Informatieprofessionals worstelen met een overvloed aan informatie die relevant kan zijn. Waar is wat te vinden? Traditionele technieken schieten te kort. Kunnen e-Discovery-technieken het recordmanagement vereenvoudigen en doeltreffender maken? De uitkomsten zijn verrassend goed, concluderen Charles Jeurgens en Hans Henseler.

door: CHARLES JEURGENS & HANS HENSELER **beeld:** SHUTTERSTOCK

RECORDMANAGEMENT **HEEFT BAAT** BIJ E-DISCOVERY

Hulp voor informatieprofessionals die verdrinken in digitalisering

Organisaties schieten te kort bij het archiveren van informatie: recordmanagement. Dat is een van de bijvangsten van zowel de commissie Elias inzake falende ICT-projecten bij de overheid als de Parlementaire Enquêtecommissie Woningcorporaties. In beide gevallen werd het onderzoek ernstig gehinderd door gebrekkige archivering en het ontbreken van relevante informatie. Dit moet en kan beter. Maar hoe? De problematiek die Elias en de enquêtecommissie zijdelings benoemen, staat niet op zichzelf. Informatieprofessionals worstelen met een overvloed aan informatie die relevant kan zijn voor verantwoording en bewijsvoering.
Toegankelijkheid en hergebruik van informatie zijn actuele thema's in de dagelijkse beroepspraktijk van informatieprofessionals en managers. Overheidsdiensten worden voortdurend gevraagd om openheid van zaken te geven met betrekking tot hun bedrijfsinformatie en werkwijze. Denk bijvoorbeeld aan governance, WOB-verzoeken, parlementaire onderzoeken, Kamervragen et cetera. Ten gevolge van de digitalisering van de informatiemaatschappij schieten traditionele technieken te kort (zie kader). Met e-Discovery-technieken blijkt regelmatig dat onvindbaar geachte informatie achteraf toch te vinden is. Kunnen deze technieken wellicht ook vooraf ingezet worden om het recordmanagement te vereenvoudigen en doeltreffender te maken?

Weglakken

Praktische vragen van informatieprofessionals lijken veel op vragen die bij e-Discovery spelen (zie kader). Zo is er een grote behoefte aan technieken die de registrerende en archiverende ambtenaar kunnen helpen om zijn informatie te registreren zodat 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-technieken worden aan de hand van voorbeelden automatisch relevante documenten gevonden. Die technieken bestonden al langer maar bijzonder is dat advocaten en rechters in de VS deze technologie volledig hebben omarmd omdat er geen alternatief meer is en de kwaliteit even goed of zelfs beter en vooral ook goedkoper is dan handmatige classificatie. Predictive coding heeft een groot potentieel voor information governance. Bijvoorbeeld 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 'redactie'. Met behulp van een e-Discovery-reviewplatform kunnen tientallen en zelfs honderden gebruikers gelijktijdig documenten reviewen en informatie weglakken. Daarnaast bevat zo'n platform uitgebreide functionaliteit om documenten en bijbehorende metadata in een afgesproken formaat automatisch te exporteren.

E-Discovery-vragen

Informatieprofessionals in overheidsorganisaties zitten met een aantal praktische vragen die ook bij e-Discovery-projecten stelselmatig aan de orde zijn. Tijdens een informatiebijeenkomst voor informatieprofessionals bij het Nationaal Archief met aansluitend een discussie, bleken vooral de volgende vier vragen telkens terug te keren in verschillende typen overheidsorganisaties (gemeenten, ministeries, inspecties en toezichthouders):

1. Op welke manier kan het informatielandschap van de organisatie efficiënt en volledig in kaart worden gebracht en hoe kan deze kaart onderhouden worden? Bedenk daarbij dat de kaart ook oude informatie moet benoemen. Sommige informatie in oude systemen wordt niet gemigreerd naar nieuwe systemen, zij blijft beschikbaar in oude systemen. Het is belangrijk om niet alleen een overzicht te hebben van medewerkers op dit moment in de organisatie maar ook van medewerkers uit 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 lastig gevallen 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 informatiebronnen van de organisatie (die voorkomen in het informatielandschap, zie punt 1) gezocht worden naar informatie over een bepaald onderwerp (denk aan een WOB-verzoek, kamervragen et cetera).
4. Indien de onder 3 gevonden informatie aan een externe partij geleverd moet worden, hoe kan dan informatie efficiënt weggelakt worden (bijvoorbeeld persoonsgerelateerde informatie)? Hierbij wordt enerzijds gedacht aan de toepassing van text-miningtechnieken om bijvoorbeeld automatisch namen van (vooraf onbekende) personen te identificeren maar ook aan efficiënte workflow en daarbij behorende tools die in e-Discovery-projecten worden gebruikt om bewijsmateriaal te produceren.

Oorzaken slecht informatiebeheer

De problemen rondom informatiemanagement zijn niet van vandaag of gisteren. Overheden proberen de kwaliteit van hun informatiebeheer op verschillende manieren te borgen. De traditionele recordmanagementtechnieken zijn er vooral op gericht om informatie te structureren zodat deze vindbaar is. Sinds eind 19e en begin 20e 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 het blijkens de steeds weer opnieuw gesignaleerde problemen allemaal onvoldoende.
Wat zijn de oorzaken hiervan?
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 in- en uitgaat.
Als gevolg hiervan is iedere medewerker steeds meer zijn eigen recordmanager geworden en in steeds meer organisaties is sterk bezuinigd op de DIV-functies. Er zijn weliswaar allerlei DMS-systemen beschikbaar, maar soms lijkt het erop dat vergeten is dat medewerkers die geen achtergrond hebben op het gebied van recordmanagement, 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. 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 en te houden. Deze verplichting geldt voor alle overheidsinformatie. Daar zit ook meteen een probleem. Een aantal organisaties is hier heel duidelijk in: het budget is gewoon te klein om dit ook daadwerkelijk waar te kunnen maken. De standaarden, normen en metadataschema's die beschikbaar zijn voor informatiebeheer worden als complex en lang niet altijd als realistisch ervaren. Het gevolg is dat of alle informatie op een onvoldoende niveau wordt beheerd, of dat differentiatie plaatsvindt waarbij belangrijke informatie beter beheerd wordt dan minder belangrijke informatie.
Soms kan dit een organisatie opbreken. In een juridische kwestie, of wanneer bijvoorbeeld 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 worden gebruikt?

Wangedrag

Op de 6de workshop voor Discovery of Electronic Stored Information (DESI workshop, 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 70 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 30 procent relevante e-mails. De uitkomsten waren verrassend goed.
De onderzoekers hadden voor deze vorm van predictive analytics de toepasselijke naam 'Apocalyptics' bedacht. Als nadeel werd genoemd dat het model een ruzie tussen medewerker en leidinggevende niet kan onderscheiden van een ruzie tussen medewerker en echtgeno(o)t(e). Daarmee raken de onderzoekers aan een ander gevoelig punt in de discussie, namelijk privacy. Advocaten en rechters zijn inmiddels overtuigd van het 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 (in het algemeen belang van de organisatie). Hoe dan ook, er zal naast een technische oplossing ook een cultuurverandering nodig zijn om de problemen van de informatieprofessional op te lossen. «

Charles Jeurgens (links) is hoogleraar archival studies aan de Universiteit Leiden en tevens 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 informa tieprofessionals bij overheidsorganisaties te ondersteunen.

ag
Voor reacties en nieuwe bijdragen van IT-experts: Henk Ester, 070 3046812 h.ester@automatiseringgids.nl