Onderzoek naar e-Discovery voor informatiebeheer

Hans Henseler en Charles Jeurgens ■

De Hogeschool van Amsterdam, Universiteit van Amsterdam, Universiteit Leiden, Nationaal Archief, Gemeente Amsterdam en een aantal departementen slaan de handen ineen om te onderzoeken welke bijdrage e-Discovery kan leveren aan het verder moderniseren van informatiebeheer bij overheidsorganen. Het consortium wil de komende twee jaar onderzoeken hoe de werelden van e-Discovery en records management elkaar kunnen versterken en op welke wijze e-Discoverytechnieken ingezet kunnen worden om het records management te ondersteunen.

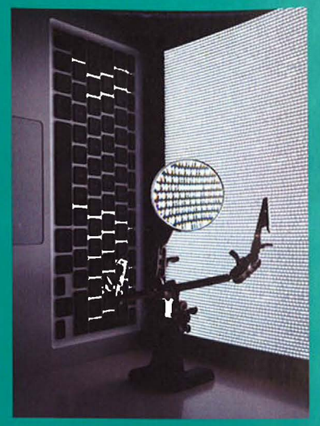
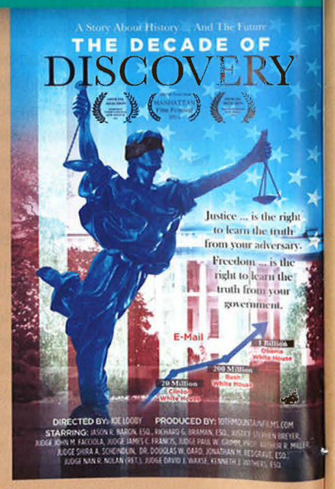


Foto: Sjoerd Knibbeler.

Jason Baron, lange tijd directeur van de afdeling *litigation* van de *National Archives and Records Administration* in de Verenigde Staten, werd vaak geconfronteerd met rechtszaken waar de federale overheid bij betrokken was en waarin grote hoeveelheden elektronische documenten op relevantie moesten worden beoordeeld. In de Verenigde Staten hebben partijen het recht om tijdens het vooronderzoek documenten en ander bewijsmateriaal van derden op te vragen en te gebruiken. In een rechtszaak aan het eind van de jaren negentig – de Amerikaanse staat tegen Philip Morris en een aantal andere tabaksfabrikanten wegens misleiding en achterhouden van informatie met betrekking tot gezondheidsrisico's van roken – moesten 32 miljoen e-mails van de Clinton-regering worden doorzocht op informatie die betrekking had op de tabaksindustrie. Met slimme zoektechnieken slaagden Jason en zijn team erin om het aantal e-mails dat relevante informatie zou kunnen bevatten, terug te brengen tot zo'n 200.000. Vervolgens duurde het overigens nog een half jaar om met een groep van 25 advocaten en archivarissen deze set handmatig te doorzoeken op informatie die werkelijk van belang was voor de rechtszaak. Het is een voorbeeld van e-Discovery: identificeren, verzamelen, verwerken, doorzoeken en analyseren van grote hoeveelheden veelal ongestructureerde elektronische informatie. Jason Baron is inmiddels een autoriteit op het gebied van dit type onderzoek en was onlangs in Nederland te gast bij een symposium over e-Discovery.



De documentaire 'The Decade of Discovery', met daarin Jason Baron, schetst hoe e-Discovery zich heeft ontwikkeld in de periode van ca. 2002–2012.

Records management

Een van de kenmerken van de digitale informatiesamenleving is de enorme snelheid waarmee de hoeveelheid informatie toeneemt. Er is berekend dat op dit moment in twee dagen meer informatie wordt gegenereerd dan er sinds de uitvinding van het alfabet tot 2003 is geproduceerd en vastgelegd. Negentig procent van alle bestaande informatie in de wereld is pas in de afgelopen twee jaar gevormd.[1] Die informatievloed proberen informatieprofessionals zodanig te managen dat de gegevens kunnen worden gebruikt als ze nodig zijn. Informatieprofessionals worstelen echter met de overvloed aan informatie die op zichzelf allemaal relevant kan zijn voor verantwoording en bewijsvoering.
Naar de letter van de Archiefwet zijn overheidsorganisaties verplicht om de informatie die uit hun werkprocessen voortvloeit in 'goede, geordende en toegankelijke staat' te hebben.

Discussiepanel met Jason Baron (in het midden) op het jaarlijkse E-Discovery Symposium, op 23 april 2015 georganiseerd door het lectoraat E-Discovery van de HvA. Het thema was 'E-Discovery en Information Governance'. (foto: Carina Klijn).

Veel organisaties slagen er niet in om hun informatie op een manier te beheren die 'in control' is. De budgetten zijn onvoldoende en het beschikbare instrumentarium in de vorm van standaarden, normen en metadata-schema's wordt als complex en lang niet altijd als realistisch ervaren. Soms kan dit een organisatie opbreken, bijvoorbeeld in een juridische kwestie of wanneer ten behoeve van een parlementair onderzoek plots alles relevant kan zijn. Hoe dan snel de benodigde relevante informatie te vinden uit de vele verschillende systemen die gebruikt worden? De gebrekkige kwaliteit van de gegevens die in dergelijke situaties aangeleverd worden, is een terugkerende constatering van tal van advies- en onderzoekscommissies. Onlangs stelden zowel de onderzoekscommissie Elias, inzake falende ICT-projecten bij de overheid, als de Parlementaire Enquêtecommissie Woningcorporaties dat hun onderzoek ernstig was gehinderd door gebrekkige archivering en het ontbreken van relevante informatie. Als oorzaak voor de ondermaatse kwaliteit van informatiebeheer wordt vaak gesteld dat digitaal informatiebeheer stoelt op technieken die in een papieren omgeving misschien voldeden, maar in een digitale omgeving achterhaald zijn. Dit moet en kan beter. Maar hoe? Kan e-Discovery het records management ondersteunen?

Automatisch classificeren

Het consortium wil onderzoeken op welke manier ongestructureerde informatie automatisch geclassificeerd kan worden door middel van technieken die binnen e-Discovery ontwikkeld zijn en worden. Met dergelijke technieken blijkt geregeld dat onvindbaar geachte informatie *achteraf* toch te vinden is. De vraag is of deze technieken ook vooraf ingezet kunnen worden om records management te vereenvoudigen en doeltreffender te maken.
Praktische vragen van informatieprofessionals lijken veel op vragen die in e-Discovery spelen (zie kader 'e-Discovery: vragen uit de praktijk'). Zo is er een grote behoefte aan technieken die de registrerende en archiverende ambtenaar kunnen helpen om zijn informatie zodanig te registreren dat deze later ook eenvoudig terug te vinden is. De afgelopen jaren is *predictive coding* in e-Discovery bezig aan een opmars. Door middel van *machine learning* kan de computer op basis van een set voorbeelden en gedefinieerde criteria leren om gelijksoortige documenten te herkennen. Die technieken bestonden al langer maar vanwege de groeiende omvang van informatie is een goed alternatief niet langer aanwezig en de kwaliteit is even goed of zelfs beter dan handmatige classificatie.

Groot potentieel

Predictive coding heeft een groot potentieel voor *information governance*, bijvoorbeeld bij het automatisch bepalen of een e-mail of document wel of niet van belang is om te archiveren en of een document wel of geen persoonsgerelateerde gegevens bevat. Bij een WOB-verzoek kan dit laatste helpen om sneller te identificeren welke documenten handmatig gescreend moeten worden om persoonsgegevens weg te lakken. Ook het weglakken is een bekend onderdeel van e-Discovery en staat bekend als *redaction*. Met behulp van een e-Discovery review-platform kunnen tientallen en zelfs honderden gebruikers gelijktijdig documenten reviewen en informatie weglakken. Daarnaast bevat zo'n platform uitgebreide functionaliteiten om documenten en bijbehorende metadata in een afgesproken formaat automatisch te exporteren (in e-Discovery beter bekend als *production*).

Wangedrag detecteren

Op de zesde workshop voor Discovery of Electronically Stored Information (DESI, onderdeel van de International Conference on Artificial Intelligence and Law conferentie) die afgelopen juni in San Diego gehouden werd, presenteerden onderzoekers van een advocatenkantoor een toepassing van *predictive analytics* om vroegtijdig wangedrag in de organisatie te kunnen detecteren. De onderzoekers hebben een groot aantal termen verzameld die verband houden met diverse vormen van fraude (boekhouding, intimidatie, omkoping et cetera) en negatieve sentimenten. Ze ontleenden hun voorbeelden uit afgeronde onderzoeken. Per onderzoek selecteerden ze willekeurig zeventig procent van de relevante e-mails. Deze selectie werd gebruikt om een model te trainen. Het model is vervolgens gevalideerd aan de hand van de resterende dertig procent relevante e-mails. De uitkomsten waren verrassend goed.

Cultuurverandering

De onderzoekers hadden voor deze vorm van predictive analytics de toepasselijke naam 'Apocalyptics' bedacht. Als nadeel werd genoemd dat het model een ruzie tussen medeweker en leidinggevende niet kan onderscheiden van een ruzie tussen medewerker en levenspartner. Daarmee raken de onderzoekers aan een ander gevoelig punt in de discussie, namelijk privacy. Advocaten en rechters zijn inmiddels overtuigd >>

Openbaar verhoor Parlementaire Enquêtecommissie Woningcorporaties. (foto: Enquêtecommissie Woningcorporaties van Tweede Kamer der Staten-Generaal, Wikimedia Commons).

>> van nut en noodzaak van predictive coding in onderzoeken waarvoor een duidelijke aanleiding is. Dat betekent echter niet dat medewerkers direct enthousiast zullen zijn over de toepassing van predictive coding om hun dagelijkse informatiestroom constant te monitoren, zonder aanleiding maar in het algemeen belang van de organisatie. Hoe dan ook zal er naast een technische oplossing ook op het gebied van cultuur en gedrag het nodige moeten veranderen om de problemen waarmee de informatieprofessional te kampen heeft op te kunnen lossen. ■

Noot

1 ■ Robert F. Smallwood, *Information Governance. Concepts, Strategies and Best Practices* (New Jersey, 2014), 3.

Charles Jeurgens is hoogleraar archival studies aan de Universiteit Leiden en werkzaam bij het Nationaal Archief. Hans Henseler is lector e-Discovery aan de Hogeschool van Amsterdam en algemeen directeur van Tracks Inspector. Samen doen zij onderzoek naar toepassingen van e-Discovery om informatieprofessionals bij overheidsorganisaties te ondersteunen.

Oorzaken

De problemen rond informatiemanagement zijn niet van vandaag of gisteren. Overheden proberen de kwaliteit van hun informatiebeheer op verschillende manieren te borgen. De traditionele technieken voor records management zijn er vooral op gericht om informatie te structureren zodat deze vindbaar is. Sinds eind negentiende en begin twintigste eeuw gebeurt dat vooral door documenten die op een en dezelfde zaak betrekking hebben bij elkaar te voegen en deze dossiers volgens een classificatiesysteem (UDC) te ordenen. Theoretisch beschouwd kan op die manier iedere snipper of iedere byte worden gearchiveerd en eenvoudig teruggevonden. Toch werkt dat, zo blijkt uit de steeds opnieuw gesignaleerde problemen, onvoldoende. Wat zijn hiervan de oorzaken?

Onvoorstelbaar groot

De omvang van de informatie die geproduceerd en gebruikt wordt is onvoorstelbaar groot geworden. Digitalisering heeft echter niet alleen tot groeiende informatiestromen geleid, maar ook tot geheel andere communicatiepatronen. Vroeger werd de post centraal ingeschreven en op die manier bestond een redelijk beeld van de informatie die door een instelling werd ontvangen en verstuurd. Ook was het betrekkelijk eenvoudig om de hiërarchische lijnen in de routering en afdoening tot uitdrukking te laten komen. Tegenwoordig communiceren medewerkers rechtstreeks via de mail of zelfs sociale media met elkaar en wordt helemaal niet meer geregistreerd welke informatie er in- en uitgaat.

Beperkt

Als gevolg hiervan is iedere medewerker steeds meer zijn of haar eigen records manager geworden. Bovendien wordt binnen organisaties in toenemende mate sterk bezuinigd op DIV-functies. Er zijn weliswaar allerlei DMS-systemen beschikbaar, maar soms lijkt het erop dat vergeten wordt dat medewerkers die geen achtergrond hebben op het gebied van records management met die systemen moeten kunnen werken. Gebruiksgemak en volledige integratie met de werkprocessen die worden uitgevoerd, zijn soms ver te zoeken. Informatie uit databases, websites en vaak ook e-mailsystemen zijn slecht vertegenwoordigd in deze documentmanagementsystemen. Dat betekent dat de informatie waarover een organisatie werkelijk 'in control' is vaak erg beperkt is.

E-Discovery: vragen uit de praktijk

Informatieprofessionals in overheidsorganisaties zitten met een aantal praktische vragen die ook in projecten rondom e-Discovery stelselmatig aan de orde zijn. Tijdens een informatiebijeenkomst die de consortiumpartners in de eerste helft van dit jaar organiseerden om met informatieprofessionals, werkzaam bij gemeenten, ministeries en toezichthouders, hun onderzoeksbehoeften te identificeren, kwamen de volgende vier vragen steeds terug:

1. *Op welke manier kan het informatielandschap van de organisatie efficiënt en zo volledig mogelijk in kaart worden gebracht, en hoe kan deze kaart worden onderhouden?* Bedenk daarbij dat de kaart ook oude informatie moet benoemen. Sommige informatie in oude systemen wordt niet gemigreerd naar nieuwe systemen maar blijft beschikbaar in oude systemen. Ook belangrijk is om niet alleen een overzicht te hebben van medewerkers op dit moment in de organisatie, maar een overzicht van medewerkers in de organisatie in het verleden.
2. *Is het mogelijk om e-mails en documenten automatisch te classificeren om te bepalen of ze wel of niet bewaard dienen te worden?* Medewerkers produceren grote hoeveelheden informatie en het is voor informatieprofessionals in de organisatie ondoenlijk om nog te bepalen welke informatie gewist mag worden en welke niet. Ook de medewerkers weten dit vaak niet en willen niet lastiggevallen worden met dit soort vragen. Met automatische classificatietechnieken is het wellicht mogelijk om automatisch te herkennen in welke categorie informatie valt.
3. *Op welke manier kan er binnen alle informatiesystemen van een organisatie gezocht worden naar informatie over een specifiek onderwerp?* Denk bijvoorbeeld aan een WOB-verzoek, Kamervragen et cetera.
4. *Indien de onder 3 gevonden informatie aan een externe partij geleverd moet worden, hoe kan de informatie die niet openbaar is eenvoudig geïdentificeerd en weggelakt worden (bijvoorbeeld persoonsgerelateerde informatie)?* Hierbij wordt enerzijds gedacht aan de toepassing van text mining-technieken om bijvoorbeeld automatisch namen van (vooraf onbekende) personen te identificeren, maar ook aan een efficiënte workflow en daarbij horende tools die in e-Discoveryprojecten worden gebruikt om bewijsmateriaal te produceren.