

Nota van Inlichtingen

IUC21-027 Inzicht in Ongestructureerde Bestanden (IOB)

In het kader van bovengenoemde Marktconsultatie was er tot uiterlijk vrijdag 23 juli 2021 (24:00 uur) de mogelijkheid voor Nadere inlichtingen/stellen van vragen aan de Belastingdienst.

Door partijen zijn onderstaande vragen gesteld. Bij elke vraag is door IUC Belastingdienst in samenwerking met het projectteam het bijbehorende antwoord vermeld. De vragen en antwoorden vormen met deze Nota van Inlichtingen een integraal onderdeel van de Marktconsultatie.

Gezien de vakantieperiode wordt het uiterste moment van het insturen beantwoording verschoven van 11 naar 20 augustus 2021 (zie vraag 16).

Nota van Inlichtingen

IUC21-027 Inzicht in Ongestructureerde Bestanden (IOB)

Nr.	Datum binnenkomst	Document Markt Consultatie	Paragraaf / Eis / vraag	Vraag	Antwoord
1	22-7-2021	Pagina 5	Paragraaf 2.1	Hoe ziet de Active Directory (AD) structuur van de Belastingdienst er uit? Gaat het om 1 of meerdere domeinen? Is het Forest met meerdere child domains? Zijn dit losstaande of nested domeinen?	De AD structuur van de Belastingdienst bestaat uit een zogenoemde multi-forest single domain structuur.
2	22-7-2021			Wie worden volgens de Belastingdienst de key gebruikers van de Oplossing? Welke functies/profielen?	Er is een team belast met operationeel houden van de oplossing. Er zijn slechts enkele personen die bij de data kunnen.
3	22-7-2021	Pagina 6	Paragraaf 2.1	"De Oplossing moet kenmerken (patronen) kunnen scannen binnen bestanden "strings" en deze als metadata kunnen opslaan." Bedoelt u hiermee dat u deze metadata wil opslaan in de Oplossing of in het bronsysteem?	De metadata moeten verrijkt kunnen worden met de gevonden "string". De uitkomsten moeten in de database van het product worden opgeslagen zodat deze data in query's gebruikt kunnen worden.
4	23-7-2021			Hoe moeten we omgaan met gezippte documenten?	Ook gezippte documenten moeten gescand kunnen worden waarbij het "pad" naar het zip-bestand herleidbaar moet kunnen zijn.
5	23-7-2021			Welke bestandsformaten moeten ondersteund worden? Verwacht u bijvoorbeeld nog bestanden als Lotus123 of WordPerfect bestanden aan te treffen?	Alle door Office gebruikte/ondersteunde formaten dienen ondersteund te worden. Overige typen bestanden zijn "nice to have".
6	23-7-2021	Pagina 6	Paragraaf 2.2	Kunt u uitleggen wat er van "Metadata in de tijd bijhouden" verwacht wordt?	Met metadata in de tijd wordt bedoeld dat verschillende tijdstippen worden bijgehouden van een file. Hierbij valt te denken aan: • Scantijdstip; • Verwijder of change tijdstip. Met deze gegevens is het mogelijk om een tijdgebonden rapportage te genereren. Zie hiervoor ook vraag 12: "Beschrijf hoe en volgens welke regels scandata worden bijgehouden en eventueel worden verwijderd?"
7	23-7-2021	Pagina 6	Paragraaf 2.2	Welke type databases moeten ondersteund worden? (voor scannen comments)?	De volgende databases dienen ondersteund te worden: • Oracle; • DB2; • MS SQL.
8	23-7-2021			Moeten ook e-mails geïndexeerd en gescanned kunnen worden?	Ja.
9	23-7-2021	Pagina 6	Paragraaf 2.3	Lotus Notes mail wordt genoemd als bron. Gaat het alleen om standaard Notes mail databases (.nsf) of ook andere Notes document/.nsf type informatie?	Ja, het gaat het alleen om standaard Notes mail databases.

Nota van Inlichtingen

IUC21-027 Inzicht in Ongestructureerde Bestanden (IOB)

10	23-7-2021			Hoe gaat de Belastingdienst om met informatie die aangemaakt wordt op C: schijven. Zijn dit fysiek lokale schijven?	C: schijven zijn meestal fysieke lokale schijven, die voornamelijk in laptops voorkomen. Uitzonderingen zijn Virtuele Desktops waarbij de bestanden op VMware vSAN clusters staan.
11	23-7-2021			Is document-level security relevant (voor het scannen en bevragen van de documenten)? Hoe verwacht de Belastingdienst hier mee om te gaan?	Er zullen omgevingen zijn die niet gescand mogen worden. De voorkeur gaat er naar uit om deze te kunnen uitsluiten van de scan. Gaarne ontvangen wij de mogelijkheden die hiervoor geboden kunnen worden.
12	23-7-2021			Hoe verhoudt deze marktconsultatie zich tot de BIO?	Voor de Belastingdienst bevat de Baseline Informatiebeveiliging Overheid (BIO) een set van normen waar alle gebruikte tooling aan moet voldoen. In de eventueel volgende RFP zal de BIO als verplicht worden gesteld. De gebruikte gegevens hebben een vertrouwensniveau van BBN2 (maximaal Departementaal Vertrouwelijk (DepV), De verzamelde metadata dienen ook als zodanig beschermd te worden in uw tooling.
13	23-7-2021	Pagina 9	Paragraaf 2.9 Vraag 36	Vraag 36 meldt "Beschrijf de interface voor het afhandelen van incidenten". Kunt u aangeven wat onder een incident wordt verstaan?	Een incident betreft een issue dat uw product heeft gedetecteerd over het eigen functioneren, alsmede de issues die tijdens het scannen worden ondervonden.
14	23-7-2021			Kunt u d.m.v. 1 of 2 use-cases aangeven hoe u verwacht de service te gaan gebruiken? Wat zou een mogelijke vraag kunnen zijn van uw klanten. Wat wilt/verwacht u dan als antwoord te moeten geven?	Use case 1: Controle of de data voldoen aan de AVG regels. Hierbij wordt in het bestand gezocht naar informatie die duidt op vallen onder AVG regelgeving. En deze gegevens worden gecombineerd met bijvoorbeeld een leeftijd van een bestand van bijvoorbeeld zeven jaar of meer. Het resultaat is een lijst met namen van bestanden die uitgesplitst (en dus ook verdeeld) kan worden over verschillende afdelingen. Use case 2: Controle of de bestanden nog gebruikt worden. Hierbij wordt een selectie gemaakt van bestanden van bijvoorbeeld 5 jaar en ouder die de afgelopen 3 jaar niet meer gelezen zijn. Het resultaat is een lijst met namen van bestanden die uitgesplitst (en dus ook verdeeld) kan worden over verschillende afdelingen.
15	23-7-2021			Moeten klanten of gebruikersgroepen ook inhoudelijk door de ongestructureerde informatie kunnen zoeken (mag een gebruiker het document ook openen)?	Dat is nu nog niet duidelijk. Kunt u aangeven of uw product hiertoe in staat is alsmede het aantal gebruikers dat gelijktijdig gebruik kan maken van uw product / deze functionaliteit.

Nota van Inlichtingen

IUC21-027 Inzicht in Ongestructureerde Bestanden (IOB)

16	23-7-2021			Is het mogelijk om uitstel te krijgen vanwege de vakantieperiode?	Ja dat is mogelijk. Gezien de vakantieperiode wordt het uiterste moment van het insturen beantwoording verschoven van 11 naar 20 augustus 2021.
17	23-7-2021			Wil de Belastingdienst voor metadatering gebruik maken van landelijke richtlijnen of modellen. Zo ja, welke?	Nee, dit is op dit moment niet in scope.
18	23-7-2021			Wil de Belastingdienst voor uitvoering van WOB en WOO gerelateerd werk gevonden gegevens automatisch kunnen lakken?	Nee, dat is niet het geval. Wel willen we graag weten of dit mogelijk is en we verzoeken partijen dit dan ook in de beantwoording op te nemen.
19	23-7-2021			Wil de Belastingdienst bepaalde documenttypen verschillend kunnen analyseren als dat vanuit verschillende proces-perspectieven geschiedt? Bijvoorbeeld het onderzoeken en de metadatering van eenzelfde klachtbrief vanuit AVG- of WOB/WOO-perspectief?	Ja.
20	23-7-2021			Wil de Belastingdienst met de Oplossing kunnen zoeken in foto's of ander beeldmateriaal dat zich in Office- of Pdf-bestanden bevindt?	Ja, indien foto's of ander beeldmateriaal in Office- of PDF-bestanden bevinden dan vragen wij conform vraag 8 of deze ook doorzocht kunnen worden.
21	23-7-2021			Wil de Belastingdienst naast zoeken naar duidelijk gestructureerde gegevens als bedragen en postcodes kunnen zoeken naar minder gestructureerde gegevens zoals adressen, persoonsnamen of thema-gerelateerde tekstfragmenten?	Ja.
22	23-7-2021	Pagina 7	Paragraaf 2.5 Vraag 14	Vraag 14 uit de vragenlijst in Hoofdstuk 2 is ons niet helemaal duidelijk. Kunt u deze toelichten?	In vraag 13 geven wij aan dat we een tool zoeken die 1,5 miljard bestanden kan doorzoeken en metadata hierover kan bijhouden. In vraag 14 vragen wij de limieten van uw tooling voor het aantal bij te houden bestanden en eventueel het aantal te scannen bestanden. In het tweede deel van de vraag willen we weten of er mogelijkheden zijn om deze limieten te ondervangen. Bijvoorbeeld uw tooling kan maximaal 1 miljard bestanden opslaan, maar door de toepassing van een andere database of splitsing van de database kan dit ondervangen worden. Bij een splitsing willen we ook graag weten hoe totaaloverzichten gemaakt kunnen worden.

Nota van Inlichtingen

IUC21-027 Inzicht in Ongestructureerde Bestanden (IOB)

23	23-7-2021			Het scannen en opslaan van gegevens van 1,5 miljard objecten naar één metadatabase en analyse omgeving maakt o.i. de herkenbaarheid van de gegevens over de eigenaren van gegevensverzamelingen vrijwel onmogelijk. Het initieel reeds “splitsen” van de omgevingen in lijn met de informatie eigenaren heeft daarbij de voorkeur. Dit ook als de informatie eigenaren zelf analyses ter hand gaan nemen. Heeft ook informatie veiligheid voordelen. Vraag: kan (bij benadering) er een verdeling van de 1,5 miljard objecten over de diverse informatie eigenaren worden gemaakt?	Nee, slechts ten dele, doel is ook om data te kunnen koppelen aan de diverse eigenaren. In vraag 14 kunt u aangeven wat het maximale aantal bestanden is dat u kunt analyseren.
24	23-7-2021			Het analyseren van de inhoud van objecten voor entiteit-extractie (bv. AVG gevoeligheid) kan op diverse manieren. Twee voor de hand liggende methoden zijn; enerzijds de analyse en inventarisatie van de individuele woorden en dataelementen in het object en anderzijds een analyse op basis van woord-in-context (Natuurlijke Taal Verwerking) technologie. De aard van het bestand / document bepaalt in hoge mate de aantrekkelijkheid en effectiviteit van een methode. Vraag: kan opdrachtgever aangeven wat de hoofdtype objecten per informatieverzameling zijn (formulieren, standaard documenten, vrije tekst documenten, rapporten, emails met bijlagen, etc) en welke inhoudelijke analyse methode zij voor iedere type verkiest?	In eerste instantie ligt de focus op Office-documenten. Voor analyse wordt vooralsnog gekeken naar kenmerken van individuele woorden. Afhankelijk van de behoefte, die deels gecreëerd wordt door het bestaan van de dienst (IOB), zullen in een later stadium andere typen bestanden worden meegenomen en zal gekeken worden naar andere methoden zoals woord-in-context
25	23-7-2021			Omvangrijke ongestructureerde informatie verzamelingen bevatten doorgaans grote hoeveelheden bestanden en documenten ZONDER leesbare tekst (TIFF, non-tekst PDF, JPG, ed). Dit of als resultaat van een “platte” scanning van het document, of als resultaat van een “platte” print/output/conversie naar non-tekst PDF. Voordat inhoudelijke analyse en entiteitextractie kan plaatsvinden zal de inhoud van deze documenten middels een OCR slag moeten worden gelezen en omgezet naar leesbare tekst. Vraag: is er inzicht dan wel een inschatting te maken van het percentage non-tekst bestanden binnen de 1,5 miljard of per informatie verzameling? Vraag: waar moet de zo gelezen tekst worden opgeslagen? Ook in de metadata database?	Het doel van de nieuwe dienst is om ook inzicht te krijgen in de inhoud van de bestanden en daarom is ook geen inschatting te maken van het percentage bestanden en documenten zonder leesbare tekst. De gelezen tekst moet op een betrouwbare en veilige manier binnen uw tooling opgeslagen worden, maar dit hoeft niet persé in de metadata database opgeslagen te worden.

Nota van Inlichtingen

IUC21-027 Inzicht in Ongestructureerde Bestanden (IOB)

26	23-7-2021	Pagina 8	Paragraaf 2.6 Vraag 25	Uw vraag 25, Integratie Oplossing in een workflowtoepassing. Vraag: Betreft dit de gehele Oplossing of bijvoorbeeld het aanroepen / integreren van deelfuncties via bv. een API call. Indien het laatste kan opdrachtgever een overzicht geven van de door u te integreren hoofdfuncties?	De Belastingdienst zoekt naar workflow integratie over de gehele keten en vraagt zich af welke mogelijkheden er in uw oplossing geboden worden.
27	23-7-2021	Pagina 7	Paragraaf 2.5 Vraag 8	Uw vraag 8. Wellicht dat sommige informatie verzamelingen ook grote hoeveelheden beeld- of videomateriaal bevatten. Vraag; wat zijn uw eisen m.b.t. analyse en ontsluiting van beeld- en videomateriaal e.d.	De Belastingdienst stelt op dit moment geen eisen aan beeld en videomateriaal, maar wil graag weten welke mogelijkheden geboden worden om deze informatie om te zetten naar metadata.
28	23-7-2021	Pagina 6	Paragraaf 2.3	2.3 Bronsystemen. Vraag; welke versie(s) van het IBM FileNet systeem zijn aanwezig bij opdrachtgever en van welke onderliggende database software wordt hierbij gebruikt.	FileNet versie 5.5.5 en 5.5.6 worden gebruikt met een DB2 database.
29	23-7-2021	Pagina 6	Paragraaf 2.3	2.3 Bronsystemen. Vraag; in welk formaat zijn de documenten in de FileNet systemen opgeslagen en zijn deze tekst leesbaar?	De in FileNet opgeslagen bestanden hebben het Binary Large Object (BLOB) formaat.
30	23-7-2021	Pagina 8	Paragraaf 2.6	2.6 Datahygiëne De aanvullende metadata gegevens worden doorgaans in de analyse omgeving opgeslagen. Vraag; wenst opdrachtgever één uniforme database met alle mogelijke metadata gegevens voor de totale verzameling van 1,5 miljard of zijn er voor iedere informatie eigenaar verschillende structuren cq deelverzamelingen noodzakelijk?	Bij voorkeur alles in één uniforme database waarbij een uitsplitsing achteraf mogelijk is.
31	23-7-2021	Pagina 9	Paragraaf 2.8	2.8 Automation Vraag; wenst opdrachtgever naast de mogelijkheid van een zelf in te stellen periodieke scan ook de mogelijkheid van een continue monitoring van (delen van) de informatieverzameling op mutatie of aanvullingen en de analyse en verwerking daarvan?	Indien uw product dit biedt kan dit een waardevolle toevoeging zijn.
32	23-7-2021	Pagina 10	Paragraaf 2.10	2.10 Hosting. Vraag; als hosting plaatsvindt binnen de infrastructuur van opdrachtgever, voorziet opdrachtgever dan zelf in backup en recovery functionaliteit?	De opdrachtgever zal de back-up en recovery van het product zelf uitvoeren. Kunt u aangeven hoe uw product bij voorkeur geback-uppt dient te worden.
33	23-7-2021			Vraag; is het mogelijk om de lijst met vragen digitaal als Word tabel of als Excel sheet te ontvangen?	De beantwoording van alle door partijen gestelde vragen is in deze Nota van Inlichtingen opgenomen.

Nota van Inlichtingen

IUC21-027 Inzicht in Ongestructureerde Bestanden (IOB)

34	23-7-2021	Pagina 5	Paragraaf 2.1	<i>Is er een reden waarom het scannen beperkt blijft tot metagegevens? Zou scannen op inhoudsniveau ook voordelig zijn?</i>	<i>Scannen blijft niet beperkt tot metadata. Zie hiervoor de tekst in paragraaf 2.1, vierde bullet: De Oplossing moet kenmerken (patronen) kunnen scannen binnen bestanden "strings" en deze als metadata kunnen opslaan.</i>
35	23-7-2021	Pagina 5	Paragraaf 2.1	<i>Wij nemen aan dat u scannen op inhoudsniveau nodig hebt voor het matchen van patronen in de inhoud zelf?</i>	<i>Dit is correct.</i>
36	23-7-2021	Pagina 5	Paragraaf 2.1	<i>U stelt dat de oplossing ter plaatse moet draaien, maar in paragraaf 2.10 bespreekt u de hosting van de oplossing - bent u in staat om cloud-gebaseerde aanbiedingen te overwegen op basis van hun vermogen om verbeterde mogelijkheden te bieden die van waarde kunnen zijn voor de overheid?</i>	<i>Op dit moment staat de regelgeving geen cloudhosting toe. Dus op dit moment moet de oplossing zonder cloudhosting kunnen draaien. Hosting op de locaties van de Belastingdienst behoort wel tot de mogelijkheden.</i>
37	23-7-2021	Pagina 8	Paragraaf 2.6 Vraag 23	<i>Vraag 23 - Wat bedoelt u met "sjabloon" in dit scenario?</i>	<i>In vraag 23 vragen wij om templates en wij gaan er vanuit dat u dit vertaald heeft naar "sjabloon". De Belastingdienst is op zoek naar standaard zoekcriteria voor bijvoorbeeld AVG gerelateerde documenten. Een voorbeeld is het voorkomen van een Burger Service Nummer (BSN), maar dit kunnen ook andere criteria zijn. Zijn er binnen uw oplossing ook standaard templates aanwezig en zo ja beschrijf deze. Hiernaast vragen wij ons af of deze aan te passen of uit te breiden zijn met zoekcriteria en of het mogelijk is om nieuwe templates aan te maken en zo ja hoe.</i>