# Smartphone en computer als plaats delict

E-Discovery is een soort van CSI (Crime Scene Investigation), maar dan voor data. Het gaat erom digitale informatie te vinden, verzamelen en doorzoekbaar te maken, om die vervolgens te kunnen gebruiken als bewijsmateriaal. InformatieProfessional sprak met Hans Henseler, die ruim twintig jaar op dit vakgebied werkzaam is.

Ronald de Nijs en Eric Sieverts

## Hans Henseler

Na een studie Informatica aan de TU Delft (1982-1987) en een baan als onderzoeker bij de vakgroep Informatica aan de Rijksuniversiteit Limburg belandt Hans Henseler (1964) in 1992 bij het Gerechtelijk Laboratorium, de voorloper van het Nederlands Forensisch Instituut (NFI). 'Hier ben ik het forensisch onderzoek ingerold,' zegt Henseler, die bij zijn indiensttreding de laatste hand legt aan zijn promotieonderzoek over artificiële neurale netwerken en patroonherkenning. Als eerste forensisch computeronderzoeker zet hij hier de afdeling Forensisch Computer Onderzoek op.

Van 1998-2000 werkt hij als leider divisie informatiesystemen bij TNO-TPD (Technisch Physische Dienst) en vervolgens gaat hij als technisch directeur aan de slag bij Zylab, waar hij zich bezighoudt met softwareontwikkeling. 'In die tijd – net na de eeuwwisseling – groeide het zoeken in digitaal bewijsmateriaal vooral in de Verenigde Staten heel hard. In Nederland maar ook in de rest van Europa kwam dit (in de civiele sector) maar langzaam van de grond. Deze ontwikkeling werd door Zylab wel al gesignaleerd – ze waren betrokken bij het doorzoekbaar maken van alle stukken rond het proces tegen O.J. Simpson. Maar hier werden hun producten vooral nog toegepast voor kennismanagement en workflowautomatisering.'

E-Discovery komt weer op zijn pad als Henseler vanaf 2006 bij PwC grote projecten gaat leiden. Het betreft bijvoorbeeld bedrijven die door toezichthouder NMA verdacht worden van het maken van prijsafspraken. Ondertussen is hij in 2009 lector E-Discovery geworden aan de Hogeschool van Amsterdam (HvA). 'Bij recordsmanagement, waar mijn lectorcollega Geert-Jan van Bussel bij de HvA zich op richt, denk je tevoren al na hoe je eventueel voor bewijsdoeleinden je informatie wilt vastleggen. Met E-Discovery ga je met terugwerkende kracht na of er iets gebeurd is; of je het terug kunt vinden in de informatie. Dat is mooi complementair.'

Na zijn PwC-tijd begint Henseler voor zichzelf en sinds 2010 is hij partner bij Fox-IT, bij de oprichting in 1999 het eerste forensisch digitaal onderzoeksbureau in Europa. De afgelopen tien jaar is Fox-IT (inmiddels 190 werknemers) met name snel gegroeid op het terrein van cybersecurity. Henseler gaat hier nu het forensisch digitaal onderzoek weer op de kaart te zetten. Daarnaast heeft hij de leiding over de ontwikkeling van het forensische uitleesprogramma Tracks Inspector, waarmee tactisch rechercheurs zelf digitaal bewijsmateriaal kunnen analyseren.

'Een politiemedewerker vloog dan naar Japan om het wachtwoord van de elektronische agenda af te laten halen'

Het klinkt als een spannend jongensboek. In de jaren negentig probeerde de politie toegang te krijgen tot de inhoud van elektronische zakagenda's van opgepakte criminelen. Daarin zat veelal een complete administratie – en dus heel veel interessante informatie, maar wel beveiligd met een wachtwoord. Met een rechtshulpverzoek van Interpol op zak vloog een medewerker van de politie dan naar de fabrikanten van deze apparatuur, zoals Sharp en Casio in Japan, om het wachtwoord van de agenda af te laten halen. Vervolgens kon de politie gemakkelijk bij de gegevens, aangezien die niet waren versleuteld. Zakagenda's hadden toen nog geen rekenkracht.

Tegenwoordig is het vele malen moeilijker om versleutelde informatie op elektronische apparatuur te kunnen kraken. Om nog maar te zwijgen over de grote hoeveelheden informatie die computers tegenwoordig bevatten. Hoe haal je in geval van bijvoorbeeld boekhoudfraude uit alle digitale bestanden van een bedrijf de interessante en relevante informatie?

Welkom in de wereld van E-Discovery, het vakgebied van Hans Henseler dat erop gericht is om digitale informatie te vinden, verzamelen en doorzoekbaar te maken, om het vervolgens te kunnen gebruiken als bewijsmateriaal.

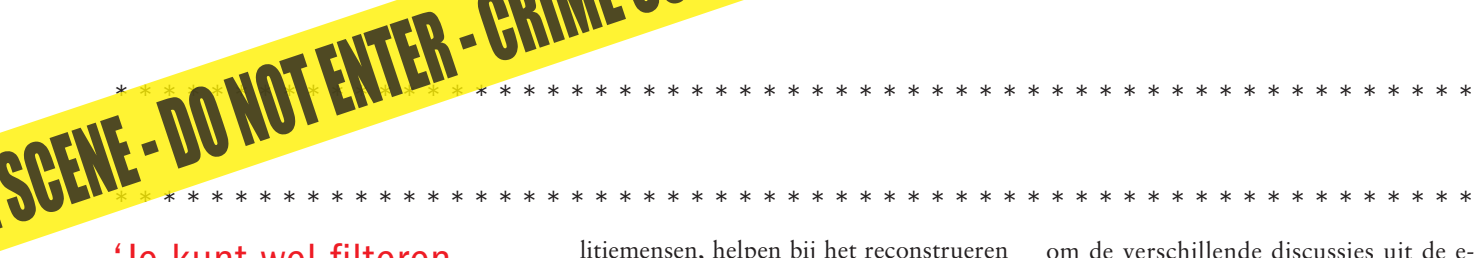
'E-mail niet altijd het ultieme, sluitende bewijs'

### Stappenmodel

'Discovery' is een juridische term uit de Verenigde Staten, legt Henseler uit. Voordat een zaak bij de rechter komt, kunnen partijen een zogeheten 'Discovery-verzoek' indienen. De tegenpartij moet dan vertrouwelijke bedrijfsdocumenten aanleveren, waarna de vragende partij, als voorbereiding op de zaak, deze informatie kan doorzoeken op bewijs. 'Omdat alle informatie tegenwoordig digitaal is, spreekt men van "E-Discovery".'

Er is een zogeheten Electronic Discovery Reference Model (EDRM). 'Dit stappenmodel omvat het identificeren, verzamelen, verwerken, analyseren en produceren van informatie. Maar eigenlijk begint het bij informatiemanagement: van tevoren dien je al na te denken over hoe je voor bewijsdoeleinden alle informatie in een bedrijf of organisatie wilt vastleggen.'

Vervolgens moet je identificeren welke informatie belangrijk is voor het onderzoek. 'Als iemand bijvoorbeeld verdacht wordt van omkoping of overtreding van de mededingingsregels, dan moet je je misschien alleen beperken tot de e-mail van de verkoopafdeling. Deze identificatie is een belangrijke fase.'




‘Je kunt wel filteren, maar je houd uiteindelijk toch een grote hoeveelheid mailtjes over die stuk voor stuk gelezen moeten worden’

Heb je eenmaal de informatie geïndentificeerd, dan moet je deze gaan verzamelen. ‘Dat moet op een forensisch verantwoorde wijze gebeuren. Want als je later conclusies gaat trekken, moet je wel de chain of custody kunnen aantonen. Dus: Waar heb je de informatie vandaan? Komt de informatie ook echt van de plek waarvan jij zegt dat-ie vandaan komt en is die in de loop van het onderzoek niet gewijzigd?’
Na verzameling volgt de verwerking, zoals attachments uit e-mails halen, indexeren en de mails misschien ook ontdubbelen. Daarna volgt de analyse en de review van die informatie.
De review en analyse resulteren in de selectie van de belangrijke informatie die je als bewijs wilt gebruiken. Soms kan een belangrijk resultaat zijn dat er niets gevonden is. In dat laatste geval zul je heel goed moeten uitleggen dat je volledig bent geweest. Als er wel relevante informatie is gevonden, dan moet die geproduceerd worden in een van tevoren afgesproken formaat zodat het aan de andere partij (bijvoorbeeld een beurswaakhond of mededingingsautoriteit) kan worden opgeleverd.
‘Een formele productie is heel belangrijk voor juristen maar in de praktijk, en zeker in Nederland, volstaat meestal een minder formele rapportage die door onderzoekers als bijlage bij hun rapport wordt gebruikt. In sommige gevallen, bijvoorbeeld in een rechtszaak, moet het bewijs tenslotte gepresenteerd worden.’

**Wanneer wordt zoiets nu geaccepteerd als bewijsmateriaal?**
*Hans Henseler:* ‘Het gaat om betrouwbaarheid natuurlijk, al is e-mail niet altijd het ultieme, sluitende bewijs. Wel kan de inhoud van de mailberichten onderzoekers, zoals forensisch accountants of politiemensen, helpen bij het reconstrueren van wat er is gebeurd. Neem bijvoorbeeld boekhoudfraude. Als je op de computer van de boekhouder bewijzen hebt gevonden, kun je in een interview met de verdachte zeggen: “Ik weet hoe het zit: je hebt dit, dit en dat gedaan.” In negentig procent van de gevallen bekent de verdachte dan.’

**Hebben jullie veel te maken met encryptie en versleuteling van informatie?**
‘Encryptie bestaat al heel lang. Als je het goed doet, zijn informatiebestanden niet te kraken. Maar de daarvoor gebruikte sleutels moet de gebruiker wel zelf invoeren én onthouden. Als een verdachte ooit in een e-mail of een chat of waar dan ook die sleutel heeft opgeschreven, dan is de kans groot dat een analyseprogramma na een weekendje draaien zo’n sleutel heeft kunnen herkennen.’

**Zijn ‘gewone’ retrievaltechnieken ook bruikbaar?**
‘Voor het doorzoeken van heel grote e-mailbestanden gebruiken we Clearwell, een concurrent van Zylab. Clearwell werkt volgens het trechterprincipe. Het is te vergelijken met de werkwijze van AutoScout, een online markt waar je 2 miljoen tweedehandsauto’s vindt. Hierin zoek je niet met steekwoorden. Wie een wagen via AutoScout zoekt, doet dat door steeds categorieën uit te sluiten: ik wil geen rode, blauwe of witte auto, geen op benzine of meer dan een jaar oud. Door steeds categorieën uit sluiten, ga je trechteren. Maar daarvoor moet wel enige semantische structuur in die gegevens aanwezig zijn.’
‘Bij E-Discovery zijn de filters bijvoorbeeld: de taal van e-mailberichten en de e-maildomeinen van de afzenders. Dergelijke tools zijn ook goed bruikbaar om de verschillende discussies uit de e-mailthreads bij elkaar te zetten, want als je dat alleen op basis van het subject van de e-mail doet, is dat niet volledig.’
‘Overigens kun je wel filteren, maar op een gegeven moment houd je toch een grote hoeveelheid mailtjes over die stuk voor stuk gelezen moeten worden. Bij patentzaken in de VS waar veel haast mee is, gebeurt het soms dat 100 mensen tegelijkertijd in plukjes 1.000.000 mails reviewen en daar een maand mee bezig zijn.’

Als een verdachte ooit in e-mail of chat de encryptiesleutel heeft opgeschreven, is de kans groot dat een analyseprogramma zo’n sleutel kan herkennen’

**Hoe bepaal je de relevantie in die gevonden (mail)bestanden?**
‘Google gebruikt voor zijn zoekresultaten relevance ranking die is gebaseerd op de populariteit van de pagina’s. Als ik digitaal bewijsmateriaal in beslag neem, dan zit daar geen populariteitsmaat aan. Dus relevance ranking in forensisch onderzoek moet op een andere manier gebeuren.’
‘Allerlei kennismanagementtechnieken waar bedrijven jaren terug vanwege de hoge kostprijs niet in wilden investeren, zie je nu terug in E-Discovery-tools. Bijvoorbeeld vector space search. Daarbij worden documenten voorgesteld als vectoren in een hoogdimensionale ruimte, die wordt opgespannen door de woorden die in de hele documentverzameling voorkomen. Hoe meer de vector van een bepaald document in dezelfde richting wijst als die van een voorbeelddocument, hoe meer die overeenkomen.’
‘Nu wil men wel voor die dure oplossingen betalen, omdat de financiële belangen bij E-Discovery heel groot zijn. Als 100 advocaten in plaats van vier maar drie weken hoeven te reviewen, dan scheelt dat veel geld. Bovendien zijn computers veel sneller geworden en ze hebben veel meer geheugen, zodat de benodigde vectorberekeningen en andere slimme algoritmes op het terrein van taaltechnologie voldoende snel kunnen worden uitgevoerd. Clearwell biedt zowel Boolean search als vector space search, waarmee ze aan predictive coding kunnen doen.’

**Kun je iets meer vertellen over predictive coding?**
‘Bij predictive coding probeer je de relevantie van een document te voorspellen, vooral op basis van taalkundige of vectormodelachtige vergelijkingen met andere documenten. Een aardig voorbeeld van de noodzaak van predictive coding zie je in


‘De politie wil er de komende twee jaar nog eens 450 digitale forensisch experts 450 bij hebben’

de VS. Elke keer als een nieuwe president aantreedt, gaan alle mails van het Witte Huis uit de ambtstermijn van de vorige president naar de National Archives. Zo kwamen er in 2001 32 miljoen mails binnen van de Clinton Administration. De verwachting is dat in 2017 de Obama Administration maar liefst 1 miljard e-mails zal inleveren.’

‘Destijds wilde de tabaksindustrie alle mailtjes van het Witte Huis uit die periode die op hen betrekking hadden. Je kunt dan niet alle mails uit die periode geven en zeggen: zoek het maar uit. Want dan zien ze ook alle andere communicatie van het Witte Huis. De National Archives hadden dus het probleem hoe een onderbouwde selectie te maken die ze “verdedigbaar” konden overdragen aan advocaten in de tabaksindustrie. Tot begin vorig jaar gebeurde zoiets met Booleaans zoeken. Aan het samenstellen van de uiteindelijke zoekvraag ging een hele onderhandeling vooraf, tussen de overheid en de advocaten van een bedrijf of organisatie. Dat leverde queries op van soms drie A4tjes lang, met eindeloze combinaties van zoekwoorden met ANDs en ORs.’

‘Het maken van zo’n query is heel moeilijk en je weet bij voorbaat al dat je nooit alles zult vinden. Als wetenschapper op het gebied van machine learning zeg je: daar hebben we toch andere tools voor, zoals supervised learning. Hier is een stapel voorbeelden van documenten die ik wel interessant vind, en dit zijn voorbeelden van wat ik niet interessant vind, ga maar rekenen met die vectortechnieken en geef me alles wat met de voorbeelden overeenkomt. De computer leert dus onder supervisie wat in een bepaald geval interessant is.’

‘Het probleem voor advocaten is dat dit een black box oplevert: het is een wiskundig model waarbij de computer over relevantie beslist. Daar kun je wel een drempel bij instellen, dat je bijvoorbeeld de beste dertig procent wilt hebben, maar rechters en advocaten vertrouwen dat niet. Ze willen transparantie en zeggen: we willen gewoon Booleaans zoeken met steekwoorden; daarvan begrijpen we precies wat er wel en niet uitkomt. Maar dat brengt enorme kosten met zich mee en blijkt inmiddels niet goed genoeg te werken. Vorig jaar is er in de VS voor het eerst een uitspraak geweest van een rechter die het goed vond predictive coding toe te passen.’

‘Allerlei kennismanagementtechnieken waar bedrijven vanwege de hoge kostprijs niet in wilden investeren, komen nu terug in E-Discovery-tools’

**Het gaat dus om acceptatie van statistische kansen als iets waarmee je kunt werken? En zo’n gigantische Booleaanse combinatie is toch ook maar schijnzekerheid; daar kan ook een term vergeten worden.**

‘Precies. De rechter zei: ik weet dat het niet perfect is, maar het gaat erom dat het proces dat je gevolgd hebt duidelijk is. En je moet ook een steekproef nemen in de bestanden die je niet gelezen hebt. Zodat je enigszins betrouwbaar weet dat daar echt geen interessante mailtjes tussen zitten. Dit raakt nu langzaam aan geaccepteerd.’

**Met predictive coding doe je in feite aan textmining. Zijn er nog andere technieken op het gebied van text en datamining die voor jullie belangrijk zijn?**

‘Met predictive coding zit je aan de kant van “supervised” learning. Ik denk dat je met unsupervised ook wel wat kunt doen. Dat wil zeggen dat je de computer laat rekenen, en kijkt wat voor categorieën die zelfstandig vindt; wat springt eruit? Het idee daarbij is dat je vaak tevoren niet weet wat je zoekt; dat is dus een krachtige techniek. Je kunt predictive coding wel zien als textmining, maar je bent toch niet echt aan het minen. Je maakt overal

## ‘Predictive coding is dé oplossing voor het privacyprobleem bij E-Discovery’

vectoren van en je probeert die slim te ordenen.’
‘Een andere vorm van textmining is “entity extraction”. Je vindt een bepaald document pas echt interessant, als je weet dat deze personen en plaatsen erin voorkomen en misschien wel een aantal relaties daartussen. Ik geloof er heel erg in om zo van ongestructureerde informatie naar gestructureerde te komen.’
‘Op het moment dat je die structuur hebt, kun je filterlijstjes gaan maken, en kun je net als bij AutoScout de gebruiker gaan prikkelen: ik heb 100.000 documenten maar deze onderwerpen komen erin voor, wat vind je interessant, “klik”, dat onderwerp is niet interessant, “klik”. Dat kan alleen als je die informatie hebt, maar met de hand is dat niet te doen, dus moet de computer dat automatisch doen.’

**Hoe zit het met privacyachtige problemen? Of heeft wat bij bedrijven gebeurt, niet zozeer met persoonlijke privacy te maken?**

‘Juist wel, omdat dergelijke E-Discovery-onderzoeken vaak vanuit de Verenigde Staten worden aangestuurd. Nu begrijpen ze het wel, maar een paar jaar geleden was er totaal onbegrip dat er een juridische basis moet zijn om bijvoorbeeld bij een Duits bedrijf e-mailberichten in beslag te mogen nemen. Het verschilt ook per land. Ze werken dus ook veel via advocatenkantoren, dat is ook een van de veranderingen die ik hier in gang heb gezet. De advocaten zorgen dan voor de juridische inbedding.’
‘We hebben het ook wel over E-Discovery readiness: wat moet je doen om als bedrijf op E-Discovery voorbereid te zijn. Staat een Nederlands bedrijf werknemers toe om de bedrijfscomputer voor privédoeleinden te gebruiken, dan mag je bij een E-Discovery-onderzoek niet zomaar alle informatie bekijken. Maar is er een zwaarwegend belang of heb je de toestemming van de medewerker, dan mag het weer wel.’

**Kun je een techniek als predictive coding gebruiken om privégevens uit informatiebestanden te filteren?**

‘Predictive coding is dé oplossing voor het privacyprobleem. Een paar jaar geleden kwam de discussie over de bodyscan op vliegvelden op gang. Nadat een man met een bom in zijn schoenzool in een vliegtuig was ingestapt, wilde men op Schiphol meer van die apparaten plaatsen. Daar waren bezwaren tegen, aangezien je op een bodyscan niet alleen objecten ziet die een reiziger bij zich heeft, maar ook diens lichaam.’
‘Leveranciers van dergelijke apparatuur hebben toen een update gedaan, waardoor men geen “naaktfoto” meer te zien krijgt. maar een cartoonachtig poppetje met eventuele verdachte objecten. Dit zie ik bij predictive coding ook als optie: het wegfilteren van privé-informatie. De computer krijgt deze informatie wel te zien, de gebruiker niet.’

**Je werkt deels bij de HvA en deels bij Fox-IT. Kun je studenten voor ‘geheime’ projecten inzetten?**

‘Ik probeer altijd heel open en transparant te zijn. Een van de businessunits van Fox-IT werkt voor banken. Als er incidenten zijn, moet dat geheim blijven. Dat geldt ook voor onze werkzaamheden voor ambassades, defensie en dergelijke. En ook als je forensisch onderzoek voor een klant doet, moet dat vanzelfsprekend geheim blijven, maar de methode waarmee je het doet, moet juist transparant zijn. Want je moet de toets der kritiek kunnen doorstaan. Dus stagiaires mogen van mij de onderzoeksmethoden die ze bestudeerd hebben of soms hebben ontwikkeld in hun scriptie beschrijven en eventueel publiceren, zolang er maar geen klantgegevens in staan.’
‘Ik ben heel erg voor het delen van kennis. Natuurlijk leest de concurrentie mee, maar vanuit mijn jarenlange ervaring met softwareontwikkeling weet ik dat de stap van een idee naar een werkend en te verkopen product wel eventjes duurt...’

**In hoeverre staat e-discovery in ons land nu al op de kaart?**

‘Er is de laatste jaren meer aandacht voor gekomen, al denk ik dat Nederlandse bedrijven er veel meer aan moeten doen. Er wordt nu vooral naar gekeken in termen van riskmanagement, maar er zijn ontwikkelingen die maken dat bedrijven er niet onderuit komen. Bijvoorbeeld door een meldplicht datalekken die eraan komt. Die dwingt bedrijven beter te weten wat voor informatie in de organisatie aanwezig is. Systeembeheerders beheren systemen maar in de praktijk is het heel lastig om te achterhalen welke applicaties op welke systemen draaien en waar ze voor gebruikt worden en dus welke informatie er in staat opgeslagen. Die kennis is in het voordeel van diegenen die zich met E-Discovery bezighouden. Want nu zijn we veelal dagen bezig om alle informatie in een bedrijf in kaart te brengen.’
‘Vervolgens is de vraag of je als bedrijf zelf mensen moet opleiden die die informatie kunnen aanleveren? Dat hangt ervan af in welke sector je zit en hoe groot je bent. Soms kan je beter afspraken maken met een leverancier. Maar het is wel goed ook bij je eigen mensen bewustzijn te creëren.’

**En de echte specialisten, zoals jij? Zijn er daar ook nog meer van nodig? Is het een groeimarkt?**

‘De politie heeft inmiddels 300 digitale forensisch experts, maar wil er nog eens 450 bij in de komende twee jaar. Dat gaat, denk ik, niet lukken. Zorg er eerst eens voor dat de “gewone” tactisch rechercheurs tools en een betere infrastructuur krijgen waarmee ze meer zelf kunnen doen, zoals het bekijken van foto’s en berichten in smartphones en computers. De huidige 300 experts kunnen zich dan richten op de echt ingewikkelde zaken.’

*Ronald de Nijs is eindredacteur en Eric Sieverts is redacteur van InformatieProfessional.*

**Hans Henseler over E-Discovery horen spreken?**

Op 27 juni spreekt Hans Henseler tijdens het ochtenprogramma van het KBenP Event in Scheveningen. Hij zal uitgebreid ingaan op de do’s en don’ts van E-Discovery. Meer informatie is te vinden op www.kbenp.nl/actueel/kbenp-zomerevent-2013.