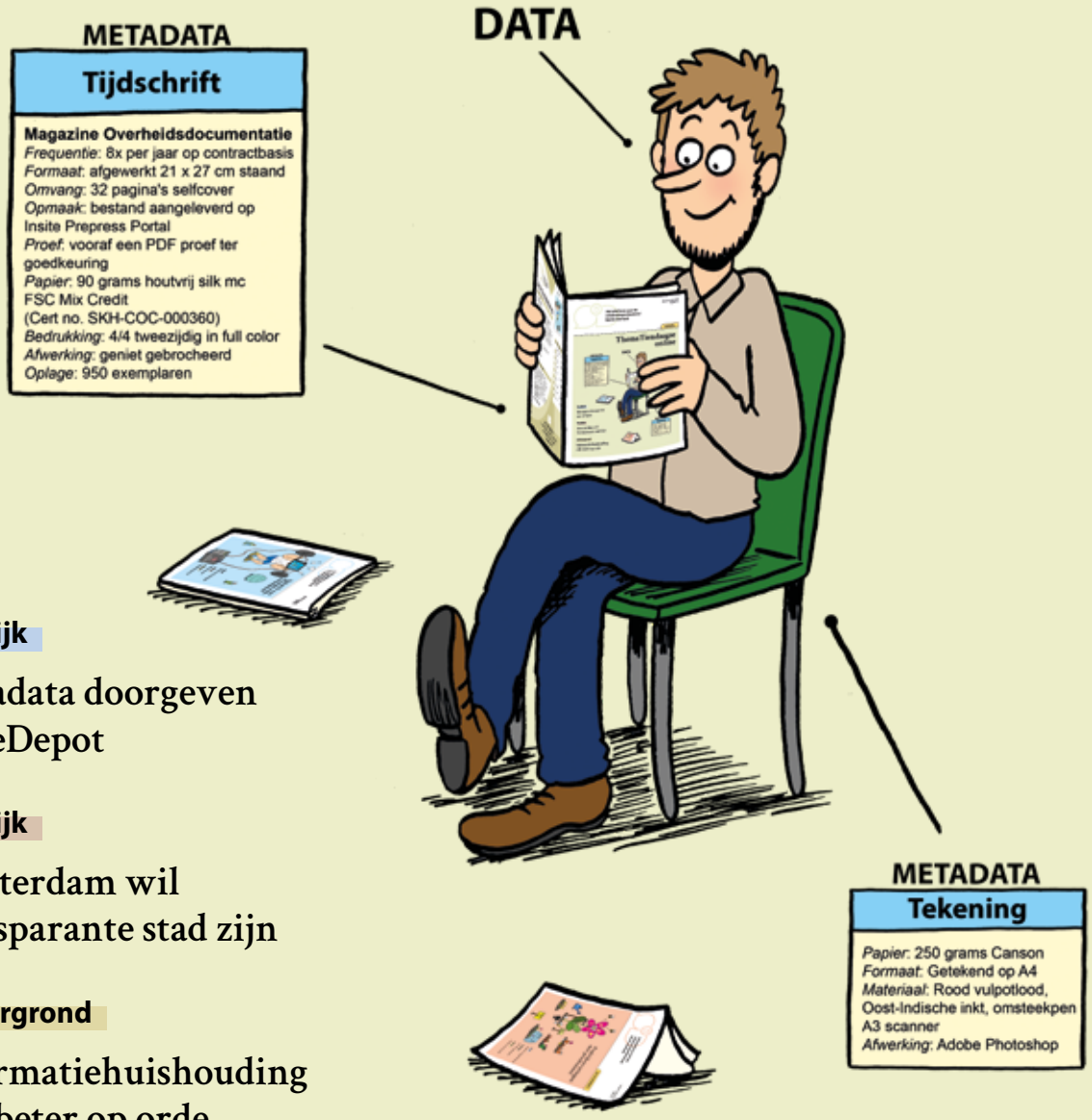
Jaargang 75 | februari 2021
Uitgave nr. 10

75 JAAR

Hét platform voor de informatieprofessional bij de overheid

METADATA

ThemaTiendaagse online



Praktijk

Metadata doorgeven aan eDepot

Praktijk

Amsterdam wil transparante stad zijn

Achtergrond

Informatiehuishouding rijk beter op orde

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

Technische metadata

Opstap naar archiveren?

Metadata worden door *IT-archiving* gebruikt als uitgangspunt voor het verplaatsen naar goedkopere opslagmedia. De achtergrond van dit proces beschreef ik in mijn eerdere artikel Hoe archiveert een computer in *Od 6*. Kunnen metadata een bijdrage leveren aan het archiveren van de 80 tot 85 procent ongestructureerde data zonder grote investeringen in technologie en met inzet van bestaande kennis? Om die vraag te kunnen beantwoorden moeten we eerst de achtergrond van technische metadata kennen.

Volgens het Nationaal Archief zijn metadata (ook wel metagegevens genoemd) gegevens die de karakteristieken van gegevens beschrijven. Het zijn eigenlijk data over data. Voorbeelden van karakteristieken zijn de *creator*, de datum van creatie, de gebruikte taal en het bestandsformaat. Metadata beschrijven niet alleen de gegevens zelf, maar ook de context waarbinnen de gegevens zijn ontstaan of ontvangen. En wat er vanaf het moment van ontstaan of ontvangst met die gegevens is gebeurd.
Volgens deskundige op het gebied van data- en metadatamanagement Donna Burbank van DAMA (The Global Data Management Community) zijn er twee soorten metadata te onderscheiden:

'METADATA PROVIDE INFORMATION ABOUT OTHER DATA, INCLUDING A DESCRIPTION OF THE DATA'
(Library Stanford)

1. Technische metadata; beschrijven de structuur, het format en de regels om data op te slaan, in een file-systeem en/of in een database-systeem.
2. Zakelijke metadata; beschrijven de organisatie-definities, -regels en -context van data.

Ik beperk me tot de achtergrond van technische metadata in het file-systeem, omdat file-data verreweg het grootste deel van de 80 tot 85 procent ongestructureerde data vormen.

Technische metadata zijn al vanaf het prille begin in gebruik, omdat ze onmisbaar blijken voor het goed functioneren van een computer. Waarom zijn ze onmisbaar?

Hoofdbestandstabel

Een computer zet bits fysiek op een opslagmedium, zoals een harddisk of een *flashdrive*. Dit wordt geregeld in de bestandssysteem(filesysteem)-software, onderdeel van de besturing(operating)-software van de computer. Leveranciers hebben veelal hun eigen filesystemstandaarden gecreëerd. De volgende uitleg is gebaseerd op die van Microsoft, omdat de meeste ongestructureerde documenten van de 80 tot 85 procent vanaf het begin van de jaren 80 door middel van Microsoft-filesys-

tem-software zijn opgeslagen.
Het bestandssysteem NTFS (New Technology File System) van Microsoft bevat een bestand dat de hoofdbestandstabel (Master File Table (MFT) wordt genoemd (voor 1993 File Allocation Table (FAT)). Er is tenminste één vermelding in de hoofdbestandstabel voor elk bestand op een NTFS-bestandssysteemvolume, inclusief de hoofdbestandstabel zelf. Alle informatie over een bestand, inclusief de grootte, tijd- en datumstempels, machtigingen en gegevensinhoud, wordt opgeslagen in hoofdbestandstabel-vermeldingen. Deze technische metadata worden door Microsoft *attribute*s genoemd.[1] Door middel van deze hoofdbestandstabel met technische metadata vindt de computer de documenten (files) op de fysieke drager terug. De creatie van deze metadata door het filesystem is daarmee essentieel.

Prioriteren

We kunnen concluderen dat technische metadata:

- sinds lange tijd binnen Microsoft-omgevingen gestandaardiseerd zijn;
- altijd beschikbaar zijn, want noodzakelijk voor de werking van een computer;
- altijd actueel zijn door de volledig geautomatiseerde aanmaak en mutatie;
- een index vormen die in de verkenner te raadplegen is.

Het blijkt dat technische metadata sinds het begin van de jaren 80 volgens een standaard in de FAT en NTFS aan files gekoppeld zijn. Dit is vanuit een archiveringsstandpunt interessant. Is hier een houvast voor indexering zonder bijvoorbeeld de directe beschikbaarheid van bij de documenten horende applicaties? Kunnen we op basis van deze metadata gaan prioriteren in het proces naar definitieve archivering en de achterstanden vervolgens handmatig of geautomatiseerd gestructureerd wegwerken?

Technische metadata / File Attributes

Engels	Nederlands	Functie
Share	Locatie	Op welke gedeelde locatie staat de folder of het bestand. vb. H:\ Dit is een logische aanduiding van een fysieke opslag schijf in een eigen datacenter of in de cloud.
Name	Naam	Door de gebruiker aan de map of het bestand gegeven naam.
Path	Pad	Het "pad" naar het bestand vb. H:\departement\afdeling\staf\
Fullname	Volledige naam	Pad plus bestandsnaam.
File extention	Extensie	De letters achter de punt in de bestandsnaam.
File Size	Bestandsgrootte	De omvang van het bestand in Bytes.
Creation Time	Aanmaak tijd	De datum en tijd van aanmaak van het document.
LastAcces Time	Opening tijd	De laatste datum en tijd van openen van het document.
Lastwrite Time	Bewerking tijd	De laatste datum en tijd van bewerking van het document.

Grip

Hoe zou dat er praktisch uitzien? Stelt u zich voor: een afdelingsshare met vele duizenden documenten die al sinds de eerste netwerken uit de tijd van Novell bestaat.[2] De documenten zijn in de loop van de tijd om de 3 tot 5 jaar fysiek naar nieuwe opslagmedia overgezet, in logische zin zijn ze nog steeds aanwezig in dezelfde afdelingsshare. De documenten zijn niet gearchiveerd volgens de archiveringsnormen.
Een insteek kan zijn dat we op folder-namen (mappen) selecteren die in eerste instantie waarschijnlijk de meest archiefwaardige documenten bevatten. In de geselecteerde folders sorteren we vervolgens op creatiedatum, waarna we direct een indruk krijgen van de oudste documenten die het eerste in aanmerking komen voor archivering. Als we de documenten eerst sorteren op bestanden waarvan de grootte, de creatiedatum en de extensie gelijk zijn, krijgen we direct een subset van documenten die waarschijnlijk snel te ontdubbelen zijn. Dit kan vervolgens door de archivaris in opdracht van, of door de informatie-eigenaar zelf worden uitgevoerd.
Na ontdubbelen van deze selectie kan gekeken worden naar documenten waarvan de bestandsnamen overeenkomen en waarvan een doc- en een pdf-exemplaar van bestaat. Zeer waarschijnlijk is de uitkomst van deze selectie relatief klein en kan vervolgens gearchiveerd worden, ook nu door de archivaris in opdracht van, of door de informatie-eigenaar zelf, handmatig of geautomatiseerd.
Deze aanpak geeft de archivaris de mogelijkheid aan de organisatie een archiveringdienst te verlenen zonder diepgaande inhoudelijke kennis van IT-technologie. Door het gebruik van technische metadata van documenten kan de afdeling archief haar grip op ongestructureerde informatie in de organisatie vergroten en met een gestructureerde aanpak de 80 tot 85 procent ongestructureerde data op de juiste wijze archiveren.

'METADATEN [...] DIENEN DER REPRODUZIERBARKEIT VON FORSCHUNG UND DER AUFFINDBARKEIT VON DATEN'
(Universität Hamburg)

1 The webarchives. 2008, Microsoft, NTFS Metadata creation.

2 Novell Netware was in de jaren 80 en 90 de meest gebruikte netwerksoftware waarmee binnen organisaties pc's en servers onderling verbonden werden.