

Nota van Inlichtingen Marktraadpleging in verband met Web scraping

Datum: 5 augustus 2020

In deze nota van inlichtingen geeft de NVWA antwoord op de in het kader van de hierboven genoemde marktraadpleging gestelde vragen en geplaatste opmerkingen. Deze nota van inlichtingen maakt onderdeel uit van de marktraadpleging 'Web scraping' met kenmerk IUC: 202001116 en NVWA/2020/024 d.d. 20 juli 2020.

Let op! De NVWA zou graag de reacties uiterlijk donderdag 13 augustus 2020 voor 9.00 uur ontvangen. Mocht de NVWA naar aanleiding van de reacties nog vragen hebben, dan zou de NVWA graag vrijdag 14 augustus 2020 tussen 13.00 en 17.00 digitaal (Via Skype voor Bedrijven) contact hebben met de Geïnteresseerde. Het verzoek aan u is om de gevraagde gegevens bij hoofdstuk 3 Informatieaanvraag volledig in te vullen, zodat er eventueel met u een afspraak ingepland kan worden op vrijdag 14 augustus 2020.

Vraag	Vraag	Antwoord
1.	Werking van de tool: 1)weet u welke websites u wilt scrapen, 2) of moet het product in staat zijn om een zoekopdracht uit te voeren (op bijvoorbeeld Google) met betreffende trefwoord(en) en de ruwe data van de zoekresultaten scrapen? 3) Zo ja, hoeveel zoekresultaten dienen gescraped te worden?	1) De NVWA wil scrapen op basis van een zoekterm. Alle websites waar deze specifieke zoekterm op voorkomt moeten aangeleverd worden. Op basis hiervan wil de NVWA aanbieders online kunnen achterhalen en ook informatieve websites in beeld krijgen. Met andere woorden; de markt moet in kaart worden gebracht. Webscraping is een middel om het werk van de NVWA te kunnen doen. Er is geen minder ingrijpend middel ter beschikking. Bij iedere webscrape zoekt de NVWA zo gericht mogelijk. En op het moment dat de NVWA onverwachts data tegenkomt die niet nodig is, wordt deze meteen verwijderd. De data bewaart de NVWA zo lang als nodig is voor het doel, niet langer. 2) Ja, hier moet het product inderdaad toe in staat zijn. 3) Het aantal zoekresultaten valt niet te bepalen op voorhand.
2.	Werking van de tool: kunt u meer informatie geven wat voor type websites u verwacht te scrapen?	De NVWA verwacht alleen voor publiek toegankelijke sites te gaan scrapen.
3.	Werking van de tool: op wat voor frequentie wilt u scraping opdrachten uitvoeren?	De frequentie is op dit moment nog niet bekend.
4.	Werking van de tool: dienen er analyses uitgevoerd te worden op de gescraped data, zoals categoriseren van data of sentiment analyse?	Nee, er hoeven geen analyses door opdrachtnemer uitgevoerd te worden.
5.	User Interface: wilt u in staat zijn om zelf (d.m.v. een UI) een scrape opdracht aan te maken o.b.v. trefwoorden? Zo ja, wie is de primaire	Nee, de NVWA hoeft niet zelf in staat te zijn om een scrape opdracht aan te maken.

	gebruiker van de UI, hoeveel gebruikers verwacht u en wat voor gebruiksfrequentie verwacht u?	
6.	User Interface: 1) wilt u in staat zijn om zelf (d.m.v. UI) de voortgang van een scrape opdracht te volgen en 2) de resultaten in te zien wanneer deze voltooid is?	1) Nee, de NVWA hoeft de voortgang van een scrape niet te volgen. 2) Ja, de NVWA wil de resultaten zien wanneer een scrape opdracht is voltooid.
7.	Scripts: dient voor elke scraping opdracht een nieuw script geschreven te worden, en is daarmee elke opdracht maatwerk?	Ja, waarschijnlijk betreft het merendeel maatwerk.
8.	Data: kunt u meer informatie geven wat voor ruwe data u verwacht te ontvangen?	De NVWA verwacht de website, de specifieke pagina en evt. aanvullende gegevens zoals adres- of bijvoorbeeld kvk-gegevens.
9.	Data: zijn er eisen aan de ruwe data behalve dat deze ontsloten moet worden d.m.v. een API of in SQL formaat?	Bij nader inzien: Data mag aangeleverd worden in een Excel-bestand.
10.	Data: 1) dient de leverancier van de scrapers een database aan te houden met gescraped ruwe data? 2) Zijn er, naast eventueel verwijderen van ruwe data en/of persoonsgegevens uit de ruwe data, andere eisen vanuit AVG of b.v. archiefwetgeving waar rekening mee gehouden dient te worden?	1) Nee. 2) Nee. Het is vanuit de NVWA een 1-malige opdracht waarna alles verwijderd mag worden. Belangrijkste eis uit de AVG is dat op zijn minst beschreven moet worden wat precies voor welk gerechtvaardigd doel gescraped zou moeten worden. En of dat proportioneel, subsidiair en niet meer dan noodzakelijk is.
11.	Data: bent u er zich van bewust dat op de gescraped ruwe data mogelijk rechten rusten zoals uitgeversrechten / IP rechten / auteursrechten of overige rechten op de inhoud? Bij wie (NVWA of leverancier) ligt eventuele juridische aansprakelijkheid omtrent het en/of voortvloeiend uit de data?	NVWA scraped niet op inhoud. Alleen op het voorkomen van de zoekterm zelf op een bepaalde pagina. NVWA slaat de inhoud van de pagina niet op. Nadrukkelijk is de NVWA er niet op uit om inbreuk te maken op IE rechten (zoals auteursrecht).
12.	Doelsysteem: kunt u aangeven wat voor systemen de ruwe data gaan inladen (d.m.v. een API of vanuit SQL)?	Data mag aangeleverd worden via een Excel-bestand.
13.	Ondersteuning: wat voor ondersteuning verwacht u behalve het schrijven van scrapers (scripts)?	De NVWA verwacht een kennispartner die veel ervaring heeft op het gebied van scrapen en actief de ins en de outs met ons wil delen. Bijv. door middel van best practices.
14.	Algemeen: overweegt u om data vanuit web scraping te combineren met andere data in één overzichtelijke tool die gebruikers de meest relevante data toont op basis van analyses en Machine Learning?	De insteek van de NVWA is niet dat er een tool beschikbaar komt voor de eindgebruiker.
15.	Samenwerking: wat voor contracttermijn heeft u in gedachten? Kunt u meer inzicht geven in de verwachte samenwerking (bijvoorbeeld frequentie / intensiteit, samenwerkingsvorm)?	De NVWA denkt er op dit moment over na om een raamovereenkomst te sluiten met één opdrachtnemer. De looptijd wordt waarschijnlijk één jaar. Daarbij wil de NVWA zich het recht voorbehouden om de

		raamovereenkomst te verlengen met maximaal drie (3) keer twaalf (12) maanden.
16.	Op welke bestaande systemen moet de informatie gaan aansluiten via API of SQL?	Zie het antwoord op vraag 9 en 12.
17.	Moet de informatie vanuit de scraper gevisualiseerd en geanalyseerd kunnen worden?	Nee.
18.	Is integratie met de bestaande systemen wenselijk?	Nee, integratie is niet wenselijk.
19.	Welke informatie dient gescraped te gaan worden? Fora? Darkweb? Social Media? Artikelen? Graag verduidelijking.	Zie het antwoord op vraag 1. Het kan gaan om alle openbare databronnen. Zodra er een toegankelijkheidsdrempel gepasseerd moet worden, wordt afgedaan aan het openbare karakter.
20.	Zou er ook interesse zijn voor gebruiksvriendelijke software matige oplossingen die on premise hetzelfde werk kunnen doen, zonder benodigde python/R kennis en dus zelfstandig beheer?	Nee, hier heeft de NVWA geen interesse in.
21.	Wanneer de keuze gemaakt wordt voor toch scripting, zou er dan interesse zijn om gebruikers van het NVWA op te leiden, zodat ze zelf nog wat beheer/aanpassingen aan eventueel geschreven scripts kunnen maken en om hoeveel gebruikers zou dit gaan?	Nee, de NVWA heeft geen interesse in een opleiding om zelf nog beheer/aanpassingen aan scripts te kunnen maken.
22.	Wordt er samengewerkt met bevriende diensten, zo ja, welke kunnen genoemd worden?	De NVWA werkt wel eens samen met andere toezichthouders.
23.	Om hoeveel gebruikers zou het idealiter gaan?	De hoeveelheid gebruikers is voor deze opdracht, volgens de NVWA, niet relevant. Het gaat om het opleveren van Excel-bestanden van de zoekresultaten.
24.	Bestaat er al budget voor dit traject en is daar iets over te zeggen?	Op dit moment is hier nog niets over te zeggen.
25.	Bedoeld u met de audittrail een interne of externe audit?	Bij nader inzien is een audittrail niet nodig.
26.	Welke websites wil de NVWA scrapen?	Zie het antwoord op vraag 1.
27.	Welke informatie wil de NVWA ontvangen?	Zie het antwoord bij vraag 8.
28.	Hoeveel wil de NVWA scrapen?	Zie het antwoord bij vraag 3.
29.	Mag er gewerkt worden met rotating proxies?	Ja, er mag gewerkt worden met rotating proxies.
30.	Is er een voorkeur voor script geschreven in Python of PHP?	PHP geniet de voorkeur.
31.	Vraagt de NVWA kwaliteitsgaranties bij de te leveren informatie?	Ja, de kwaliteitsgaranties zullen beschreven worden in het programma van eisen. Belangrijk daarbij is dat data openbaar, betrouwbaar en volledig moet zijn.

32.	Heeft de NVWA de wens om JavaScripted-based websites te laten scrapen, zoals ReactJs of Angular die HTML-gegevens dynamisch creëren?	De NVWA focust zich hier niet specifiek op. De verwachting van de NVWA is dat de websites die opdrachtnemer moet scrapen dit niet hebben.
33.	Wat moet de frequentie van het scrapen zijn?	Zie het antwoord bij vraag 3.
34.	Heeft de NVWA de behoefte om opdracht aan te gaan op maandelijkse basis, waarbij facturering op maandelijkse basis plaatsvindt?	De NVWA zou een raamovereenkomst willen aangaan. De frequentie van het aantal opdrachten is op voorhand niet aan te geven. De frequentie van facturering heeft de NVWA bij voorkeur zo laag mogelijk en moet gekoppeld zijn aan een volledig afgeronde opdracht.
35.	1) Heeft de NVWA de voorkeur om de webscraping resultaten van een externe dienst te ontvangen of 2) wenst men een webscraping product / oplossing die in eigen beheer genomen kan worden?	1) Ja, dit heeft de voorkeur. 2) Op dit moment is deze wens er niet, maar de NVWA heeft wel interesse om de eventuele mogelijkheden te bespreken.
36.	Voor welke soort websites verwacht de NVWA webscraping te gaan gebruiken? Publiekelijk toegankelijke sites of interne en beveiligde sites, waar rekening gehouden moet worden met gevoelige data?	Zie het antwoord op vraag 19.
37.	Zijn de bronnen die gescraped moeten worden bekend of moeten die ad-hoc opgezocht worden?	De bronnen moeten ad-hoc opgezocht worden.
38.	Hoeveel bronnen is de verwachting dat er zijn?	Dit is op voorhand niet aan te geven. Het internet heeft oneindig veel bronnen.
39.	1) Welke informatie moet er gescraped worden. 2) Is het hierbij nodig dat er business intelligence in de scripts wordt ingebouwd? In andere woorden: wat voor gegevens zijn jullie naar op zoek (wat zijn x en y?).	1) Zie het antwoord bij vraag 8. 2) Business intelligence is niet nodig.
40.	Voor welke doeleinden wil de afdeling Handel- en Digitaal Toezicht Web Scraping toepassen?	De afdeling wil de markt in kaart brengen, zodat er toezicht gehouden kan worden op de digitale handel. Op grond van de taak van de NVWA om toezicht te houden, moet de NVWA dit middel inzetten om het werk te kunnen doen.
41.	Welke data beoogt de afdeling Handel- en Digitaal Toezicht te verzamelen met Web Scraping?	Zie het antwoord bij vraag 8.
42.	Welke data beoogt de afdeling Handel- en Digitaal Toezicht te verrijken met de door Web Scraping verzamelde data?	Dit betreft informatie die de NVWA, in ieder geval op dit moment, niet met derden wil delen.
43.	Over welke data subjecten (natuurlijke personen, organisaties etc.) verwacht de afdeling Handel- en Digitaal Toezicht data te willen verzamelen?	Aanbieders van producten- en diensten online in kaart brengen.
44.	In hoeverre zijn de beoogde databronnen legaal te kopiëren en te gebruiken voor de doeleinden van de NVWA?	De NVWA is slechts geïnteresseerd in de URL van de website en de specifieke pagina's waarop de zoekterm te vinden is.

45.	In hoeverre is bekend of voor de beoogde databronnen een autorisatie nodig is, of dat deze openbaar toegankelijk zijn?	Zie het antwoord op vraag 19.
46.	Als er een wijziging plaatsvindt in een databron, kan het voorkomen dat de Web Scraping niet (juist) meer werkt. Hoe zal worden om gegaan met een veranderende databron en onderhoud van Web Scraping? Bij wie wordt deze verantwoordelijkheid belegd?	De opdrachten die de NVWA heeft, zullen vooral maatwerk zijn en een eenmalig karakter hebben.
47.	In hoeverre hebben andere afdelingen binnen de NVWA, behalve de afdeling Handel- en Digitaal Toezicht, interesse en/of baat bij de toepassing van Web Scraping?	Er is geen interesse van andere afdelingen, de afdeling Handel en Digitaal Toezicht behandelt ook de vragen van andere afdelingen binnen de NVWA.
48.	Wat is het beleid van de NVWA ten aanzien van cloud-gebaseerde diensten?	Het beleid van de NVWA ten aanzien van Cloud-gebaseerde diensten is voor deze marktraadpleging ons inziens niet relevant. Dit vanwege het feit dat er gevraagd wordt naar het opleveren van Excel-bestanden.
49.	Welke kennis en/of ervaring is binnen de NVWA al aanwezig over de toepassing van Web Scraping?	Welke kennis en/of ervaring ten aanzien van Web scraping exact aanwezig is, is voor deze marktraadpleging ons inziens niet relevant.