1 Juridiska aspekter på robots.txt - bakgrund

För några dagar sedan blev jag uppringd av domstolsverkets chefsjurist, som hade sett min webbplats lagen.nu. På webbplatsen, som innehåller samtliga författningar som ingår i SFS, är paragrafer i lagtexterna sammanlänkade med rättsfall som i sina domskäl åberopar dessa lagrum; på så sätt kan en användare snabbt hitta rättsfall som illustrerar och förtydligar en viss given paragraf. Samma system används i både ”den blå” och ”den röda” lagboken, dock med den betydande skillnaden att listorna över rättsfall där är resultatet av ett redaktionellt urval, snarare än en automatisk process. För att kunna göra sammanställningen hämtar jag information från domstolsverkets webbplats ”Domstolsväsendets rättsinformation”[1] genom ett datorprogram, en sk webspindel, -crawler eller –robot. En webrobot är ett samlingsnamn för alla sorters program som hämtar data från en webbplats utan att vara direkt styrda av en användare vid varje sidhämtning. En webrobot är en mycket viktig delkomponent i de flesta sorters sökmotorer på Internet, men de har även många andra användningsområden. Den webrobot som jag handhar kan exempelvis inte sägas vara en del av ett sökmotorsystem.

Domstolsverket var kritiskt till att jag direktlänkar till rättsfallsreferaten på deras webbplats; genom mina länkar hade nämligen Google och andra sökmotorers robotar hittat till rättsfallen och indexerat dessa. Detta ansågs oönskat, då vissa referat kan innehålla personuppgifter, och om dessa indexeras så att referaten kan hittas vid en Google-sökning på exempelvis en bostadsadress, så kan dessa personers integritet kränkas.

Det finns alltså två robotar inblandade i den här frågeställningen: Den robot som jag administrerar, som inte behandlar referaten i vidare utsträckning än verifiera att de finns; referatstexten indexeras inte eller behandlas på annat sätt, samt Googles robot, som indexerar och gör sökbart precis allt den kommer över. Domstolsverket hade ursprungligen tänkt att Google och andra sökmotorer inte skulle kunna indexera rättsfallen genom att man valt en relativt komplicerad teknisk lösning för att göra dessa tillgängliga. De tekniska förutsättningar som gäller för att en generell webrobot ska kunna indexera information på en webbplats är att informationen måste gå att nå genom vanliga hyperlänkar; om det, som på domstolsverkets hemsida, krävs att man matar in termer i ett sökformulär, kommer inte en generell webrobot att kunna hitta den. Den etablerade termen för webbinnehåll som på det här sättet är onåbart för generella sökmotorer är ”the deep web”[2]

En specialiserad webrobot har däremot möjlighet att fylla i sökformulär efter givna regler och därigenom kartlägga en specifik webbplats innehåll i mycket större utsträckning. Då jag lade in direktlänkar till rättsfallsdetaljerna och –referaten på lagen.nu upphörde denna information att vara en del av ”the deep web”. Jag tipsade domstolsverket om den tekniska möjlighet som fanns att spärra ut oönskade robotar från deras webbplats, allmänt kallad ”robots.txt”, vilken de samma eftermiddag började använda.

2 Robots.txt – teknisk bakgrund och betydelse

Redan när de första försöken att indexera webben någon gång 1993 inleddes, insåg man att det borde finnas ett sätt för webbplatsägare att, på ett maskinläsbart sätt, klargöra sin ståndpunkt i frågan om de ville tillåta robotar att gå igenom webbplatsen. Lösningen blev en standard kallad ”The Robots Exclusion Protocol” (REP), allmänt känd som robots.txt[3]. Standarden föreskriver att en webplatsägare som önskar kontrollera robotars access till platsen skapar en fil med namnet robots.txt och placerar

[1] http://www.rattsinfosok.dom.se/lagrummet/index.jsp
[2] ”The Deep Web: Surfacing Hidden Value”, http://www.brightplanet.com/technology/deepweb.asp
[3] ”Robots Exclusion”, http://www.robotstxt.org/wc/exclusion.html. Notera att REP är en betydligt inofficiellare standard än de som tas fram inom ramen för IETF, W3C och i RFC-arbetet.

den i ”roten” på webserverns resursrymd; för en webplats med URL:en ”http://foo.com/” så ska filen finnas på URL http://foo.com/robots.txt. Filens innehåll beskriver vilka delar av webplatsen som får hämtas av vilka robotar. Det enklaste exemplet på en restriktiv robots.txt är följande:

```
User-agent: *
Disallow: /
```

Denna talar om att ingen robot (”User-agent”) överhuvudtaget (”*”) får besöka någon del av webplatsen (”Disallow: /”). Termen ”User-agent” används på samma sätt som i HTTP-protokollet[4]; en textsträng som identifierar programvaran som, på en användares begäran, används för att hämta sidor.

Det hör till god sed bland sökmotorer och andra respektabla webrobotadministratörer att respektera den här standarden, och att namnge sin robot på ett sådant sätt att en webplatsägare sedan kan ta ställning till om han önskar besök från roboten i fråga.

Det bör noteras att detta inte är den enda metod som kan användas av en webplatsansvarig för att stänga ute oönskade robotar. Även mer avancerade metoder som inte förutsätter webrobotens samarbetsvillighet, exempelvis User-agent-matchning och analys av sidhämtningsbeteende, står till buds.

3 Länklegaliteten och personuppgiftslagen

Fråga uppstod om huruvida jag inte borde ha kontaktat domstolsverket innan, och även om min hantering kunde tänkas bryta mot personuppgiftslagen. Vad gäller upplysning/ansökan om tillstånd hade jag inte funnit någon anledning till det, då jag trodde att webplatsens ”deep web”-utformning var en bieffekt av den tekniska lösning man valt, inte en medveten strategi i sig.

Vad gäller PUL så var min ursprungliga bedömning att behandling, som enbart befattar sig med den data som finns i varje domsluts detaljvy, inte faller under dess bestämmelser, då jag trodde att personuppgifter enbart kunde förekomma i referaten. Jag har dock blivit uppmärksammad på att även målnummer kan anses vara en personuppgift. Därmed finns en anmälningsplikt enligt PUL 36 §, och då uppgifterna i vissa fall kan hänföras till en brottsmålsdom får dessutom bara myndigheter behandla uppgifterna (21 §). För att behandlingen skall bli tillåten måste jag därmed skaffa ett utgivningsbevis för webbplatsen enligt YGL 1:9 så att denna grundlag blir tillämplig framför PUL. Av flera olika anledningar har detta inte varit aktuellt tidigare, men i ljuset av den uppkomna situationen blir det nog nödvändigt.

4 Olovlig tillgång till upptagning

Som nämnts ovan valde domstolsverket valt att lösa sitt problem genom att införa en restriktiv robots.txt-policy, som stänger ute samtliga webrobotar. Vad är då den juridiska betydelsen av REP? Det är flera olika lagar som kan tänkas spela in.

Personuppgiftslagen har vi behandlat ovan. Upphovsrättslagen kan i vanliga fall spela in; i det här fallet hämtar min robot enbart material som inte lyder under upphovsrätt (Upphovsrättslagen 9§ p 2). Vad som däremot är intressant är brottet dataintrång (BrB 4:9 c).

Dataintrång innebär att man ” olovligen bereder sig tillgång till upptagning för automatisk databehandling” – att det i det aktuella fallet är fråga om ”upptagning för atomatisk databehandling” är givet. Frågan är hur ”olovligen” ska tolkas.

I det här sammanhanget passar det bra att presentera ytterligare en teknisk term: Sceenscraping (även kallad webscraping), dvs. processen att från webbsidor som hämtats med en webrobot, utvinna specifik information. Det ligger i screenscrapingens natur, till skillnad från den mer generella sökmotorindexeringen, att programmen

[4] Se RFC 2616 (tillgänglig från http://www.w3.org/Protocols/Specs.html), sektion 14.43

anpassas efter specifika webbplatser. Den webrobot som används för lagen.nu har screenscraper-egenskaper i det att den utvinner information såsom titel, målnummer, lagrum med mera från domslutsdetaljvyn.

För två år sedan var jag inblandad i ett open source-projekt kallat XMLTV[5]. Jag underhöll ett program, en screenscraper med tillhörande specialiserad robot, som utvann tv-tablåinformation från webbsidorna hos dagenstv.com, och konverterade denna till ett XML-baserat format för vidare användning i HTPC-lösningar[6]. Programmet kördes av varje användare på dennes egna dator, tablådatat (som annars skulle kunna tänkas falla under URL 49§) distribuerades aldrig, och användes enbart för enskilt bruk (URL 12 §).

En dag införde webbplatsägaren ett skydd gentemot den här roboten (en sk User-agent-blockering); istället för tablådata returnerade den en skarpt formulerad text om att handlingen som programmet försökt sig på var olaglig. Det är i och för sig inte sannolikt att den var olaglig, åtminstone inte med avseende på upphovsrätten, men i och med att skyddet infördes hade webbplatsägaren tydligt markerat att han ansåg robotens agerande som olovligt. Skyddet hade lätt kunnat kringgås genom att ändra den User-agent-sträng som roboten presenterade sig med, men då hade jag verkligen medvetet olovligen berett mig tillgång till en upptagning (sedermera löstes problemet genom att någon annan skrev en ersättare som hämtade information från ett annat ställe).

Det finns för övrigt mycket som talar för att en webbplatsägare som inte, medelst robots.txt eller user-agent-skydd, sätter upp policies mot webrobotar, ska anses tycka att webrobotar är välkomna. Flera framstående webbplatser (Amazon[7], Livejournal[8], Yahoo[9], Google[10], Flickr[11]), uppmuntrar aktivt hobbyprogrammerare att, medelst webrobotar, använda webbplatsernas information och funktion för egna tillämpningar. Sökmotorer skulle vara praktiskt omöjliga om access som inte explicit tillåtits var att anse som otillåten.

Då REP är en etablerad sedvänja på Internet anser jag att det är rimligt att säga att en robot som respekterar robots.txt inte olovligen bereder sig tillgång[12]. För att en gärning ska kunna anses som olovlig måste den rimligtvis bryta mot en mer eller mindre klart uttryckt riktlinje. Då REP är den standard som gäller för att ge sådana riktlinjer till robotar, och då REP är eller borde vara bekant för en webbmaster, är det rimligt att säga att han bör begagna denna mekanism för att uttrycka sin access-policy.

5 Robots.txt som eventuellt förvaltningsbeslut

Som tidigare angetts har domstolsverket numera en REP-uttryckt policy att webrobotar inte får accessa deras tjänst. Det löser deras problem med Google, men gör samtidigt att jag inte, enligt den sedvänja som gäller på Internet, får hämta data från tjänsten med min webrobot. Detta kommer att leda till att listorna med rättsfall under paragraferna på lagen.nu gradvis kommer att bli mer och mer irrelevanta.

Jag skulle kunna utföra samma arbete manuellt, då det rör mellan 10 och 50 nya rättsfall i veckan, vilket inte är otänkbart att ladda ner ”för hand” – det skulle ta kan-

---

[5] http://membled.com/work/apps/xmltv/

[6] ”Home Theater Personal Computer”: en dator, byggd på standard-PC-arkitektur som har till uppgift att spela in och upp film, musik etc i en vardagsrumsmiljö. Det är vanligt att datorentusiaster bygger sina egna HTPC-lösningar från grunden med hård- och mjukvarukomponenter.

[7] http://www.amazon.com/gp/aws/landing.html

[8] http://www.livejournal.com/doc/server/ljp.csp.xml-rpc.protocol.html

[9] http://developer.yahoo.net/

[10] http://www.google.com/apis/

[11] http://www.flickr.com/services/api/

[12] Att det i ett senare skede kan medföra upphovsrättsintrång är en annan fråga.

ske 1-2 timmar i veckan. Det kommer dock, av praktiska skäl, inte bli gjort om det inte låter sig göras manuellt.

Frågan är; är beslutet att införa denna policy förenligt med TF 2:1, FL 4 § och verksförordningen 7 § 2 st 5? Samt, är själva instiftandet av policy en del av domstolsverkets faktiska verksamhet, eller är det möjligtvis ett förvaltningsbeslut? Beroende på svaret på den frågan finns det sedan två vägar framåt.

5.1 Förvaltningsbeslut

Vad som talar för att det är ett förvaltningsbeslut är att det kan ses som ett ”uttalande varigenom en myndighet vill påverka andra förvaltningsorgans eller enskildas handlande”[13]. Det innehåller ett direkt handlingsmönster. Som sådant kan det överklagas genom förvaltningsbesvär. När domstolsverket tar ställning till besväret bör de iaktta tre stadganden:

- TF 2:1 om varje medborgares rätt att ta del av allmänna handlingar – vilket det är fråga om i det här fallet.
- FL 4 § som stadgar att myndigheter har skyldighet att bland annat lämna upplysningar till enskilda i frågor som rör myndighetens verksamhetsområde. FL gäller inte domstolarnas dömande verksamhet, men domstolsverket är i sig en förvaltningsmyndighet (Förordning (1988:317) med instruktion för Domstolsverket, 1 §)
- Verksförordningen 7 § 2 st 5, som stadgar att verksamhetens chef skall se till att allmänhetens och myndigheten underlättas genom bland annat god service och tillgänglighet.

Å andra sidan bör domstolsverket även iakttaga två riktlinjer:

- Enligt statskontorets riktlinjer bör myndigheter som gör diarier tillgängliga på Internet se till att ”uppgifter i diarierna åtminstone inte bör kunna sökas med hjälp av konventionella, allmänt tillgängliga sökmotorer, t.ex. Altavista och Google”[14].
- Enligt EU’s artikel 29-grupp bör myndigheter som publicerar information med personuppgifter ”Främja användandet av tekniska hjälpmedel som är till för att förhindra automatiskinsamling av uppgifter som finns tillgängliga online”[15] – REP nämns specifikt som ett sådant hjälpmedel.

Dessa bägge riktlinjer talar för att det, om det är ett beslut, i alla fall inte är ett självständigt sådant fattat av Domstolsverket.

5.2 Faktisk verksamhet

Vad som talar för att det inte är ett förvaltningsbeslut, utan faktisk verksamhet, är att syftet med ändringarna i första hand är gjorda för att undvika att information med personuppgifter blir sökbart via Google och andra sökmotorer, och även kanske att skydda systemet från överlastning på grund av ”aggressiva” webrobotar. Detta kan anses falla inom ramen för en systemadministratör eller webbmasterns yrkesroller, två roller som vanligtvis inte tar förvaltningsbeslut.

Om det inte är att se som ett förvaltningsbeslut, följer att policyn inte är avsedd att påverka enskildas handlande. Med det som utgångspunkt kan man anse att en webrobot som ignorerar robots.txt därmed inte olovligen bereder sig tillgång till data – för att handlingen ska anses som olovlig måste ju den policy som robots.txt uttrycker vara avsedd att påverka enskildas handlande, och då måste det i sin tur ha varit frågan om ett förvaltningsbeslut.

---

[13] RÅ 2004 ref. 8

[14] Statskontorets rapport 2003:1, s 12

[15] http://europa.eu.int/comm/justice_home/fsj/privacy/docs/wpdocs/1999/wp20sv.pdf - Yttrande 3/99 s 9 f.