Tekst **Frank Jan Bertram**
Frank Jan Bertram is metadata-onderzoeker bij HOTCOLDFROZENDATA

Hoe archiveert een computer?

Hoe archiveert een computer? Dat is natuurlijk een onzinnige vraag. Een computer doet wat een mens erin stopt. Maar toch zijn we computers als entiteiten gaan zien die los van ons staan en die hun eigen leven leiden, vandaar deze titel. Ik probeer te beschrijven wat wij in computers gestopt hebben om ze te laten archiveren.

Het duurzaam toegankelijk maken van informatie kan benaderd worden vanuit het perspectief van Informatie Management (IM) en dat van Informatie Technologie (IT). In de loop van de tijd is er op dit vlak tussen deze twee een kloof ontstaan. Dit artikel kiest een uitleg vanuit het IT *archiving*-perspectief, omdat het begrijpen van elkaars wereld en een bijbehorend aha-erlebnis over het algemeen een aanzet geeft tot het overbruggen van een kloof.

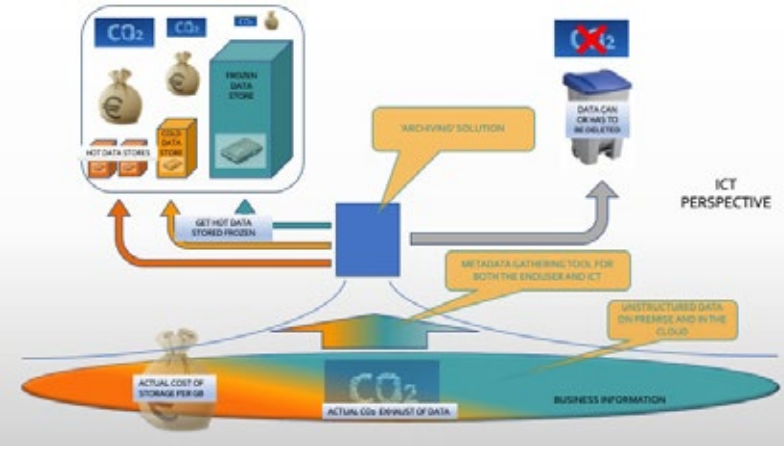
Betrouwbaar en beschikbaar

Het opslaan van informatie op media als tape, floppy, cd-rom, harddisk en flashdisk, kende vanaf de start het probleem van fysieke begrenzing en transitie: vol was vol. Papier kende dat probleem niet, mits er uiteraard voldoende papier en archiefruimte beschikbaar waren. Bovendien konden data corrupt (onbruikbaar) raken en dan was het nodig dat een niet-corrupte versie aanwezig was, een back-up. Op de back-up tape stonden dan achter elkaar de "nog werkende" laatste data van de dag of de week daarvoor, afhankelijk van de frequentie van het maken ervan. Dat waren de digitale archieven van de organisatie in het kader van technische betrouwbaarheid. Maar de informatie op de back-up was niet direct vindbaar, beschikbaar of leesbaar, laat staan interpreteerbaar. Pas na het terugzetten van de data was dat wel weer het geval. De back-up was en is voor informatie die niet op papier staat het laatste vangnet. Data moesten voortdurend beschikbaar zijn. Voor de komst van internet was een hoge beschikbaarheid van minder belang, maar het ontstaan van de 24/7-economie vroeg om de continue beschikbaarheid van data. Daarom werden data op *storage disks* gespiegeld, bij voorkeur op verschillende locaties want dan kon na een calamiteit op de ene locatie de andere het per direct overnemen. Gespiegelde data mogen niet verward worden met geback-upte data. Corrupte data op één disk/locatie betekent namelijk automatisch ook corrupte data op de andere. Dan blijft het noodzakelijk dat er goede, niet-corrupte data aanwezig zijn. Tegenwoordig is het principe van gespiegelde data overigens ongewijzigd, in het eigen datacenter en in de cloud[1].

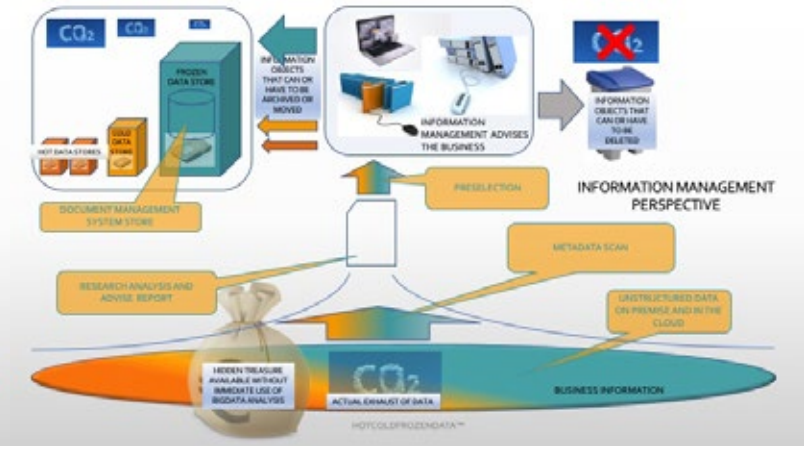
Leesbaar, vindbaar en interpreteerbaar

Data waren leesbaar voor een mens mits de juiste dataopslaglezer en bijpassende applicatie beschikbaar waren. Data waren vindbaar via een indexeringssysteem dat de metadata en de inhoud van files bijhield (bijvoorbeeld Explorer). Data werden interpreteerbaar via de computer door specifieke algoritmes – door de mens geschreven – op data los te laten (big data-analysis) met als meest recente stap zelflerende computers die zelf de algoritmes samenstellen (kunstmatige intelligentie/AI).

1 Ontbreken van back-up is nu aan de orde bij opslag in de cloud. Een met de C:\ schijf gesynchroniseerde OneDrive, Dropbox of Google Drive, etcetera, is een min of meer gespiegelde versie van de data, maar let op: weggooien op de C:\ schijf betekent direct weggooien in de cloud, er is geen back-up! Breng dit risico zo snel mogelijk in kaart en neem actie door de betaalde versie te gebruiken met de back-up-optie aan of verplaats de data naar een eigen share die wel elke dag geback-upt wordt.



Archiving gezien vanuit een Informatie Techonologie-perspectief



Archiving gezien vanuit een Informatie Management-perspectief

Eind jaren negentig werden op het gebied van bewaren van data vier belangrijke ontwikkelingen zichtbaar:

1. Transport van data via e-mail nam significant toe.
2. De hoeveelheid digitaal beeldmateriaal nam enorm toe.
3. Verschillende digitale dragers kregen uiteenlopende specificaties, de één sneller, betrouwbaarder en beschikbaarder dan de ander, met navenante prijsverschillen per eenheid opgeslagen data.
4. De eigenaar van de data kon de data niet meer opruimen. Door de grote hoeveelheid was er geen tijd meer voor interpretatie.

Archiving

Door toename van data en kosten werd de transitie van de berg van ongebruikte data naar een ander goedkoper opslagmedium interessant, want er was fors geld mee te besparen. Het ESSDC/Ozzodata project[2] toonde in 2009 bovendien aan dat hiermee flinke IT-gerelateerde CO_2-reducties behaald konden worden. De oplossing voor die transitie naar goedkopere opslagmedia werd door Engelse, Amerikaanse en Japanse hard- en softwareleveranciers *archiving* genoemd.

Hoe werkte die archiving?

Het werkte als volgt. Als de datum van creatie (metadata) van data in de vorm van files en/of mails een in te stellen periode overschreed, gebeurden er twee dingen: er werd een "punaise" (een stub) gecreëerd die achterbleef op het huidige (kostbare) opslagmedium waarna de oorspronkelijke file en/of de e-mail met bijlage ongewijzigd naar een ander opslagmedium verplaatst werd. De punaise verwees naar de nieuwe locatie. De data bleven beschikbaar en vindbaar. Als de bovengenoemde datum de datum overschreed van een, door regelgeving opgelegde periode van bewaarplicht, werd deze vernietigd, automatisch en zonder te kijken naar inhoud. Tegenwoordig zijn deze *archiving*-principes ongewijzigd.

Rol van metadata

Vindbaar bleven de files via de punaise, ongeacht op welk medium ze waren opgeslagen, maar op basis van welke regel moest er vernietigd worden en wie werd er verantwoordelijk voor het instellen van die regel? IT zeker niet. In de praktijk bleek dat er met name één afdeling bij betrokken werd, de juridische. Deze werd gevraagd om een wettelijk kader. Informatie Management stond er verder vanaf. Immers, zij regelden het vernietigen van documenten voor de organisatie in het documentmanagementsysteem (DMS), zij hadden dat al lang geregeld. Maar uit diverse metingen[3] bleek dat slechts gemiddeld 15-20 procent van alle data in het DMS zat. De resterende 80-85 procent was niet gearchiveerd, niet duurzaam toegankelijk. Wie was daar verantwoordelijk voor?

80-85 Procent opstap naar archiveren by design?

Er is veel ondernomen om dit probleem op te lossen, leveranciers gingen onder andere DMS-achtige containers bouwen zoals Microsoft dat bijvoorbeeld deed met SharePoint. Maar feit blijft dat er nog veel ongestructureerde data aanwezig blijven. Daarom is het interessant te weten dat parallel aan deze ontwikkelingen de oorspronkelijk al door *archiving*-systemen gebruikte metadata een betrouwbare opstap bleken te bieden om belangrijke delen van bovengenoemd probleem van 80-85 procent op te lossen zonder grote investeringen in technologie en met inzet van bestaande kennis. Door voorselecties op basis van deze metadata te realiseren, werd de data-berg meetbaar, inzichtelijk en konden specialisten (IT'ers samen met IM'ers) de informatie als interne dienstverleners voor de eindgebruiker duurzaam toegankelijk maken.

2 Ozzodata was de projectnaam van het Energy Self Sufficient Data Center (ESSDC), een project gestart vanaf het GreenIT Amsterdam platform in het kader van het vergroenen van IT.

3 MDES deed datametingen bij organisaties in de periode van 2002 -2012. Door het meten van de totale data van een organisatie bleek dat de DMS-database gemiddeld 10-15 procent van de totale data bevatte.