Scannen papieren archief is monsterklus

Metadata ontsluit digitaal papier

Het digitaliseren van papieren archieven bespaart ruimte en vereenvoudigt het zoeken naar documenten. Bedrijven moeten daarbij goed overwegen welke digitale standaard ze willen gebruiken. Consultant Bart van Wanroij waarschuwt dit soort projecten niet licht op te vatten omdat de meeste bedrijven niet zijn voorbereid op het scannen van 15.000 pagina's.

Het lijkt erop alsof papier moeilijk is uit te dammen. Er zijn nog maar weinig kantoren zonder archiefkasten en zelfs in de met de duurste Business Process Management (BPM)-software gemaakte workflow zit vrijwel altijd een velletje papier. Een drietal ontwikkelingen kan de overgang naar digitaal papier vergemakkelijken: de TabletPC's, het Portable Document Formaat (PDF) en het gebruik van zogenaamde layers. Het is de vraag of deze ontwikkelingen ervoor kunnen zorgen dat digitaal papier algemeen wordt geaccepteerd.

TabletPC

Microsoft introduceerde onlangs de TabletPC, waarmee documenten met een staande of liggende paginaoriëntatie kunnen worden gelezen terwijl de gebruikelijke computerschermen over het algemeen alleen een liggend scherm tonen. Deze eigenlijk zeer voor de hand liggende nieuwe functie kan mogelijk grote gevolgen hebben voor digitale archieven. Bedrijven kunnen hun facturen opbergen in virtuele multomappen die via de TabletPC draadloos beschikbaar zijn, eenvoudig kunnen worden doorgebladerd en direct duidelijk leesbaar worden getoond. Daarbij profiteren werknemers van de voordelen van het digitaal zoeken op trefwoorden. Het zou overigens naïef zijn om alleen de voordelen van het gebruik van de TabletPC te belichten, zonder aandacht te besteden aan het monnikenwerk dat voorafging aan het omzetten van stapels mappen naar digitaal papier. Digitaal papier is niets meer dan een computerbestand met dezelfde vorm als het papieren origineel. De paginagrootte van een eBook past zich automatisch aan de breedte van het scherm aan en de bijgeleverde software beschermt de auteursrechten van de schrijver. Op die manier kunnen uitgevers contracten aanbieden die recht doen aan de oorspronkelijke papieren vorm van de eBooks en zekerheid bieden omtrent hun authenticiteit. Bovendien kunnen de eBooks op tekstdelen worden doorzocht.

Digitalisering

Een digitaal archief kan niet van de ene op de andere dag worden gerealiseerd. Organisaties die hun kastruimte willen vervangen door schijfruimte moeten eerst de papierkraan dichtdraaien voordat ze gaan dweilen met scanners. Zo kunnen bedrijven binnenkomende faxen automatisch omzetten in e-mailberichten en inkoopfacturen scannen en vervolgens routeren zodat ze niet worden gekopieerd en opgeslagen. Ook kunnen organisaties intern geproduceerde documenten standaardiseren via briefmodellen, wat de doorzoekbaarheid vergroot en de gestructureerde opslag bevordert.
In deze fase wordt de keuze voor een standaard bestandsformaat voor digitaal papier van groot belang. Er zijn in deze beslissing drie formaten die bijzondere aandacht vereisen: Tagged Image File Format (TIFF), PDF en eXtended Markup language (XML). Welk van deze drie formaten het beste is, is afhankelijk van de eisen die bedrijven stellen aan het digitale papier wat betreft vormgeving, doorzoekbaarheid van de tekstdelen en de inhoudbeschrijvende gegevens.

Formaten

Het TIFF-formaat legt rasterinformatie vast zonder gegevensverlies. De scanner plaatst de informatie van het papier in een raster met een door de gekozen scanresolutie bepaalde rasterdichtheid. Het resultaatbestand wordt vervolgens gecomprimeerd via de ITU T4/6-compressiemethode, ook wel faxcompressie genoemd.
PDF is op dit moment vermoedelijk het meest bekende digitale papierformaat. Het woord portable verwijst naar de kleine bestandsomvang en de eenvoudige uitwisselbaarheid van gegevens in het PDF-formaat. PDF ontstond tien jaar geleden als opvolger van Adobe Postscript en is sindsdien nog steeds vrijwel 100 procent backwards compatible. Het PDF-bestandsformaat voldoet ook aan de eisen die de overheid stelt aan de

duurzaamheid van digitale archieven. In de toelichting op de archiefwet uit 1995 wordt verwezen naar PDF als een geschikt en duurzaam opslagformaat. Softwareleveranciers hebben in-middels begrepen dat hun producten onverkoopbaar zijn als ze niet tien keer het woord XML noemen, dus moet ook het digitale papier in de vorm van een XML-bestand worden opgeslagen. Nadeel hiervan is dat XML de oorspronkelijke vorm niet behoudt. De meeste XML-archieven bestaan uit de tekstdelen en een TIFF-bestand met de oorspronkelijke afbeelding. Maar daarmee wordt voorbijgegaan aan het oorspronkelijke doel van digitale gegevensopslag, namelijk het aanbrengen van structuur. Een groot deel van het papier in de archiefkasten kenmerkt zich juist door het gebrek aan structuur. Handgeschreven notities worden in XML niets meer dan een XML-envelop rond een TIFF-plaatje zonder enig beschrijvend gegeven over de inhoud van de afbeelding. In heel veel gevallen besluiten bedrijven omwille van het beheer van deze XML-bestanden zelfs om de binaire afbeeldingen in een tekstvorm op te slaan in het XML-bestand, zodat ze de inhoud gemakkelijk kunnen blijven uitwisselen. Dit lijkt nogal omslachtig, vooral omdat deze binaire bestanden minimaal 30 procent toevoegen aan de totale bestandsgrootte van documenten.

Metadata

Computers kunnen iets bijzonders doen met digitaal papier in de vorm van layer-techniek. Het digitale papier bestaat uit een laag met de afbeelding van het document, een laag met de teksten en een aantal lagen die de inhoud van het papier beschrijven. Deze beschrijvende gegevens worden aangeduid als metadata en kunnen als verpakking om het papieren document worden geplaatst of worden ingesloten in de inhoud ervan. Door gebruik te maken van metadata krijgen gebruikers informatie over de auteur en de inhoud van documenten en hoe vaak de informatie is geraadpleegd. Deze gestructureerde manier van werken, helpt mensen met het zoeken naar informatie en zorgt er tevens voor dat software bij de zoekopdracht kan worden ingezet. Een PDF-bestand is in deze layerstructuur niets anders dan een TIFF-afbeelding met daaronder de tekstdelen en als verpakking een aantal velden die de inhoud van het document beschrijven. Hierbij is PDF een herbenoemd XML-bestand met een door Adobe gedefinieerde functie en vorm in de PDF-namespace. In feite is het nieuwe PDF-formaat dan ook een voor de gebruikers van digitaal papier gemaakt XML-bestand met gratis software om de bestanden te lezen, te doorzoeken en af te drukken. De TabletPC voegt aan dit digitale papier een nieuwe dimensie toe met digitale inkt. Het is namelijk mogelijk om aantekeningen als laag bovenop de afbeelding aan het digitale PDF-papier toe te voegen. Hiermee profiteren gebruikers van alle voordelen van het werken met papier en van de voordelen uit de digitale wereld. Aantekeningen kunnen namelijk met één druk op de knop worden verwijderd of naar digitale tekst worden vertaald en aan bestanden worden toegevoegd.

PDF voldoet aan de eisen die de overheid stelt aan digitale archieven

Obstakels

Naast de vele voordelen van digitalisering van het papieren archief kleeft er ook een aantal nadelen aan. Bovenaan de lijst van bezwaren staat de enorme investering in tijd en dus geld. De kosten voor het converteren van een papieren document naar digitaal papier kunnen oplopen van 8 tot 40 cent per pagina en op een gemiddelde kastplank liggen 15.000 pagina's aan informatie. Bedrijven hebben in de meeste gevallen simpelweg de tijd en het geld niet om papier rendabel te converteren naar digitale informatie. De meeste bedrijven hebben de afgelopen jaren vermoedelijk onvoldoende rekening gehouden met het feit dat alle papieren archieven door een scanner moeten worden verwerkt met een invoerbak van honderd vel die 50 losse pagina's per minuut aan voor- en achterzijde kan scannen. Tot overmaat van ramp bevatten vrijwel alle dossiers afwijkende papierformaten, nietjes, ingebonden stukken en post-it notes. Nadat al deze problemen zijn opgelost moet nog de nodige tijd worden geïnvesteerd in het toevoegen van de juiste beschrijvende metagegevens aan het document en de gegevens die nodig zijn om alle teksten achter de plaatjes te herkennen. Als klap op de vuurpijl laat de mate van tekstherkenning veel te wensen over, zodat medewerkers weinig zekerheid hebben over de gevonden zoekresultaten in het digitale archief. Bovendien kunnen bedrijven omwille van mogelijk ontbrekende pagina's in contracten vrijwel nooit de originele papieren stukken definitief vernietigen.

Het scheppen van realistische verwachtingen in een scanproject is een eerste prioriteit. Op die manier kunnen bedrijven afgewogen investeringsbeslissingen nemen, waarvan de resultaten niet tegenvallen.

Bart van Wanroij

Bart van Wanroij (consultancy@broekhuis.nl) is consultant bij Broekhuis Consultancy en adviseert in projecten waarin papieren archieven worden geconverteerd naar digitale archieven.