Statistikcentralens allmänna verksamhetsprinciper för webbskrapning

När Statistikcentralen skrapar information från internet för statistikproduktionen följs följande allmänna principer:

För användning av materialet som insamlas med webbskrapning görs ett beslut om datainsamling och materialet beskrivs i Datainsamlingsregistret. I anslutning till webbskrapning som baserar sig på uppgiftsskyldighet följs statistikmyndighetens normala uppgifts- och samrådsskyldigheter. Risker förknippade med kvaliteten på det material som samlats in med webbskrapning identifieras och bedöms innan beslut om webbskrapning fattas. I samband med publiceringen av statistiken anges om webbskrapning har använts som datainsamlingssätt.

Lagenlighet. Lagstiftningen beaktas fr.o.m. planeringsskedet och den följs fullskaligt. Eventuella förändringar i rättsläget följs upp.

Transparens. Webbskrapningar meddelas offentligt på Statistikcentralens webbplats och samtidigt anges följande uppgifter: syftet med webbskrapning, uppgiftstyper som är föremål för webbskrapning samt kontaktinformation som den webbansvarige kan använda. Vid insamling av uppgifter som baserar sig på uppgiftsskyldigheten beaktas informationsskyldigheten enligt statistiklagen. Om materialet innehåller personuppgifter, publiceras uppgifter som gäller behandlingen av personuppgifter öppet och lättillgängligt på Statistikcentralens webbplats.

Principen om minsta olägenhet. Webbskrapningen genomförs till alla delar så att webbplatsens funktion och dess ägare förorsakas så lite olägenheter och kostnader som möjligt.

Rätt att förbjuda. Den webbansvarige ges rätt att förbjuda skrapning genom att kontakta Statistikcentralen. Begäran om förbud respekteras och meddelas på en lista.

Iakttagande av statistikföringsprinciper. De förfaringssätt och principer som tillämpas på framställning av statistik samt de yrkesetiska principerna inom statistik- och forskningsbranschen iakttas.

Revidering av användningsvillkoren. Webbskrapning riktas endast till sådana webbplatser i vars användningsvillkor webbskrapning inte uttryckligen är förbjuden eller där förbudet tydligt har begränsats till att gälla kommersiell verksamhet. I oklara situationer kan man kontakta den webbansvarige. Om man inte får något svar inom rimlig tid, kan webbplatsen skrapas.

Utöver de allmänna principerna följer Statistikcentralen följande praktiska verksamhetsprinciper vid webbskrapningen:

Uppgifternas nödvändighet. De uppgifter som skrapas ska på goda grunder vara nödvändiga för framställning av statistik och tillföra ett mervärde vid statistikproduktionen.

Användningsändamål. Material som samlats in med hjälp av webbskrapning får bara lämnas ut för ändamål som anges i 13 § i statistiklagen.

Avslöjande av identitet (user agent string). Statistikcentralens identitet, kontaktställe för kontakttagning samt en länk till meddelandet om webbskrapning finns på Statistikcentralens webbplats.

Belastningsminimering. Webbplatserna belastas inte med överflödiga förfrågningar utan förfrågningarna görs med rimliga intervaller. Skrapningen sker i mån av möjlighet vid sådana tidpunkter då man inte kan vänta sig mycket trafik på webbplatsen. Ytterligare förfrågningar görs inte.

Förhandssamråd i undantagsfall. Den webbansvarige ska höras i sådana fall där webbskrapningen kommer att vara exceptionellt omfattande eller betungande.

Situationsspecifik bedömning. Ändamålsenligheten med webbskrapning beroende på situation utreds innan skrapningen inleds.

Robots.txt. Om webbplatsen har en robots.txt-fil som förbjuder webbskrapning, respekteras den. För att frångå robots.txt-filen krävs i förväg ett skriftligt godkännande av den webbansvarige.

Om material som anskaffats med hjälp av webbskrapning skaffas från en tredje part, följs följande principer:

Statistikcentralen kan skaffa webbskrapat material bara i det fall att verksamhetssätten hos den leverantör som producerat materialet inte strider mot Statistikcentralens verksamhetsprinciper för webbskrapning. Innan ett kontrakt upprättas ska det säkerställas att anskaffningen inte ens indirekt leder till att man bryter mot de principer som definierats i denna anvisning eller till anskaffande av uppgifter som skrapats olovligt.

Leverantören ska visa att materialet har anskaffats med tillstånd antingen så att det inte har skrapats i strid med villkoren för användning av sidorna eller så att man har kommit överens med den webbansvarige som förbjuder skrapning om skrapningen och rätten att sälja materialet. Dessutom ska materialet också i övrigt ha inhämtats på ett etiskt hållbart sätt och det får inte t.ex. ha kopierats från en databas som skyddas av upphovsrätt.

Material som innehåller personuppgifter kan i regel inte skaffas färdigt, utan skrapningen ska ske på basis av Statistikcentralens uttryckliga uppdrag så att Statistikcentralen och uppdragstagaren har kommit överens om behandlingen av personuppgifter.

Marjo Bruun Generaldirektör

Timo Koskimäki Överdirektör för statistikproduktionen

Bilaga: Verksamhetsprinciper för webbskrapning vid Statistikcentralen

Verksamhetsprinciper för webbskrapning vid Statistikcentralen

Inledning

Digitaliseringen och ökningen av datavolymerna utmanar och möjliggör också statistikproduktionen och utvecklingen av den på ett nytt sätt. På internet finns tillgängligt allt mer omfattande information, både om företag och personer, som också skulle kunna användas i statistikproduktionen.

Även om användningen av uppgifter som tagits från internet är en utmaning både för statistikproduktionssystemen och för de metoder som används, finns det också många goda skäl att använda dem.

Enligt statistiklagen ska de uppgifter som behövs för framställning av statistik samlas in så effektivt som möjligt och uppgiftslämnarbördan minimeras: vad vore då bättre än att använda uppgifter som redan finns på internet.

Uppgifter som skrapats från nätet kunde också vara en lösning på datainsamlingarnas sjunkande svarsgrader och med hjälp av dem kunde man minska de arbetsintensiva och därmed kostsamma direkta datainsamlingarna. Statistikproduktionen kunde också effektiveras på detta sätt. Till exempel i konsumentprisindexet har man med goda resultat prövat på att skrapa prisuppgifter från företagens webbsidor.

Nya informationsbehov, för vilka man överväger att skaffa material, uppstår hela tiden. Färdiga register finns inte alltid att tillgå, eller så skulle det vara för dyrt eller besvärligt att samla in uppgifter med hjälp av direkta datainsamlingar. Webbskrapning kunde också vara en lösning på sådana situationer.

Även om många saker talar för utnyttjande av uppgifter som skrapats på internet, är det inte problemfritt att använda uppgifterna i statistikföringen. När det gäller webbskrapning stöter man förutom på brister i kvaliteten på information också på både etiska och juridiska problem. Kan man i statistikproduktionen använda uppgifter som finns på webbsidorna och som enligt webbsidans användningsvillkor är förbjudet att skrapa? Och ska man be om tillstånd av dem som uppgifterna gäller att använda det material som de producerar? Hur ska man förhålla sig till en situation där den webbansvarige tillåter webbskrapning endast mot betalning?

Inom det europeiska statistiksamarbetet funderar man på samma utmaningar. Den första riktlinjen för användning av uppgifter som skrapats på nätet (ESS Web scraping policy template) publicerades i juli 2019 och följer praxis som presenteras i detta dokument.

I denna anvisning finns en sammanställning av Statistikcentralens allmänna verksamhetsprinciper för användningen av uppgifter som skrapats från internet inom statistikproduktionen. Dokumentet uppdateras om lagstiftningen ändras eller ny information eller nya anvisningar om webbskrapning annars är tillgängliga.

1. Lagstiftning som inverkar på webbskrapning

Statistiklagen (280/2004) styr framställningen av statistiken. Statistiklagen eller annan lagstiftning innehåller inga egentliga bestämmelser om webbskrapning. Det enda undantaget är lagen om deponering och förvaring av kulturmaterial (1433/2007), enligt vilken Nationalbiblioteket har till uppgift att söka och lagra webbmaterial som är tillgängligt för allmänheten. Det finns inga vedertagna tolkningar eller rättspraxis som gäller webbskrapning inom statistikproduktionen och praxisen håller först nu på att utformas även på internationell nivå. Webbskrapning ska för närvarande granskas mot bakgrund av den allmänna lagstiftningen och lagstiftningen om framställning av statistik. Vid bedömningen av webbskrapningens lagenlighet bör man särskilt beakta tre synvinklar: upphovsrätt, dataskydd och användarvillkor.

Upphovsrätter som eventuellt hänför sig till innehållet på webbsidorna begränsar inte webbskrapningen i sig, eftersom Statistikcentralens intresse är i stället för verk, information eller material som publicerats på webbsidorna den information som förmedlas om dem. Upphovsrätterna inverkar inte t.ex. på skrapning av uppgifter om företaget på företagets webbsidor. Vissa onlineplattformar anser dock att de utgör en databas. Enligt upphovsrättslagen (404/1961) har en framställare av en databas som kräver väsentlig insats ensamrätt till databasen i enlighet med EU: s databasdirektiv samt rätt att hindra kopiering eller återanvändning av databasen (sui generis-rätt). För att kunna åtnjuta sui generis-skydd har EU-domstolens rättspraxis dock krävt så betydande ekonomiska eller tidsmässiga resurser vid framställning av databaser att ytterst få plattformar uppfyller dessa kriterier. Även då bör man beakta att om databasen är öppen för allmänheten kan den som upprättar databasen enligt domstolen inte förbjuda utomstående att söka information där.

Om de sidor som ska skrapas innehåller personuppgifter, ska också dataskyddslagstiftningen beaktas innan webbskrapningen inleds. Om behandlingen av personuppgifter föreskrivs i EU: s allmänna dataskyddsförordning (Europaparlamentets och rådets förordning (EU) 2016/679) och i dataskyddslagen som kompletterar den (1050/2018). Den person som personuppgifterna gäller kallas registrerad. Särskild uppmärksamhet ska fästas vid att de personuppgifter som ingår i det material som ska skrapas till alla delar är ändamålsenliga, väsentliga och nödvändiga med tanke på framställningen av statistiken. Om man laddar ned webbplatsen i samband med skrapningen kan det dock leda till att man får med också annat material än det som behövs vid statistikproduktionen. Om sådant temporärt material som lagrats som biprodukt av en teknisk process oavsiktligt innehåller personuppgifter, ska det förstöras eller anonymiseras utan dröjsmål.

De registrerade behöver inte informeras personligen om webbskrapning av deras uppgifter, eftersom det skulle medföra oskäligt besvär och förhindra att det statistiska målet uppnås i enlighet med det undantag som dataskyddsförordningen tillåter. I sådana fall är kollektiv information tillräcklig och görs genom att man meddelar om webbskrapning och därtill hörande behandling av personuppgifter på Statistikcentralens webbplats. Uppgifterna ska täcka kraven i dataskyddsförordningen och de ska vara tydligt angivna och lätta att hitta.

Också de registrerades rätt att rätta, kontrollera, begränsa och motsätta sig uppgifter kan undantagsvis begränsas med stöd av dataskyddslagen, om framställningen av statistiken eller behovet av information kräver det. En avvikelse från den registrerades rättigheter förutsätter alltid att Statistikcentralens dataskyddsombud hörs och att en konsekvensbedömning genomförs.

Användningen av uppgifterna på webbplatsen kan begränsas med användarvillkor. För att användningsvillkoren ska bli tillämpliga måste webbplatsens användare godkänna dem t.ex. i samband med registrering eller besök på webbplatsen. Vissa onlineplattformar förbjuder webbskrapning i sina användarvillkor genom att t.ex. nämna att ”användaren har inte rätt att använda automatiska system för mångfaldigande” eller att ”det är förbjudet att använda automatiserad programvara för kopiering av information från en webbplats” Eftersom det huvudsakliga syftet med sådana villkor har varit att begränsa den kommersiella användningen av uppgifter, finns det olika tolkningar av deras betydelse för statistikmyndigheternas webbskrapning. En stark huvudregel är dock att villkoren för användning av webbplatser inte tar ställning till webbskrapning.

2. Allmänna principer för webbskrapning

Vid användning av material som skaffas genom webbskrapning fattas alltid beslut om datainsamling och materialet beskrivs i datainsamlingsregistret på samma sätt som annat material som skaffas. I anslutning till webbskrapning som baserar sig på uppgiftsskyldighet följs statistikmyndighetens normala uppgifts- och samrådsskyldigheter.

Risker förknippade med kvaliteten på det material som samlats in med webbskrapning identifieras och bedöms innan beslut om webbskrapning fattas. Det är ofta inte möjligt att till alla delar kontrollera att materialet är korrekt och aktuellt. Användningen av webbskrapning som datainsamlingssätt ska därför uttryckligen anges i samband med publiceringen av statistik som bildats på basis av det skrapade materialet. Om man vill använda webbskrapade uppgifter t.ex. inom maskininlärning, måste man också kunna försäkra sig om att materialet är opartiskt.

För att säkerställa etisk hållbarhet ska följande principer följas i all webbskrapning som Statistikcentralen bedriver:

Lagenlighet. Lagstiftningen, inklusive dataskydd, kommer att beaktas redan vid planeringen av skrapningen och tillämpas fullt ut.

Eventuella förändringar i rättsläget (lagstiftning, rättspraxis, vedertagna tolkningar) följs upp.

Transparens. Webbskrapningarna tillkännages offentligt på Statistikcentralens webbplats. Samtidigt ska följande uppgifter ges för varje webbskrapning: syftet med skrapningen, de datatyper som skrapningen gäller samt kontaktuppgifter för den webbansvarige för att begära ytterligare information eller begränsa skrapningen. Om det är fråga om insamling av uppgifter som baserar sig på uppgiftsskyldighet, då uppgifterna samlas in beaktas dessutom informationsskyldigheten enligt statistiklagen. Skyldigheten att informera om behandlingen av personuppgifter iakttas.

Principen om minsta olägenhet. Webbskrapningen genomförs till alla delar så att webbplatsens funktion och dess ägare förorsakas så lite olägenheter och kostnader som möjligt.

Rätt att förbjuda. Den webbansvarige har rätt att förbjuda skrapning (opt-out) genom att kontakta Statistikcentralen. Begäran om förbud respekteras och meddelas på en gemensam lista (s.k. black list).

Iakttagande av statistikföringsprinciper. De förfaringssätt och principer som tillämpas på datainsamling samt på utveckling och framställning av statistik iakttas också vid webbskrapning. Detsamma gäller de yrkesetiska principer inom statistik- och forskningsbranschen som ligger till grund för Statistikcentralens verksamhet.

Revidering av användningsvillkoren. Webbskrapning riktar sig tills vidare endast till sådana webbplatser vars användningsvillkor har kontrollerats. Skrapning anses vara tillåten om den inte uttryckligen förbjudits eller om förbudet tydligt har begränsats till att gälla endast kommersiell verksamhet.

I oklara situationer kan man kontakta den webbansvarige. Om man inte får något svar inom rimlig tid, kan webbplatsen skrapas.

3. Praxis vid webbskrapning

Utöver de allmänna principerna följer Statistikcentralen följande praktiska verksamhetsprinciper vid webbskrapningen:

Uppgifternas nödvändighet. Webbskrapning riktas bara till sådana uppgifter som med goda grunder är nödvändiga för framställningen av statistik. Uppgifterna måste tillföra statistikproduktionen ett mervärde.

Användningsändamål. Material som samlats in med hjälp av webbskrapning får bara lämnas ut för ändamål som anges i 13 § i statistiklagen.

Avslöjande av identitet (user agent string). Statistikcentralens identitet, kontaktställe för kontakttagning samt en länk till meddelandet om webbskrapning finns på Statistikcentralens webbplats.

Belastningsminimering. Webbplatserna belastas inte med överflödiga förfrågningar utan förfrågningarna görs med rimliga intervaller. Skrapningen sker i mån av möjlighet vid sådana tidpunkter då man inte kan vänta sig mycket trafik på webbplatsen (t.ex. nattetid). Ytterligare förfrågningar genomförs inte, utan skrapningarna genomförs så att de bara söker nödvändiga uppgifter.

Förhandssamråd i undantagsfall. Den webbansvarige ska höras i förväg i sådana fall där webbskrapningen kommer att vara exceptionellt omfattande eller betungande.

Situationsspecifik bedömning. Ändamålsenligheten med webbskrapning beroende på situation utreds innan skrapningen inleds. Uppgifterna kan också sökas via API, om en sådan har erbjudits.

Robots.txt. Om webbplatsen har en robots.txt-fil som förbjuder webbskrapning, respekteras den. Vid behov kan man fråga den webbansvarige (skriftligt) om lov för att frångå Robots.txt-filen. Webbskrapningen ska inte inledas innan ett jakande svar har erhållits.

4. Anskaffning av material som tredje part webbskrapat

På marknaden finns företag vars affärsverksamhet baserar sig på försäljning och utnyttjande av webbskrapat material. Att skaffa intressant information direkt av

kommersiella aktörer som bedriver webbskrapning kan vara ett lockande alternativ. Utöver utredningen av kostnadseffektiviteten ska man dock säkerställa att man vid webbskrapningen följer samma principer oberoende av vem – Statistikcentralen eller en tredje aktör – som har skaffat materialet.

Innan ett kontrakt upprättas ska det säkerställas att anskaffningen inte ens indirekt leder till att man bryter mot de principer som definierats i denna anvisning. Särskild uppmärksamhet ska fästas vid att materialet inte innehåller uppgifter som skrapats utan tillstånd. Leverantören ska antingen visa att materialet inte innehåller uppgifter som skrapats i strid med användningsvillkoren eller att man har kommit överens om webbskrapningen och rätten att sälja materialet med den webbansvarige som förbjuder skrapning. Dessutom ska materialet också i övrigt ha inhämtats på ett etiskt hållbart sätt (belastningsminimering, robots.txt, transparent identitet) och det får inte t.ex. ha kopierats från en databas som skyddas av upphovsrätt.

Material som innehåller personuppgifter kan i regel inte skaffas färdigt, utan skrapningen ska ske på basis av Statistikcentralens uttryckliga uppdrag så att Statistikcentralen och uppdragstagaren har kommit överens om behandlingen av personuppgifter.