Documentuitwisseling (1)
Documenten

Door Diederik Gerth van Wijk[1]

Documenten zijn voor veel bedrijven belangrijke produktiemiddelen en vaak ook belangrijke produkten. Niet alleen binnen de uitgeverij: in allerlei andere industrietakken vormt documentatie een belangrijke toegevoegde waarde bij hun eigenlijke produkt. In een serie van vier artikelen beschrijft de auteur enkele mogelijkheden en moeilijkheden bij de uitwisseling van elektronische documenten.
In dit eerste artikel gaat hij in op het begrip „document", dat van een verzameling tekens groeit naar een onderdeel van een samenhangend documentatiesysteem, waarbinnen tekst, cijfers, grafieken, vector- en rastertekeningen, gestructureerde gegevens en verwijzingen geïntegreerd worden. In de volgende artikelen zal hij ingaan op de problemen die ontstaan als je dergelijke complexe onderdelen wilt gaan vastleggen, beschrijven, uitwisselen en aan de gebruiker presenteren. Twee kernvragen spelen door de hele serie een rol. Wisselen we vorm of inhoud uit, en moet de standaardisatie die voor uitwisseling nodig is ten koste van de flexibiliteit gaan?

Op twee manieren worden produkten ingewikkelder. Ze bestaan uit meer onderdelen, en ze moeten aan meer andere produkten gekoppeld kunnen worden om zo samen weer een nog complexer produkt te vormen. Een toepasselijk voorbeeld is de ontwikkeling van pen en papier via tikmachine en papier naar toetsenbord, computer, magneetschijf, beeldscherm en printer.
Op eveneens twee manieren is deze stelling voor een verhaal over documenten relevant. In de eerste plaats leiden complexere systemen en produkten tot een exponentiële toename van de documenten die nodig zijn om ze te beschrijven. Zowel de afzonderlijke onderdelen moeten immers beschreven worden, als de manier waarop ze verbonden zijn en samenwerken. In de tweede plaats gaat de stelling ook op voor documenten zelf. Ook deze worden alsmaar complexer, en moeten hoe langer hoe meer deel uitmaken van een complex informatiesysteem.
De eerste reden verklaart waarom een steeds grotere groep mensen en bedrijven er belang bij heeft om documentatie te structureren en (elektronisch) beheersbaar te maken. Zoals we de afgelopen jaren bezig zijn geweest met het verwerken van getallen en namen in relationele databases, zo zullen we de komende jaren bezig zijn met het elektronisch opslaan en toegankelijk maken van langere teksten, plaatjes, tabellen en geluid.
In deze korte serie ga ik eerst in op dat wat we willen opslaan: documenten. In de volgende aflevering gaan we kijken naar de moeilijkheden die we tegenkomen als we deze gaan uitwisselen. In de derde aflevering licht ik een van de oplossingen voor die problemen (en veroorzaker van andere problemen) toe: de *Standard Generalized Markup Language* (SGML). In de laatste aflevering kijken we naar het uiteindelijke doel van al die documentatie: de presentatie aan de gebruiker.

Wat is een document

1Diederik Gerth van Wijk is stafmedewerker redactionele automatisering bij Wolters Kluwer Rechtswetenschappen.

Het is goed om eerst vast te leggen wat we onder een document zouden moeten verstaan. Vroeger was dat gemakkelijk: een document was van papier, en er stonden letters en plaatjes op.

Toen de computer aan zijn opmars begon, was zijn capaciteit niet toereikend om die beide onderdelen tegelijk aan te kunnen, en ontstond er een zekere specialisatie in programma's die met tekst, getallen, plaatjes of gestructureerde gegevens om konden gaan, waarbij de term document over het algemeen tot tekst beperkt bleef.

De volgende stap maakte het mogelijk om een tabel uit een spreadsheet als tekst in een tekstverwerker in te lezen, en zelfs om een plaatje over te nemen uit een tekenpakket, echter meestal met verlies van de achterliggende gegevens die nodig zouden zijn om het plaatje of de cijfers te manipuleren.
Vervolgens gingen de diverse pakkettenbouwers allerlei toeters en bellen aan hun programma's hangen, zodat je ook in een tekstverwerker kon rekenen, sorteren en tekenen. De programma's groeiden en groeiden, vooral tot geluk van de schijvenfabrikanten.

Elektronische documenten worden door programma's gemaakt, en elk programma gebruikt daarbij zijn eigen manier om de onderdelen van een document als zodanig te coderen. De documenten zoals we die tot nu bekeken hebben, zijn elk aan één moederprogramma gebonden. Nu gaat de tendens naar een andere richting: het document wordt de drager van heterogene gegevens, die door verschillende programma's gemaakt en veranderd worden. Een document is zo een object, dat weer objecten bevat. Elk object bestaat dan uit gegevens (tekst, getallen) en verwijzingen naar de methoden om die gegevens te manipuleren (het externe programma).
We komen dus weer dichter bij de oorspronkelijke betekenis, en kunnen in één document zowel

tekst als getallen als grafieken als tekeningen als plaatjes als databases verwerken en opgslaan. Een belangrijke verschuiving is echter, dat we bij een „document" nu niet meer in de eerste plaats kijken naar de verschijningsvorm zoals die de lezer bereikt, maar naar de achterliggende betekenis van de over te dragen gegevens, en die proberen we in elektronische vorm ondubbelzinnig vast te leggen.

De onderdelen van een document

De onderdelen die ik hiervoor al genoemd heb, ga ik hier stuk voor stuk nader uitwerken. Ik kijk daarbij naar de mate waarop vorm of inhoud opgeslagen wordt, en ook naar de mate waarop de onderdelen geordend kunnen worden.

●*tekst*: in het algemeen de hoofdmoot van een document. Tekst is geschreven taal, en bestaat uit letters die uit een beperkt alfabet gekozen kunnen worden, en op een bepaalde manier getoond worden (cursief, groter). Met die vorm wordt een bepaalde afwijkende betekenis weergegeven, zoals „dit is een vreemd woord" of „hier begint een nieuwe paragraaf". Deze codes werden vroeger meestal direct in de tekst ingebracht; nu bieden de meeste tekstverwerkers de mogelijkheid via stijlbladen de vorm te scheiden van de inhoud. Een hiërarchische ordening is vaak via outlines of de opties waarmee inhoudsopgaven gegenereerd kunnen worden aan te geven. Daarnaast zijn vaak verwijzingen naar een pagina- of hopofdstuknummer mogelijk, en kan de tekst vaak naar verschillende uitvoerstromen geleid worden (bijvoorbeeld hoofdtekst en voetnoten).

●*getallen*: constante getallen en formules waarmee een getal uit andere getallen te berekenen valt. In rekenbladprogramma's worden deze gegroepeerd tot een tweedimensionale tabel, terwijl gelijkvormige tabellen op elkaar gelegd kunnen worden zodat zelfs een driedimensionale tabel ontstaat. Vorm en inhoud zijn per definitie goed gescheiden: voor elke cel of groep cellen kan worden aangegeven hoe het getal weergegeven moet worden: zoveel cijfers voor de komma, zoveel erachter, terwijl het getal zelf weer uit andere getallen berekend kan worden.

●*grafieken*: in feite niets anders dan een andere manier om een reeks getallen weer te geven.

●*vectortekeningen*: tekeningen zoals die door CAD/CAM-programma's gemaakt worden, bestaande uit lijnen en cirkelbogen die door de coördinaten in een twee- of driedimensionale ruimte bepaald worden. Vectortekeningen zijn een afbeelding van een achterliggende betekenis. Ordening is mogelijk door onderdelen aan elkaar te plakken, en daarmee figuren te construeren die in hun geheel manipuleerbaar zijn.

●*rastertekeningen*: tekeningen die bestaan uit verschillend gekleurde puntjes, zonder achterliggende betekenis, zoals gescande plaatjes, of (in bewegende vorm) videobeelden. Puur vorm dus, zonder enige inhoud of ordening, en daarmee in feite voor computerbewerking weinig interessant.

●*gegenereerd geluid*: het acoustische equivalent van een vectortekening of een grafiek. We kunnen puur geluid, op basis van een partituur, onderscheiden, dan wel gesproken tekst, op basis van geschreven tekst.

●*opgenomen geluid*: het equivalent van een rastertekening. Het bekendste voorbeeld is de CD. Diverse onderdelen kunnen weliswaar in een soort inhoudsopgave gebracht worden, maar binnen een nummer kun je niet meer de verschillende muziekinstrumenten of stemmen onderscheiden.

Vorm en inhoud van tekst

Als we de genoemde onderdelen met elkaar vergelijken, valt op dat het onderscheid tussen vorm en inhoud bij tekeningen en geluid enerzijds en bij tekst en getallen anderzijds van een heel andere orde is. Dat komt natuurlijk, doordat letters, cijfers en formules een al duizenden jaren toegepaste vorm van digitalisering betreffen.

Vergelijken we dan weer tekst enerzijds en getallen anderzijds, dan valt op dat bij tekst nauwelijks sprake is van formules. Als we kijken hoe nu een goed gestructureerde tekst in de computer wordt opgeslagen, dan zit daar vaak een hiërarchie in, maar het laagste niveau daarbinnen is meestal de alinea. Daarbinnen vinden we dan letters en leestekens, en eventueel stukjes tekst met een bepaalde nadruk, maar die letters vormen geen expliciet gemarkeerde woorden en die woorden geen zinnen. Zelfs tussen de punt na een initiaal of andere afkorting en de punt die een zin beëindigt wordt geen onderscheid gemaakt, wat het moeilijk maakt aan de computer een opdracht te geven als „geef me de artikelen waar in één zin de woorden *punt* en *zin* voorkomen".

Zeker bij teksten die in meer talen verspreid worden, zoals technische handleidingen, zou het te overwegen zijn de dieptestructuur van de zinnen expliciet te coderen, en onderdelen als *onderwerp*, *gezegde*, *lijdend* en *meewerkend voorwerp* als zodanig te markeren, en in plaats van de letters waaruit de woorden bestaan, een verwijzing naar een woordenboek te coderen. Om een zo gecodeerde tekst af te drukken heb je dan dat woordenboek en een grammatica nodig. Dat is onmiskenbaar een stap extra, maar als je om een vertaling te maken alleen maar een ander woordenboek en een andere grammatica aan dezelfde dieptestructuur hoeft te hangen, bereik je een revolutie die vergelijkbaar is aan het structureel coderen in combinatie met een stijlboek.

Voor de goede orde moet ik er echter helaas aan toevoegen, dat deze scheiding van vorm en inhoud het stadium van fictie nog niet voorbij is.

Ordening in ruimte en tijd

Er zijn drie belangrijke verschillen tussen papieren en elektronische documenten die nog wat nadere toelichting behoeven. Zij hangen beide samen met het feit dat electronische documenten met behulp van een machine geraadpleegd worden, en dat ze pas op het moment van raadpleging toonbaar gemaakt worden.

Het eerste aspect is dat van tijd. Dat maakt het niet alleen mogelijk om geluid en video te presenteren, maar bijvoorbeeld ook om niet alleen een blaadje met 12 *piecharts* met maandgemiddelden van de omzet van een aantal produkten te drukken, maar om in een soort filmpje in tien seconden de cijfers van dag tot dag te tonen, waarbij we de taart zien groeien en krimpen, en de punten van vorm veranderen.

Het tweede aspect is dat van de interactie. Zo kan voornoemd filmpje voor- en achteruit gespoeld worden, stilgezet op een interessante dag; maar je kunt ook de mate van verfijning instellen: wil je van alle 25 produkten een eigen puntje, of wil je ze groeperen tot vier produktgroepen. Of je bent alleen maar in één van de produkten geïnteresseerd, en wilt zien hoe de afzet daarvan zich in de verschillende regio's heeft ontwikkeld: dus niet naar gegroepeerd naar regio en gesplitst naar produkt, maar één produkt gesplitst naar regio.

Het derde aspect hangt daar weer mee samen: de ordening en selectie van de verschillende onderdelen van het document. Bij een papieren document bepaalt de auteur of de redacteur de enige juiste volgorde waarin alles staat, ook al is die niet voor elke lezer de meest relevante. Dit speelt uiteraard niet bij elk soort document even sterk: bij een roman of een monografie, waarin de auteur een bepaald betoog opbouwt, is er inderdaad maar één goede volgorde. Maar bij een collectie onderling vergelijkbaar gestructureerde gegevens, zoals een produkten- of een adressenbestand, is het wel degelijk plezierig deze interactief te kunnen groeperen. Soms ben je immers in mensen met een zelfde naam geïnteresseerd, en soms weer in mensen die in een bepaalde buurt wonen, en dan weer in een bepaalde inkomensgroep. Om dit te kunnen doen, is het noodzakelijk de structuur expliciet te coderen.

Tot slot

We hebben gezien waaruit een document kan zijn samengesteld, en wat je allemaal met een elektronisch document zou willen kunnen doen. We gaan ons in de volgende aflevering buigen over de manier waarop we zulke documenten kunnen verspreiden: van de eenvoudige tekst die een auteur naar zijn redacteur wil sturen tot de multimediale CD-I die de consumentenmarkt moet bereiken.