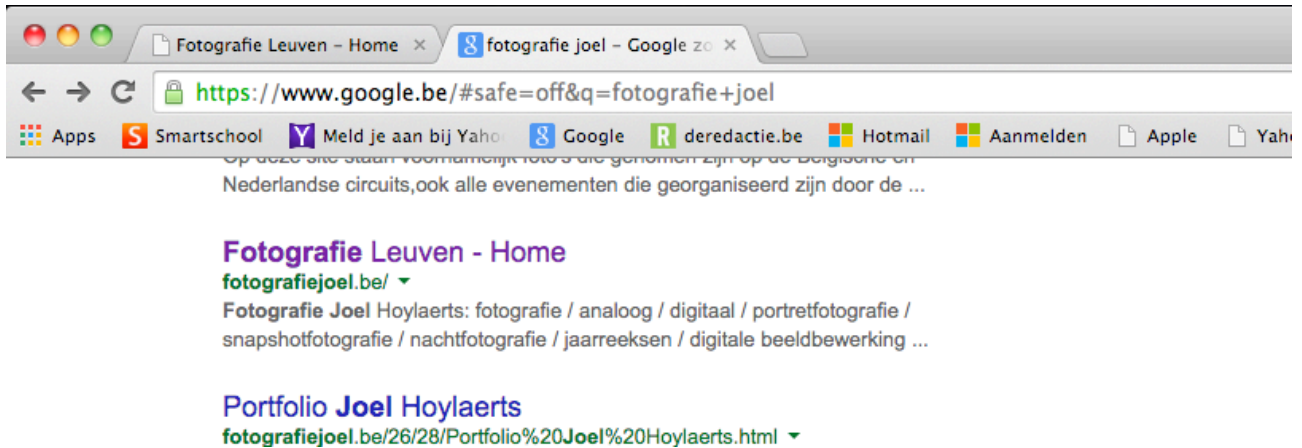
```
<meta name="author" content="Joel Hoylaerts"/>
```

```
<meta name="description" content="Fotografie Joel Hoylaerts: fotografie / analog / digitaal / portretfotografie / snapshotfotografie / nachtfotografie / jaarreeksen / digitale beeldbewerking / cross-over digitaal / over digitaal"/>
```

```
<meta name="keywords" content="fotografie, fotostudio, leuven, heverlee, communie, huwelijk, concert, leuven actueel"/>
```

```
<meta name='designer' content='Kris Merckx, Steven Geys'/>
```



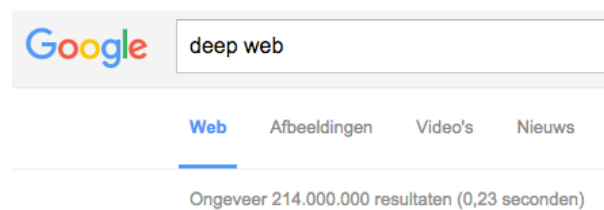


5.3.4 Het deep web

Zoekmachines vinden slechts een fractie van alle op het internet aanwezige informatie. Volgens ruwe schattingen komt in zoekmachines slechts 1/500ste deel van alle beschikbare data naar boven drijven.

De redenen:

1. Heel wat informatie zit verborgen achter gebruikersnamen en wachtwoorden. Digitale leeromgevingen, banken, overheden... publiceren hun data wel op het internet, maar die informatie zit afgesloten en vereist wachtwoordtoegang.
2. Heel veel sites (kranten, winkels, blogs..) stoppen heel veel data in een achterliggende databank. Net zoals bij Google het geval is, bestaan alle webpagina's niet echt als fysieke bestanden, maar zitten alle artikels verborgen in een databank. Wanneer de bezoeker een zoekopdracht invoert, worden de webpagina's door een stuk software op de server, aangemaakt. Vermits zoekmachines hoofdzakelijk op zoek gaan naar "hyperlinks", kunnen ze deze informatie niet (of in elk geval moeilijk) vinden.
3. Heel wat informatie wordt ook bewust verborgen omdat ze illegaal is. **Opgelet:** illegale data komen soms wel bovendrijven. Denk bijvoorbeeld aan de roemruchte Pirate Bay. Die kwam wel aan het oppervlak, ook al stond het netwerk vol met illegaal gekopieerde data.





Bron ⁵⁰

Het web telt op die manier drie niveaus: **het surface web, het deep web en het dark web.**

Uiteraard bestaan er technieken om het deep web en het dark web af te speuren. Bedoeling is dat je geen sporen achterlaat, want al je zoekopdrachten (sorry jongens als je “anoniem” surft om speelse sites te ontdekken) blijven zichtbaar. Misschien is elk spoor weg op jouw computer, maar je telecomoperator houdt lijsten van het internetverkeer bij. Anoniem surfen kan wel, maar dan moet je goed weten waar je mee bezig bent en de juiste technieken hanteren. Eén daarvan is TOR (<https://www.torproject.org/>).

*"Tor (kort voor The onion router) is een open netwerk voor anonieme communicatie gebaseerd op een techniek genaamd Onion routing. Onion routing is een technologie ontwikkeld in 1995 door het US Naval Research Laboratory. Het Tor-netwerk is bedoeld om te voorkomen dat anderen door analyse van het berichtenverkeer kunnen achterhalen wat de herkomst en bestemming van berichten is."*⁵¹

50 "Massive deep web links", (securityzap.com), Geraadpleegd op 5 september 2015.

51 TOR netwerk, ([https://nl.wikipedia.org/wiki/Tor_\(netwerk\)](https://nl.wikipedia.org/wiki/Tor_(netwerk))), Geraadpleegd op 5 september 2015.



Bron ⁵²

52 Deep web vs. Dark web, (www.deepwebtech.com), Geraadpleegd op 5 september 2015.