

# Web Caching

Een TA-studie

**Door :**

Mark Bergsma (s483608)

# Inhoudsopgave

Inleiding .....	3
Technologische kaart .....	4
Algemene werking .....	4
Technische voordelen van het gebruik van caching .....	4
Browser .....	5
Web server .....	5
HTTP .....	5
Soorten caches .....	6
Browser cache .....	6
Proxy caches .....	6
Site caches .....	8
RSS .....	8
Toekomstige ontwikkelingen .....	9
Actorenkaart .....	10
Technologiegebruikers .....	10
Technologieontwikkelaars .....	11
Technologieregulators .....	11
Effectenkaart .....	12
Beschikbaarheid informatie .....	12
Auteursrechtelijk beschermde informatie .....	13
Aanpassingen informatieoverdracht t.b.v. web caching .....	14
Inbreuk privacy en vrijheid gebruikers .....	14
Maatregelen .....	16
Technische aanpassingen .....	16
Uitbreiding HTTP standaard voor caching .....	16
Standaardisatie robots.txt .....	16
Cache verwijderingsprotocol .....	17
Juridische maatregelen .....	17
Wetgeving m.b.t. web caching .....	17
Wetgeving m.b.t. privacy .....	19
Conclusie .....	20
Referenties .....	21
Appendix A: technologische kaart .....	22
Appendix B: Actorenkaart .....	23
Appendix C: Effectenkaart .....	24
Appendix D: Maatregelenkaart .....	25

## ***Inleiding***

Dit is het verslag van een Technology Assessment studie naar *Web Caching*. Hierbij worden de technische aspecten en de daarbij behorende maatschappelijke effecten en invloeden onderzocht.

De term *cache* is afkomstig uit het Frans, en betekent *opslaan*. In de computerwereld betekent caching in het algemeen het opslaan van een kopie van brongegevens, meestal met het doel de performance van het opvragen van de gegevens te verhogen, of de bereikbaarheid ervan te vergroten. *Web caching* is een algemene benaming voor caching van objecten die op het World Wide Web beschikbaar zijn, in veel gevallen voor een groot publiek (meestal: het gehele Internet).

In veel gevallen is caching (grotendeels) transparant voor de gebruiker/opvrager van de informatie, en vaak ook voor de aanbieder ervan. Omdat deze caching plaats kan vinden zonder dat de aanbieder van de informatie hier controle over heeft, ontstaan er echter maatschappelijke problemen. Deze problemen zijn in te delen in een aantal categorieën:

- Auteursrecht
- Privacy
- Censuur
- Problemen van technische aard

Al deze probleemcategorieën zullen in deze studie worden behandeld. Gestart zal worden met een technologische kaart, waarbij de puur technische ontwikkelingen en toekomstverwachtingen worden doorzocht.

## **Technologische kaart**

Bij web caching wordt gebruik gemaakt van een groot aantal verschillende technologieën, waarvan enkel de meest voorkomende en meest verschillende (met name in de invloed ervan op de maatschappij) hier zullen worden behandeld.

Eén aspect hebben echter al deze verschillende technologieën gemeen: de manier waarop ze gebruikt worden. Ergens op het web is informatie beschikbaar, meestal in HTML-vorm, maar dit is niet noodzakelijk. Deze informatie is toegankelijk voor het gehele Internet, of slechts voor een besloten gebruikersgroep. Middels een *web server* wordt deze informatie ter beschikking gesteld aan het publiek.

Daarnaast zijn er de gebruikers, die deze informatie willen bezichtigen. In de computerwereld betekent “bezichtigen” dat de betreffende informatie opgehaald moet worden (doorgaans door deze te *downloaden*), oftewel dat er een *electronische kopie* moet worden gemaakt van de informatie. Op het World Wide Web wordt dit doorgaans gedaan met behulp van een *web browser*.

## **Algemene werking**

In het eenvoudige geval, zonder caching, maakt de web browser direct contact met de web server, en dient een verzoek in naar de gewenste informatie. Indien dit verzoek wordt gehonoreerd, wordt de informatie direct verzonden in het antwoord.

Indien gebruik wordt gemaakt van caching, vindt dit altijd plaats tussen de twee hiervoor genoemde entiteiten, de web server en de web browser (in de vorm van gebruiker). Het verschil is dat niet in alle gevallen het gehele traject tussen gebruiker en aanbieder doorlopen wordt, en de informatie bij de bron wordt opgehaald, maar eerst ergens in het traject gekeken wordt of de betreffende informatie (in de vorm van een kopie) reeds beschikbaar is. Indien dit het geval is en deze informatie geschikt wordt bevonden als zijnde gelijk aan het origineel, dan zal gebruik worden gemaakt van de kopie in plaats van een het oorspronkelijk verzoek aan en antwoord van de bron. Alternatief – als de informatie niet in de cache aanwezig is of om andere redenen niet kan worden gebruikt – wordt het verzoek doorgeleid naar de bron en afgehandeld alsof er geen cache aanwezig is.

## **Technische voordelen van het gebruik van caching**

De grootste voordelen van en meestgebruikte redenen voor het gebruik van caching zijn het verbeteren van de performance van het verkrijgen van de informatie, het verbeteren van de beschikbaarheid ervan, en het

ontlasten van de bron. Indien een cache dicht(er) bij een gebruiker wordt geplaatst, is het mogelijk dat de gebruiker hier sneller toegang toe heeft, doordat de afstand tussen gebruiker en cache kleiner is en zodoende de *latency* (reactiesnelheid) lager zal zijn. Omdat bij caching niet het gehele oorspronkelijke traject van gebruiker naar bron hoeft te worden afgelegd, kan *bandbreedtebesparing* plaatsvinden: in plaats van een bepaald percentage van de beschikbare capaciteit te gebruiken over het gehele traject, hoeft slechts capaciteit te worden gebruikt op een kleiner deel daarvan, zodat de totale gebruikte capaciteit kleiner is. Omdat dataverkeer (door eindige capaciteit) geld kost, kan hiermee een kostenbesparing worden bewerkstelligd. Bovendien is de netto haalbare snelheid in het algemeen hoger indien de communicatiepartner dichter in de buurt zit, zodat ook het gebruiksgemak verhoogd wordt: de informatie is sneller beschikbaar.

## **Browser**

Zoals reeds genoemd wordt informatie op het WWW doorgaans bezichtigd met behulp van een web browser, hoewel dit ook met veel andere, meer specialistische programma's mogelijk is, die het HTTP-protocol ondersteunen. De meest gebruikte web browser is *Microsoft Internet Explorer*, en een veel gebruikt *open source* alternatief (voor meerdere computerplatformen) is de *Mozilla suite*, en ook daaruit voortgekomen projecten als *Firefox*.

## **Web server**

Aan de aanbieder-kant draait een web server, die de verzoeken tot informatie van *web clients* (zoals web browsers) beantwoorden. De veruit populairste web server is *Apache*, op enige afstand gevolgd door *Microsoft's IIS* (Bron: Netcraft [NC]).

## **HTTP**

Het WWW is gebouwd op basis van het *HTTP-protocol*. Dit protocol is een (oorspronkelijk) relatief eenvoudig, op leesbare tekst gebaseerd request/response protocol. Zodra een HTTP-client een TCP-verbinding maakt met de server, kan deze één of meerdere verzoeken versturen, die de server direct zal proberen te beantwoorden. Het meest gebruikte verzoek-type is de *GET-request*, waarbij de client een zogenaamde *URL* (Uniform Resource Locator, uniek plaatsbepalende naam) van de gewenste informatie opgeeft, met de bedoeling deze informatie in zijn geheel te ontvangen.

De oorspronkelijke HTTP standaard, HTTP 0.9, is ontwikkeld in 1992 en was een vrij eenvoudig protocol dat enkel kernfuncties ondersteunde. In de huidige 1.1 versie, zoals beschreven in [RFC2616] en afgerond in

1999, is het protocol echter behoorlijk uitgebreid met geavanceerde functies, onder andere om caching te ondersteunen. Sindsdien is het protocol stabiel gebleven, en momenteel zijn er weinig tot geen wijzigingen in te verwachten.

## **Soorten caches**

Er is een groot aantal verschillende soorten caches die gebruikt worden voor het WWW. Slechts de meest gebruikte en meest belangrijke caches zullen hier worden besproken, samen met enkele voorbeelden ervan, indien van toepassing.

### ***Browser cache***

Een cache waar vrijwel elke Internetgebruiker mee te maken heeft, is de *browser cache*. Een web browser heeft in het algemeen een eigen, persoonlijke cache, die gebruikt wordt om opgehaalde web-objecten voor beperkte tijd op te slaan. Dit opslaan gebeurt zowel in het geheugen als op schijf, waarbij een maximale capaciteit in acht genomen wordt.

De browser cache is nuttig, omdat iemand die op het web surft regelmatig dezelfde informatie bekijkt, door bijvoorbeeld regelmatig dezelfde site te bezoeken, of de *back-knop* van de browser te gebruiken, om daarmee de vorige pagina opnieuw op te vragen. Bovendien bestaan veel websites uit elementen die in elke pagina herhaald worden. Door deze elementen op te slaan in de browser cache, hoeven ze niet voor elke pagina opnieuw opgehaald te worden bij de bron.

Doordat de browser cache zich zeer dicht bij de gebruiker bevindt - namelijk op dezelfde computer - is de snelheid van het opvragen van objecten die zich reeds in de cache bevinden zeer hoog. De gecachte informatie is bovendien volledig persoonsgebonden, aangezien de cache niet met anderen gedeeld wordt. Hierdoor wordt een hoge mate van efficiëntie bereikt, en treden er geen problemen op met privacy.

### ***Proxy caches***

Een tweede vorm van web caching die veel gebruikt wordt, is de *caching proxy server*. Een *proxy* is als het ware een tussenpersoon, iemand die handelt in naam van een ander. Een *proxy server*, is een server die op verzoek van een client een bepaalde handeling verricht, en het resultaat ervan teruggeeft aan de client. Zodoende is een *web proxy* een server die op verzoek van een web browser (of andere HTTP client) een object ophaalt van een web server, en dit vervolgens doorgeeft. De objecten die worden opgehaald door de proxy server kunnen uiteraard worden opgeslagen in een cache, en eventueel later worden herbruikt indien een

nieuw verzoek voor hetzelfde object gedaan wordt. Een proxy server die dit doet, is een *caching proxy*.

Caching proxy servers (voor het gemak vanaf nu “proxy server”, “proxy” of “proxy cache” genoemd) bevinden zich vaak relatief dicht bij de gebruiker. Meestal betekent dit: in hetzelfde netwerk (bijvoorbeeld binnen het netwerk van het bedrijf, de school of instelling), of in het netwerk van de Internet Provider. Hierdoor is de beschikbare capaciteit van het netwerk tussen de gebruiker en de proxy relatief groot, maar kan deze toch door meerdere (en soms: zeer vele) gebruikers worden gebruikt. Omdat populaire sites vaak door meerdere gebruikers worden bezocht, bouwen de proxies een cache op van populaire en dus vaak opgevraagde objecten, die daarmee snel beschikbaar zijn en veel bandbreedte kunnen besparen. In plaats van vele herhaalde verzoeken tot informatie/web-objecten naar de bron, hoeft de data slechts éénmaal over de Internetverbinding, en is vervolgens voor het gehele netwerk beschikbaar.

Bij websites die slechts door één persoon worden bezocht zijn caching proxies echter minder efficiënt, aangezien dit al door de browser cache opgevangen wordt.

Proxy servers kunnen zowel optioneel als vereist in het netwerk aanwezig zijn. In het eerste geval kan een gebruiker er zelf voor kiezen om de proxy al dan niet te gebruiken. Dit is vaak het geval bij Internet Providers, die de proxy aanbieden om de snelheid en beschikbaarheid van het surfen voor hun klanten te verhogen, maar de traditionele manier niet willen blokkeren. In netwerken van bedrijven en instellingen worden proxy servers vaak wel verplicht gesteld, door webtoegang zonder proxy server te blokkeren of zelfs HTTP-connecties naar web servers automatisch om te leiden naar de proxy server. Dit heeft voor het bedrijf of de instelling het voordeel dat hiermee een efficiënter gebruik van de (vaak kostbare) Internetverbinding bewerkstelligd wordt, maar bovenal dat men enige controle kan uitoefenen op het surfgedrag van de werknemers en/of andere gebruikers in het netwerk. Zo biedt een proxy server een eenvoudige, centrale mogelijkheid om het surfgedrag enerzijds vast te leggen (voor latere controle), anderzijds te sturen (door bijvoorbeeld bepaalde sites te blokkeren omwille van hun inhoud).

Een andere vorm waarin caching proxy servers soms worden ingezet, is als *reverse caches*, of ook wel *surrogates* genoemd. Hierbij worden één of meerdere caches vlak vóór web servers van een bepaalde website geplaatst, en ervoor gezorgd dat (enkel) alle verzoeken voor objecten van die website via deze proxy servers geleid worden. Dit heeft tot doel de toegang tot deze betreffende website te versnellen, doordat de proxies de statische inhoud vanuit hun caches kunnen leveren, en zo de web servers, die vaak hun handen vol hebben aan het genereren van dynamische content, kunnen ontlasten. Het is van belang op te merken dat deze proxies volledig gericht en geoptimaliseerd zijn voor de betreffende website, en onder controle staan van de aanbieder van de informatie, in

tegenstelling tot de meeste andere web caches. Daarom zal deze vorm van caching niet verder worden behandeld.

Een veel gebruikte caching proxy server is van oudsher het open source programma *Squid*. Inmiddels zijn er echter ook veel commerciële alternatieven, waarbij het accent niet altijd ligt bij de caching, maar op de beveiligingsfuncties (proxying).

## **Site caches**

De web caches die tot nu toe behandeld zijn, waren allen min of meer transparant aanwezig in het proces. Ze zijn technisch aanwezig in het traject van het ophalen van de informatie, maar veranderen niets aan de perceptie van de inhoud van de informatie zelf door de gebruiker. Bij de soorten caches die ik onder het kopje 'site caches' zal scharen is dat anders.

Er zijn websites die complete kopieën opslaan van andere websites, met het doel deze voor de gebruiker beschikbaar te maken. De redenen hiervoor lopen uiteen. Vaak wordt dit primair gedaan om de beschikbaarheid van de gecachte site te verhogen, in het geval dat de eigenlijke site zelf down is, of eenvoudigweg niet meer bestaat. Een prominent voorbeeld hiervan is de *Google cache* [GOOG], de cache van complete websites die geïndexeerd worden door wat velen beschouwen als de beste zoekmachine op het Internet. De Google cache slaat kopieën op van grote delen van alle webpagina's die in de index zitten, en biedt deze aan als aparte link bij de zoekresultaten, zodat een gebruiker ook de inhoud kan bekijken als de site zelf niet bereikbaar is. Een ander voorbeeld zijn de *short link services*: websites die verwijzingen maken voor inhoud met een zeer lange URL, door gebruikers hiernaar om te leiden vanaf een veel kortere URL. Vaak wordt hierbij de inhoud van de oorspronkelijke pagina's ook gecached.

In andere gevallen wil men de gebruiker de mogelijkheid bieden om de inhoud van meerdere sites te kunnen bezichtigen door slechts één site te bezoeken. Dit gebeurt veel op nieuwssites, zoals *Google News* [GNWS], wat één grote verzameling is van stukjes tekst van andere websites, die volledig geautomatiseerd worden verzameld. Indien bezoekers de volledige teksten willen lezen, worden zij echter wel doorgeleid naar de originele bron.

Dit zijn ook standaardpraktijken bij zogenaamde *verzamellogs*, weblogs (blogs) die stukjes tekst verzamelen van andere weblogs of nieuwssites.

## **RSS**

Dit laatste wordt enigszins vergemakkelijkt door een technologie die *RSS* heet. RSS is een *XML*-standaard die nieuwsstukjes aanbiedt in een voor

andere programma's leesbare standaard, zodat deze vervolgens in willekeurige verschijningsvorm opnieuw kunnen worden gepresenteerd. Zo kunnen web browsers deze informatie in 'n alternatieve, door de gebruiker aangepaste manier weergeven, en bestaan er ook speciale *RSS-readers*, die meerdere *RSS-feeds* samen kunnen weergeven. Maar het opnieuw publiceren van deze informatie op 'n andere website is hiermee uiteraard ook eenvoudig mogelijk.

## **Toekomstige ontwikkelingen**

Wat betreft toekomstige ontwikkelingen is er op het gebied van de overdracht van web-informatie, met behulp van het HTTP-protocol, voorlopig weinig te verwachten. De HTTP-standaard is al enige jaren stabiel, onveranderd, en in algemeen gebruik. Bij de laatste revisie, protocol 1.1, zijn allerlei geavanceerde uitbreidingen toegevoegd, waaronder vele die uitgebreide controle over caching mogelijk maken. Hierdoor zijn verdere wijzigingen voorlopig grotendeels onnodig.

Wel nog in beweging is de manier waarop de gegevens zelf zijn gestructureerd, de (X)HTML-standaarden. Hierbij is een duidelijke trend waarneembaar richting XML, wat zodanig gestructureerd is dat de gegevens geïnterpreteerd kunnen worden door programmatuur, in plaats van slechts door mensen. Hierdoor zou het kunnen dat in de toekomst caching meer gaat plaatsvinden op inhoudsniveau, in plaats van het huidige verschijningsniveau-gerichte caching.

## **Actorenskaart**

Deze sectie beschrijft de relevante *actoren*, oftewel *betrokken partijen*, bij web caching.

### **Technologiegebruikers**

Allereerst zijn er de *cachegebruikers*. Zij maken, gewild of ongewild, gebruik van caching. In een privéomgeving geschiedt dit vaak naar eigen wens, om het surfen sneller te maken, en bandbreedte te besparen. Met de toenemende opkomst van breedbandverbindingen wordt dit echter steeds minder gedaan.

In bedrijfsomgevingen is dit wel gemeengoed, en eerder in toenemende dan afnemende mate. Hier speelt het bedrijf, of het management ervan, de rol van *cache provider*, en biedt de caching al dan niet verplicht aan aan medewerkers of andere betrokkenen (de gebruikers) om bandbreedte te besparen, de internetverbinding te ontlasten, maar vooral ook om enige controle te hebben over het surfgedrag. Deze laatste trend, het *beveiligings- en controleaspect* van webcaches, is in feite een *technologie pull-mechanisme* vanuit de cache providers. Cache ontwikkelaars spelen hierop in door dit *onverwacht gebruikersinitiatief* verder te ondersteunen en ontwikkelen.

In het geval van caching websites (zoals de Google cache) is het duidelijk dat de gebruiker er zelf expliciet voor kiest deze caches te gebruiken.

Indirect gebruikers van deze vorm van caching zijn de *informatieaanbieders*. Zij kiezen er meestal niet voor om hun gebruikers gebruik te laten maken van caching, maar hebben hier geen controle over. Omdat caching in sommige gevallen technische problemen geeft, dienen zij hier echter rekening mee te houden, en hun *contentside ontwikkelaars* opdracht geven de software zodanig te ontwikkelen dat dit goed samenwerkt met de caching.

Sommige grotere, zwaar belaste websites maken wel direct gebruik van caching, en plaatsen zelf caches direct voor hun web servers die de content aanleveren. In dit geval zijn de informatieaanbieders dus wel directe gebruikers van web caching, en hebben zij hier ook volledige controle over.

Bij site caches is de situatie echter volledig anders. Omdat de aangeboden informatie hierbij nadrukkelijk in de context van een andere site wordt geplaatst, en dit niet op een voor de gebruiker transparante manier gebeurt, zijn de implicaties voor de informatieaanbieders volstrekt verschillend.

Zijdelings betrokken zijn de *systeembeheerders* van de caches. Zij doen in opdracht van de cache providers de caches beheren en configureren naar hun wensen. Hierbij is echter wel het privacyaspect van belang, omdat zij door inzage in logfiles een goed beeld kunnen krijgen van wat gebruikers doen.

## Technologieontwikkelaars

Web caches zijn in het algemeen geïmplementeerd in software, en wordt geschreven door software *cache ontwikkelaars*. Deze werken in het algemeen voor bedrijven die cacheoplossingen verkopen, zoals NetApp, BlueCoat, Inktomi, etc. Ontwikkeling van open source producten (zoals Squid) met onbetaalde ontwikkelaars komt echter ook voor. Open source ontwikkelaars hebben duidelijk andere belangen dan hun betaalde collega's.

Aan de informatieverstreckende kant moeten de *contentside ontwikkelaars* rekening houden met de effecten van caching. Sommige content mag bijvoorbeeld helemaal niet gecached worden, omdat elk verzoek andere informatie oplevert. Andere content mag niet worden gecached door een publieke cache, die door meerdere gebruikers gebruikt wordt (bijvoorbeeld omdat de informatie privacy gevoelige en gebruikersspecifieke informatie bevat), maar wel in een browser cache. Dit alles dient door de contentside ontwikkelaars gestuurd te worden volgens algemeen geldende standaarden.

## Technologieregulators

Vooraf betrokken bij de technische ontwikkeling van protocollen en andere standaarden m.b.t. web caching zijn *standaardisatieorganisaties*, zoals het IETF en het W3C. Zij spelen in op de behoeften van de Internetgemeenschap (waar de hiervoor genoemde gebruikers, content providers en cache providers ook onder vallen), en creëren nieuwe standaarden, of passen oude aan. Zo is het HTTP 1.1 protocol [RFC2616] een op nieuwe behoeften aangepaste verfijning van het oudere HTTP 1.0 protocol, waarin ook vele veranderingen zijn opgenomen om web caching te ondersteunen. De resultaten van publicaties van *onderzoeksinstituten*, zoals (academici van) universiteiten worden hierin verwerkt.

Tenslotte heeft de overheid een rol in de vorm van *wetgevende macht*. Hierbij gaat het vooral om overheden op landelijk en/of hoger niveau (zoals de Europese Unie). Bestaande wetgeving, vooral betrekking hebbende op *auteursrechten* en *privacy*, zijn van toepassing op relatief nieuwe technologieën als web caching en het Internet in het algemeen, en kan ontoereikend blijken. Het is dan de taak van zowel de overheid (in de vorm van nieuwe wetgeving) als de rechterlijke macht (in de vorm van jurisprudentie, precedentschepping) om hier duidelijkheid en verfijning te brengen.

## **Effectenkaart**

Deze sectie beschrijft de directe gevolgen van de toepassing van de hierboven besproken technologie en door de in de vorige sectie genoemde actoren.

### **Beschikbaarheid informatie**

Het belangrijkste directe en ook gewenste gevolg, is uiteraard dat de beschikbaarheid van de informatie verbetert. Informatie is sneller beschikbaar, en in geval van problemen op het netwerk tussen de gebruiker en de aanbieder, ook vaker, aangezien de informatie al dicht bij de gebruiker aanwezig is. Het gevolg hiervan is dat gebruikers een vergroot surfgemak hebben, en ook de informatieaanbieders minder klachten krijgen over storingen.

Een ongewenst bijkomend effect is echter, dat (gewilde) informatie die om een of andere reden verboden is, snel over het Internet wordt verspreid, waardoor Internetters op vele verschillende plaatsen toegang hebben tot deze informatie terwijl dit uitdrukkelijk niet de bedoeling is. In veel gevallen is het wel mogelijk om de oorspronkelijke bron van de informatie “van het net af te halen”, maar is het kwaad al geschied, en heeft het zich al over het Internet verspreid. Vervolgens is het vrijwel onmogelijk de informatie alsnog onbeschikbaar te maken.

Een goed voorbeeld hiervan is de verspreiding van de broncode van het kraken van het CSS-algoritme, DeCSS. Het CSS-algoritme zorgt ervoor dat de inhoud van een film DVD versleuteld is, zodat deze niet eenvoudig gekopieerd kan worden. Dit algoritme werd gekraakt in 1999, met een kort stukje broncode genaamd DeCSS. Gerechtelijke procedures om de verspreiding hiervan te voorkomen volgden, en ondanks juridische successen hierbij kon niet worden voorkomen dat de DeCSS-code voor eenieder beschikbaar bleef, doordat het op talloze plaatsen verspreid was – zij het handmatig, en in mindere mate door web caching [WP].

Een recenter voorbeeld is de Radikal-zaak, waarbij Deutsche Bahn rechtzaken aanspande tegen onder meer Google en XS4ALL. Hierbij ging het om de verspreiding van een radikaal links tijdschrift, waarin beschreven werd hoe het Duitse spoorwegennet onbruikbaar gemaakt kon worden (bijvoorbeeld tijdens protesten). Ondanks dat de betreffende editie van dit tijdschrift al 5 jaar bestond en (in Nederland) op papier niet verboden was, werd bij ontdekking op een homepage bij XS4ALL onmiddellijk gesommeerd de betreffende pagina's te verwijderen [XS1]. Na een gerechtelijk bevel deed XS4ALL hieraan gehoorzamen, maar door web caching bleef de site toch enige tijd bereikbaar, bijvoorbeeld via de Google cache, die deze pagina geïndexeerd en gearchiveerd had.

## Auteursrechtelijk beschermde informatie

Bij web caching is het vrijwel onvermijdelijk dat auteursrechtelijk beschermde informatie wordt gecached. In veel gevallen is dat ook wenselijk, met name indien door de web caching niets verandert aan de perceptie van de informatie zelf door de gebruiker. Dit geldt voor proxy caches, die de HTML-informatie ongewijzigd doorgeven, maar wel de beschikbaarheid ervan versnellen en verhogen. Dit is een gewenst effect, en hierbij treedt in het algemeen geen probleem op.

Anders is het bij integratie van stukken auteursrechtelijk beschermde informatie in andere sites, zonder toestemming. Sommige websites publiceren, handmatig geplaatst dan wel op automatische wijze, integraal stukken tekst van andere sites (veelal newssites). Dit is uiteraard geen probleem als hiervoor toestemming is verleend, zoals waarschijnlijk het geval is met Google News [GNWS].

Vaak is die toestemming er niet, en wordt –onterecht- aangenomen dat door het aanbieden van bijvoorbeeld een RSS-feed dit impliciet toegestaan is. Bij ontdekking hiervan volgt doorgaans een expliciet verzoek tot verwijdering van de content bij respectievelijk de webmaster, en of Internet provider van de betreffende site. Wordt hieraan geen gevolg gegeven, kunnen juridische procedures volgen, of technische maatregelen die de caching proberen te verhinderen. In gevallen waarbij de originele eigenaar van de informatie geen grip kan krijgen op de verspreiding ervan, kan het echter ook voorkomen dat, bij dermate hoog inkomsten verlies, minder nieuwe content wordt geproduceerd, of de beschikbaarheid dermate wordt ingeperkt. Dit heeft een *analogie* met de muziek- en filmindustrie, waarin momenteel grote inkomstenderving plaats vindt doordat het traditionele inkomstenmodel omvergeworpen wordt door grootschalige duplicatie op Internet.

De momenteel gangbare procedure voor het laten verwijderen van informatie op Internet, is het klagen over auteursrechtelijk beschermd materiaal bij de betreffende Internetprovider, waarna deze verplicht is het materiaal zo snel mogelijk te verwijderen (in het engels: 'Cease and desist'). Veel Internet providers zijn echter in de praktijk niet goed genoeg op de hoogte van de bestaande auteursrechtwetgeving, en verwijderen materiaal reeds na een enkele, slecht gefundeerde klacht. Zeer duidelijk werd deze gang van zaken door een onderzoek door het bureau Bits of Freedom, dat een tekst van Multatuli online plaatste bij verschillende providers. Deze tekst is al vele jaren niet meer auteursrechtelijk beschermd, aangezien dat recht 70 jaar na de dood van de auteur vervalst. Ondanks dit feit, en de gebruikte procedure waarbij klachten naar providers werden verstuurd met een anoniem Hotmail-adres, verwijderden 7 van de 10 providers deze tekst klakkeloos [BOF2].

## **Aanpassingen informatieoverdacht t.b.v. web caching**

Voor optimaal gebruik van web caching dient de originele informatieoverdacht hierop (technisch) voorbereid te zijn. Bij naïeve informatieopdracht, zonder rekening te houden met web caching, kan het gebeuren dat gebruikers verouderde informatie krijgen (omdat de originele inhoud vaak gewijzigd wordt, en de caches hier geen rekening mee houden), of verkeerde inhoud (doordat deze bijvoorbeeld gegenereerd wordt aan de hand van meer informatie naast de URL). In extremere gevallen kan het zelfs voorkomen dat een gebruiker privacygevoelige informatie bedoeld voor een andere gebruiker ontvangt (bijvoorbeeld bij telebankieren), en uiterlijk dan wordt duidelijk dat dit 'n groot probleem kan vormen en hier aandacht aan dient te worden besteed.

Voor vrijwel alle problemen en nadelen die content providers hebben door web caching bestaan min of meer acceptabele oplossingen en compromissen. Zo zijn er technische aanpassingen gemaakt in de HTTP 1.1 standaard [RFC2616] die de controle over de *cacheability* van content vergroten. Dit vergt echter zorgvuldige aandacht bij de ontwerp van de contentoverdacht, en dus additionele kosten. Veel contentproviders vinden het makkelijker om web caching zoveel mogelijk te verhinderen door 'n paar eenvoudige technische trucs, met nadelige gevolgen voor gebruikers van dien.

## **Inbreuk privacy en vrijheid gebruikers**

Bij web caching wordt het opvragen van gecachte informatie in de meeste gevallen *gelogd*; oftewel, het feit dat een specifiek stuk informatie op 'n bepaald tijdstip wordt opgevraagd door een gebruiker met identiteit X wordt genoteerd. Dit geeft de beheerders van de betreffende cache een extra macht; namelijk wetenschap over wat hun gebruikers uitvoeren op het Internet. Hoewel de opgeslagen informatie relatief beperkt lijkt te zijn geeft dit voldoende mogelijkheden om uitgebreide verbanden te kunnen leggen en 'n goed beeld te krijgen van wat specifieke gebruikers doen. Daarnaast komt dat, specifiek bij caches, niet alleen een omschrijving van de opgevraagde informatie wordt opgeslagen, maar ook de *inhoud van de informatie zelf* (in de cache!), zij het dat beheerders in het algemeen iets meer moeite moeten doen om deze te bezichtigen.

In veel bedrijven is het zelfs beleid dat het management inzicht krijgt in de (Internet-)verrichtingen van hun werknemers. Het management vraagt dan bijvoorbeeld wekelijkse rapporten op van de meest bezochte websites, in zijn geheel of individueel per werknemer. Indien de resultaten hieruit niet stroken met de wensen van het management kunnen er reprimandes volgen, en in extremere gevallen ook ontslagen.

Naast de inbreuk van de privacy van cachegebruikers kan er ook sprake zijn van inperking van hun vrijheid. Een proxy cache staat in het middelpunt van een bedrijfsnetwerk, en ontvangt al het webverkeer van

gebruikers (werknemers), en is daarmee de ideale plaats om gebruikers te controleren. Hoewel niet hun oorspronkelijke functie, hebben veel proxy cache producten inmiddels uitgebreide mogelijkheden om websites met bepaalde inhoud te blokkeren, om daarmee werknemers te proberen te weerhouden andere dingen te doen naast hun directe werkzaamheden.

Hoewel het beargumenteerbaar is dat “in de baas zijn tijd” werknemers zich aan deze voorschriften zouden moeten houden, voelen velen zich in hun vrijheid aangetast, en proberen de beperkingen te omzeilen. Afhankelijk van de kennis en zorgvuldigheid van gebruikers en systeembeheerders lukt dat al dan niet, en kan dit bij ontdekking tot reprimandes leiden.

Deze vorm van *censuur* wordt in veel extremere vorm toegepast in China, waar de regering haar burgers kritische informatie t.o.v. de Chinese regering wil onthouden. Daartoe is het gehele Internetgebruik in China gereguleerd, en verloopt alles via centrale firewalls en web caches die niet toegestane content blokkeren. [ITW1] Afgaande op de gevolgen van andere kritische uitlatingen t.o.v. de Chinese regering, zullen er waarschijnlijk ook relatief hoge straffen staan op het (proberen te) omzeilen van deze beperkingen.

## **Maatregelen**

In deze sectie worden enkele mogelijke maatregelen besproken die het gedrag van de actoren zou kunnen beïnvloeden, en een positieve invloed zouden kunnen hebben om de genoemde problemen op te lossen. Allereerst worden een aantal technische wijzigingen besproken, gevolgd door een aantal mogelijke wijzigingen op juridisch gebied. Voor deze volgorde is gekozen omdat de juridische maatregelen de technische wijzigingen zouden kunnen ondersteunen.

## **Technische aanpassingen**

### ***Uitbreiding HTTP standaard voor caching***

Voor goede (technische) werking van web caching onder alle omstandigheden moest de HTTP-standaard aangepast worden in een verbeterde versie. Eigenlijk blijken deze aanpassingen voldoende voor content providers om controle te houden over wat er precies, onder welke omstandigheden en met welke duur gecached wordt. In feite is dit dus een maatregel die reeds genomen is. De HTTP 1.1 standaard [RFC2616] definieert zogenaamde *headers*, die tijdens de informatieoverdracht specificeren *of* het betreffende informatieobject gecached mag worden (en zo caching volledig kan verhinderen), of dit voor *publieke* of alleen *gebruikersspecifieke* caches geldt (en zo het privacy aspect onder controle kan houden), en met welke *duur* het object gecached mag worden (om te verhinderen dat een gebruiker verouderde informatie ontvangt). Deze headers gelden vooral voor (vanuit het opzicht van de gebruiker) transparante proxy caches, en lossen hiervoor de belangrijkste technische problemen op, mits content providers bereid zijn deze ook toe te passen.

### ***Standaardisatie robots.txt***

Sinds het bestaan van *web crawlers* en andere robots die het web doorzoeken op informatie, en deze eventueel gebruiken en cachen, is er het bestand *robots.txt*, dat instructies bevat voor deze robots. Hierin kan worden aangegeven dat bepaalde pagina's op de betreffende webserver niet door deze robots geïndexeerd en gecached mogen worden. Dit 'protocol' is echter te onstandaard en ad-hoc, dat het onduidelijk is welke toepassingen precies van dit protocol gebruik zouden moeten maken en waar ze zich aan dienen te houden. Momenteel is het protocol te onflexibel en ongestandaardiseerd om alle gevallen afdoende te kunnen afdekken.

Het zou goed zijn als er een nieuw (bijvoorbeeld gebruik makend van XML-technologie), dan wel verbeterd maar op het oude robots.txt gebaseerd protocol komt, dat flexibeler, beter ontworpen is en ook meer

algemeen de problemen met auteursrechten oplost. Dit protocol zou dan (indirect) in nieuwe wetgeving kunnen worden gebruikt als houvast voor zowel content providers als cache providers.

### **Cache verwijderingsprotocol**

Soms bieden sites die caches van informatie aanbieden, ook 'n mogelijkheid voor de eigenaar ervan om deze te verwijderen. Google bijvoorbeeld, ondersteunt het verwijderen van pagina's uit hun index door het plaatsen van een bestandje in de originele webruimte, en het verwijzen daarnaar op een formulier op de Google website.

Een dergelijk protocol kan enigszins helpen om het probleem dat reeds verwijderde informatie via caches bereikbaar blijft op te lossen. Een vereiste is dan wel dat dit gestandaardiseerd wordt, en alle caching sites en proxy caches zich aan deze eenduidige standaard kunnen houden. Het is immers ondoenlijk voor een informatieaanbieder om alle caches op te sporen en handmatig, op verschillende wijzen, te wissen.

Het is echter niet mogelijk om technisch het probleem van de schending van auteursrechten tegen moed en wil aan te passen. In andere gebieden dan web caching waarin dit ook speelt, zoals bijvoorbeeld het kopiëren van CDs en DVDs, wordt grof geschut in stelling gebracht om dit te proberen technisch (m.b.v. encryptietechnologie) te verhinderen, maar hierbij blijkt enerzijds dat dit te grote beperkingen op legt voor legitiem gebruik, en anderzijds dat het serieuze overtreders niet werkelijk tegenhoudt. Een dergelijk protocol kan dan ook slechts worden gebruikt op basis van goede wil, en voor juridische houvast.

### **Juridische maatregelen**

Als techniek niet werkelijk maatschappelijke problemen kan verhinderen, of ze zelfs veroorzaakt, kunnen deze worden aangepakt met juridische middelen.

### **Wetgeving m.b.t. web caching**

Omdat web caching, en algemener het steeds kopiëren van informatie zo duidelijk en onmisbaar aanwezig is in de werking van het Internet, maar de juridische aspecten ervan in veel gevallen onduidelijk zijn, dient de wetgeving hierover te worden aangescherpt.

In artikel 1 van de Auteurswet staat:

*“het uitsluitend recht van de maker van een werk van letterkunde, wetenschap of kunst, of van diens rechtverkrijgenden, om dit openbaar te maken en te verveelvoudigen, behoudens de beperkingen, bij de wet gesteld.”*

Het is duidelijk dat enkel deze stricte definitie niet houdbaar is in dit

technisch tijdperk, waarin het kopiëren noodzakelijk is voor de technische basiswerking van de gehele informatietechnologie. Er moet dus een uitzondering te worden gemaakt waarbij het cachen van informatie dat uitsluitend dient ter (efficiënte) overbrenging van betreffende informatie, waarbij de wettige belangen van de auteur niet worden geschaad. Dit zou minstens moeten gelden voor proxy caches, aangezien deze zoveel mogelijk transparant zijn voor zowel gebruiker als aanbieder.

Moeilijker ligt dit met de zogenaamde 'site caches', waarbij de informatie in enigszins gewijzigde omstandigheden aangeboden wordt. Het verdient aanbeveling dit via technische middelen op trachten te lossen, aangezien hier geen gemene deler te vinden is. In sommige gevallen heeft een caching service nadrukkelijk voordelen voor de content provider (meer gebruikers, betere beschikbaarheid), in andere gevallen juist niet. Dit verschilt per caching service, informatieaanbieder, en groep gebruikers, en dient daarom zoveel mogelijk op basis van samenwerking te worden uitgewerkt. Waar dit niet lukt, kan het recht op individuele basis uitkomst bieden.

Aangezien er bestaande technische voorzieningen (protocollen) zijn waar zowel cache providers als content aanbieders zich aan kunnen houden, en nieuwere, verbeterde protocollen kunnen worden ontwikkeld, kan in de rechtspraak op individuele basis gekeken worden naar de mate van redelijkheid waarin de partijen moeite hebben gedaan samen te werken met behulp van deze technieken. Het is in de wetgeving wel ongewenst om naar specifieke technieken te verwijzen, aangezien deze snel wijzigen en daarmee ook de wetgeving veroudert. Toch kan een meer algemene verwijziging naar "geldende huidige standaarden" worden gemaakt.

Het is interessant om een analogie te leggen naar de situatie in de VS, waar deze problematiek met de enigszins omstreden maar voor web caching wel belangrijke DMCA-wet is aangepakt. In DMCA wordt (transparante proxy-)caching expliciet genoemd, en – voor zover de cache providers zich aan de geldende (technische) afspraken houden, uitgezonderd van inbreuk op het auteursrecht en reproduceerrecht. Van belang is ook dat het materiaal ter beschikking wordt gesteld door 'n andere partij dan de cache provider, en dat alle regels met betrekking tot de verversing en vernieuwing van de inhoud worden gevolgd. Daarmee legt DMCA zeer nauwe banden met de eigenschappen en mogelijkheden van de huidige technologieën, zonder deze expliciet bij naam te noemen.

In de VS is web caching juridisch dus sinds de komst van de DMCA-wet relatief duidelijk geregeld. In Europa dient dit ook op minstens landelijk, maar liever op EU-niveau te worden gereguleerd. De EU heeft daartoe een aantal richtlijnen geformuleerd, die in de landelijke wetgevingen overgenomen dienen te worden. [IVIR1] In Nederland is ook gebeurd, met de nieuwe Auteurswet die op 5 juli 2004 is aangenomen door de Eerste Kamer [AW1]. Cache providers worden hierin gevrijwaard van

onrechtmatig handelen, aangezien het reproduceren van materiaal met uitsluitend een functioneel, technisch doel binnen computernetwerken expliciet wordt toegestaan. Hierin wordt echter, in tegenstelling tot de DMCA, geen verwijzing gemaakt naar de zorgvuldigheid waarmee deze caching plaats vindt met betrekking tot geldende afspraken/protocollen die de cacheability van informatie sturen.

Veel andere EU-landen passen momenteel hun auteurswetten overeenkomstig aan.

In al deze voorbeelden blijft de rechtmatigheid van site caches (zoals de Google cache) onduidelijk, waarbij de DMCA-wet nog de meeste houvast biedt. Tot hier verandering in komt (op dit moment onwaarschijnlijk), dienen site cache providers ermee rekening te houden dat zij onrechtmatig kunnen handelen.

### ***Wetgeving m.b.t. privacy***

Het probleem dat systeembeheerders en hun werkgevers inzicht kunnen krijgen in het (surf)gedrag van hun werknemers, is niet zinnig technisch oplosbaar. Net zoals in andere privacygevoelige gebieden is er echter wetgeving die dit reguleert, en mogelijk aangepast moet worden om effectief te kunnen zijn met de nieuwe technieken, waaronder web caching, maar ook bijvoorbeeld e-mail op de werkplek.

In Nederland is het recht op privacy geregeld in de Grondwet, en in verdragen/lagere wettelijke regelingen. Voor werknemers onder werktijd gelden enkele beperkingen van deze wetten, maar zijn ze niet geheel afwezig. Om meer duidelijkheid te bieden heeft het College Bescherming Persoonsgegevens (CBP) een rapport opgesteld [FD1], dat enkele vuistregels stelt waarmee werkgevers kunnen toetsen in welke gevallen ze gerechtigd zijn om het surf- en e-mailgedrag van hun werknemers te controleren, en in welke mate.

Voornaamste punten uit dit rapport zijn:

- Werkgevers moeten heldere en eenduidige regels opstellen én publiceren zodat werknemers weten in welke mate Internet-gebruik op de werkplek toegestaan is.
- Controle van Internetgebruik mag alleen als de werkgever het vermoeden heeft dat een werknemer het bedrijfsbelang schaadt, en dat vermoeden achteraf ook kan aantonen.
- De integriteit van deze controle moet zoveel mogelijk gegarandeerd zijn.

De behandeling van deze verkeersgegevens is dus wel degelijk reeds aan wetten en beperkingen verbonden, zij het dat dit in de praktijk slechts moeilijk controleerbaar is.

## **Conclusie**

De belangrijkste conclusie die mijn inziens getrokken kan worden is, dat de belangrijkste maatschappelijke problemen rondom web caching eigenlijk al zijn opgelost. Meer dan ik vooraf verwacht had, zijn op zowel juridisch vlak (met nieuwe wetgevingen) als technisch vlak (technische aanpassingen aan protocollen) de belangrijkste hordes reeds genomen.

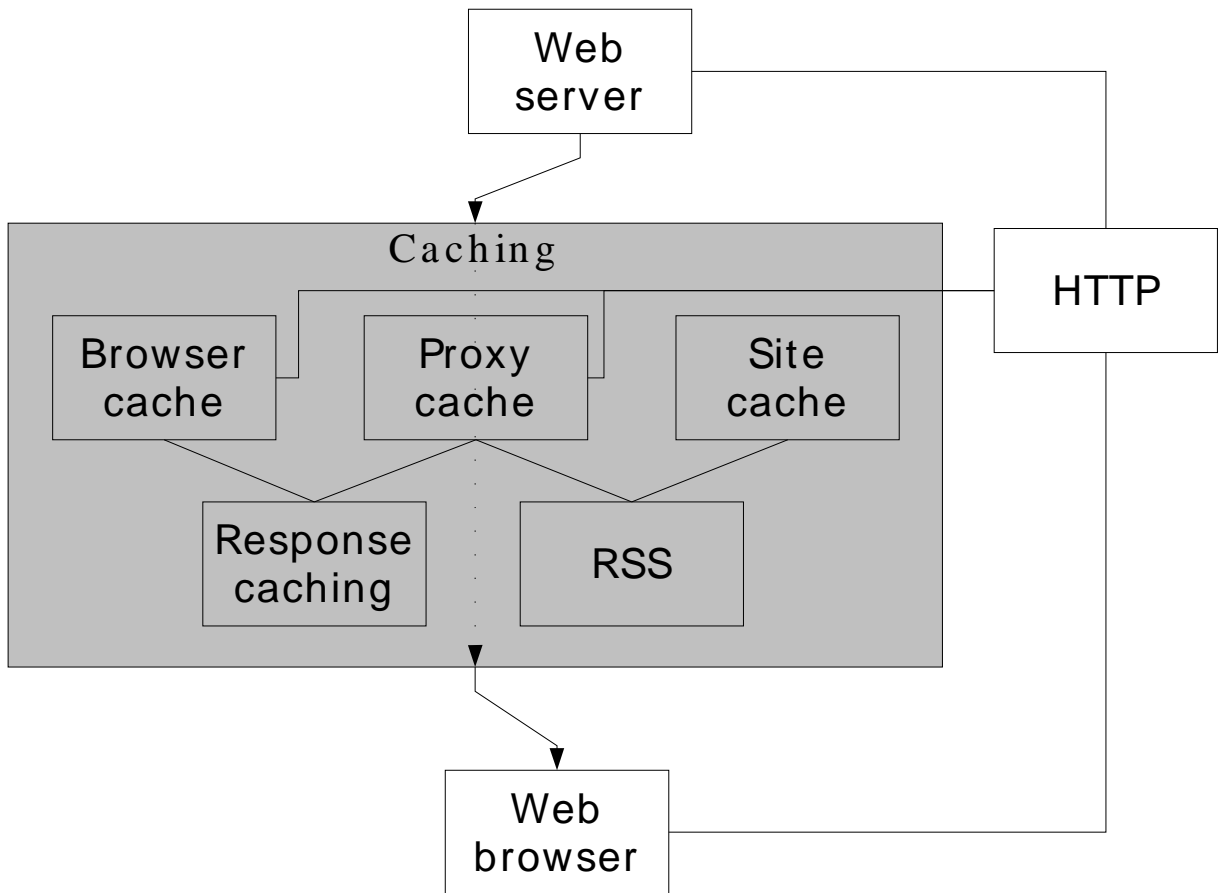
Enkele problemen blijven echter overeind, zoals de rechtmatigheid van site caches (waarbij ik steeds de Google cache noem als voorbeeld). Naar verwachting zullen deze problemen vooral op individuele basis opgelost worden, zeker in gevallen waarbij meerdere partijen voordelen kunnen behalen bij de betreffende services.

Verder vormt web caching eigenlijk relatief weinig problemen voor de maatschappij, en speelt dit (niet onverwacht) in veel mindere mate dan bijvoorbeeld de problematiek rondom RFID, spam, het kopiëren van Cds/DVDs, etc. Desondanks was het toch interessant om een dergelijk beperkt probleemgebied te onderzoeken.

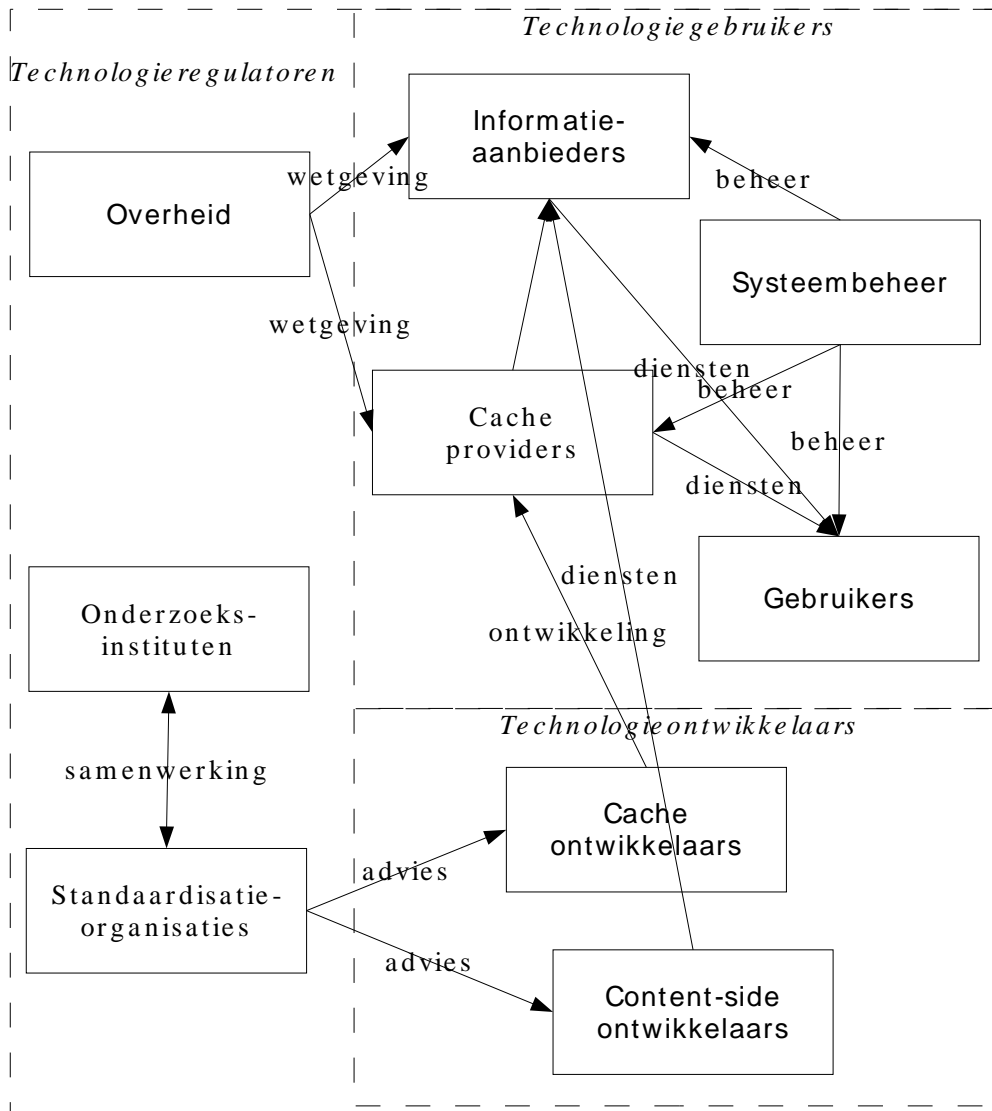
## **Referenties**

- [WP] Wikipedia, the free encyclopedia – <http://www.wikipedia.org>  
Opgezochte begrippen: [[cache]], [[proxy cache]], [[web cache]],  
[[DeCSS]].
- [SMI03] Wim A. Smit, Ellen C.J. van Oost (1999), De wederzijdse  
beïnvloeding van technologie en maatschappij, uitgeverij coutinho.
- [WC01] Duane Wessels (2001), Web Caching, uitgeverij O'Reilly &  
Associates
- [LS01] Lorre Smith, Intellectual Property, Copyright, and Fair Use  
Resources - <http://www.albany.edu/~ls973/copy.html>
- [NC] Netcraft Web Server Survey – <http://www.netcraft.com>
- [RFC2616] IETF Networking Working Group, Hypertext Transfer  
Protocol – HTTP/1.1 - <ftp://ftp.isi.edu/in-notes/rfc2616.txt>
- [GOOG] Google – <http://www.google.com>
- [GNWS] Google News – <http://news.google.com>
- [XS1] Radikal-zaak na 5 jaar voor de rechter -  
[http://www.xs4all.nl/nieuws/bericht.php?id=159&taal=nl&msect=nieuws  
&year=2002](http://www.xs4all.nl/nieuws/bericht.php?id=159&taal=nl&msect=nieuws&year=2002)
- [ITW1] Chinese censors block access to Wikipedia -  
<http://www.itworld.com/Tech/2987/040614wikipedia/>
- [IVIR1] P.B. Hugenholtz (2001), Brussels Broddelwerk. Recht en Krom in  
de Auteursrechtlijn - [http://www.ivir.nl/publicaties/hugenholtz/ami-  
auteursrechtlijn.html](http://www.ivir.nl/publicaties/hugenholtz/ami-auteursrechtlijn.html)
- [BOF1] Bits of Freedom (2003) Nieuwsbrief 7 -  
[http://www.bof.nl/nieuwsbrief/nieuwsbrief\\_2003\\_7.html](http://www.bof.nl/nieuwsbrief/nieuwsbrief_2003_7.html)
- [AW1] Staatsblad van het Koninkrijk der Nederlanden 336 (2004) -  
[http://www.eerstekamer.nl/9324000/1/j9vvgh5ihkk7kof/vgscdlujvmv0/f=y  
.pdf](http://www.eerstekamer.nl/9324000/1/j9vvgh5ihkk7kof/vgscdlujvmv0/f=y.pdf)
- [FD1] Factsheet digitaal - <http://www.doxis.nl/docs/No203.pdf>
- [BOF2] Bits of Freedom (2004), 7 van de 10 providers verwijderen tekst  
Multatuli - <http://www.bof.nl/takedown/>

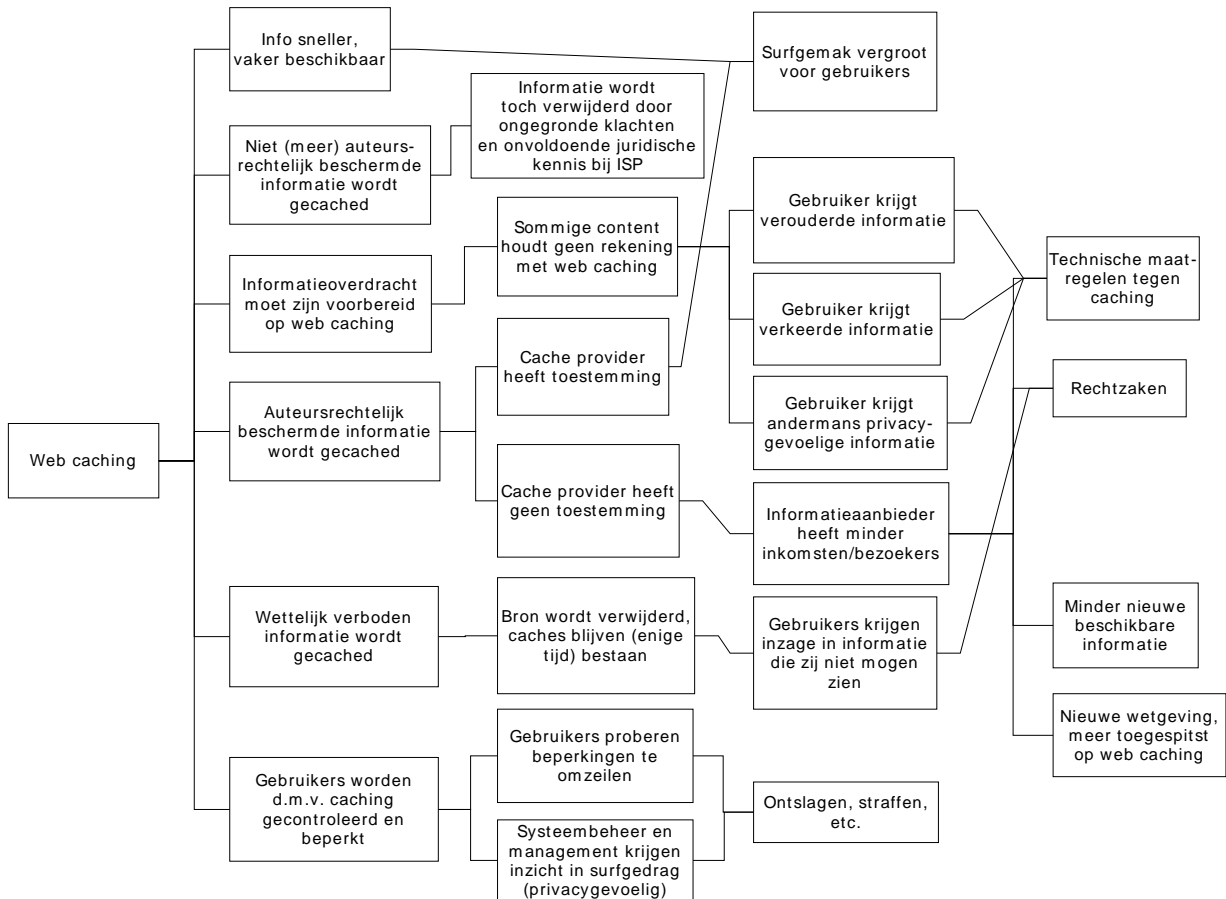
**Appendix A: technologische kaart**



## Appendix B: Actorenkaart



## Appendix C: Effectenkaart



# Appendix D: Maatregelenkaart

