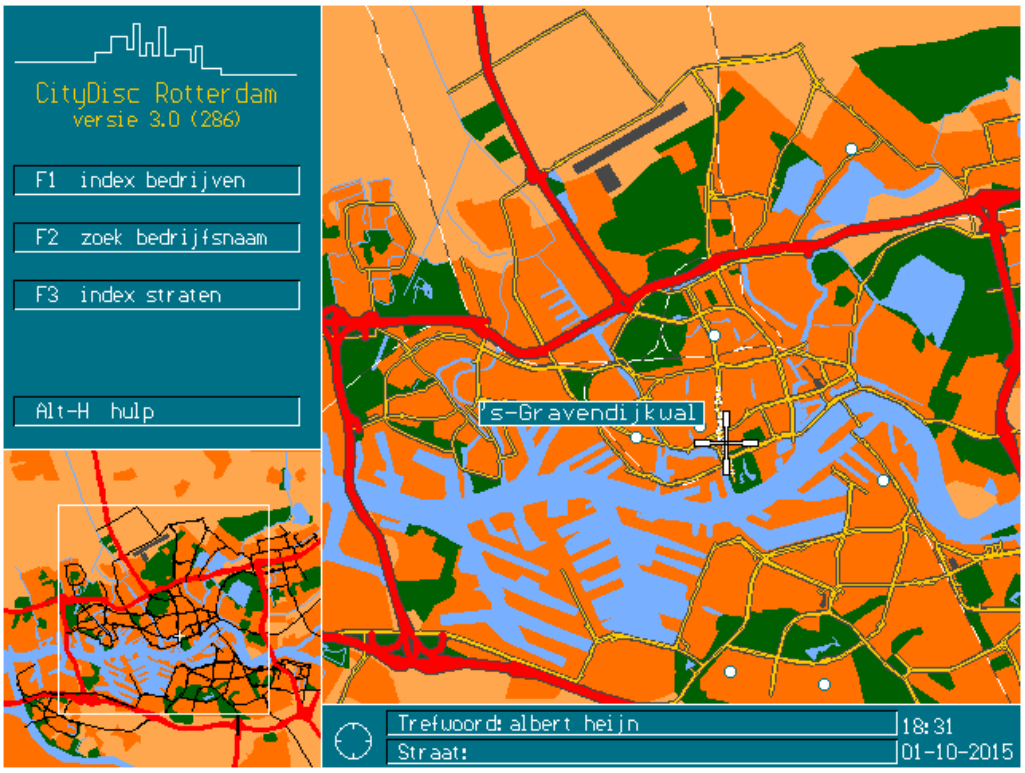
# Behoud offline optische dragers

Johan van der Knijff, Afdeling Onderzoek (met input van Barbara Sierman, Yvonne van der Steen, Han Dammann, Mark van Egmond)

Externe versie

19 November 2015



# Inhoud

# 1. Managementsamenvatting

DE KB heeft een omvangrijke collectie offline optische dragers. Naar schatting gaat het hierbij om ongeveer 15.000 fysieke dragers. Zo'n 60% hiervan bestaat uit CD-ROMs, en 30% uit audio CD's. De resterende 10% bestaat vooral DVD's. Veel van onze CD-ROMs bevatten software, die bedoeld is om in een specifieke omgeving (bijvoorbeeld Windows 95) geïnstalleerd te worden. De toegankelijkheid van deze materialen wordt bedreigd door een combinatie van fysieke achteruitgang van de dragers zelf (alle binnen dit onderzoek geraadpleegde CD Recordables uit onze collectie leverden uitleesproblemen op of bleken nu al compleet onleesbaar te zijn), de teruglopende beschikbaarheid van hardware om de dragers uit te lezen, en het vaak niet beschikbaar zijn van de hard- en softwareketen om de *inhoud* van de dragers correct te renderen. Dit rapport is een verkennend onderzoek naar de beschikbare mogelijkheden om de toegankelijkheid van deze materialen op de lange termijn te kunnen waarborgen. Het combineert hiermee *Preservation Watch* (wat zijn de risico's waaraan deze collectie blootstaat) en *Preservation Planning* (wat is er nodig om deze risico's het hoofd te bieden).

## Disk imaging

De informatie op de (vergankelijke) CD-ROMs en DVD's kan worden veiliggesteld door het maken van disk images in ISO 9660 formaat. Deze images kunnen dan via emulatie aan de gebruiker beschikbaar worden gesteld. Images van video DVD's kunnen met reguliere mediaplayer software worden afgespeeld. Voor audio CD's kunnen de afzonderlijke audiotracks "geript" worden naar audiobestanden (bijvoorbeeld in het WAV formaat). Gebruikers kunnen deze bestanden dan met reguliere mediaplayer software afspelen.

## Automatisering imaging-procedure en inzet discrobot

Vanwege de omvang van de collectie is volledig handmatig imagen ondoenlijk, wat het gebruik van discrobot hardware onontbeerlijk maakt. Met een discrobot kunnen (afhankelijk van het model) telkens tientallen tot enkele honderden dragers automatisch verwerkt worden. Discrobots die in aanmerking komen zijn de *Acronova Nimbie* (gebruikt door The British Library) en het *Ripstation* systeem (in gebruik bij o.a. Library Of Congress en The British Library). Vóór aanschaf van één van deze machines is het raadzaam om zowel bij de fabrikanten als de collega-instellingen die de machines gebruiken navraag te doen naar een aantal specifieke aspecten. Het gaat hierbij vooral om de verwerking van multisessie/mixed-mode CD's, de kwaliteit van de bijgeleverde software voor audio extractie, en de mate waarin het verwerken van verschillende typen dragers (bijvoorbeeld CD-ROMs en audio CD's) binnen één batch wordt ondersteund. Een andere belangrijke overweging is de mogelijkheid om de discrobot *tijdens* het draaien bij te vullen. Verder is uit de documentatie op de websites van beide fabrikanten niet duidelijk of de machines ook kwaliteitscontroles op de gemaakte images uitvoeren (en zo ja, welke).

Hoewel je met een discrobot het eigenlijke imaging-proces deels kunt automatiseren, brengt de verwerking hoedanook veel werk met zich mee dat handmatig zal moeten plaatsvinden: ophalen van dragers uit het magazijn en weer terugbrengen, vullen en leeghalen van de discrobot, metadata invoeren, en het afhandelen van dragers die fouten opleveren. Afgaande op ervaringen van de BL gaat in zulke handelingen uiteindelijk de

meeste tijd zitten. De BL rapporteert voor een soortgelijke collectie als de onze een verwerkingssnelheid van 1050 dagers per maand. De KB collectie zou bij zo'n snelheid in ongeveer 14 maanden verwerkt zijn (dit is inclusief alle bijkomende menselijke handelingen, maar exclusief voorbereidingen zoals ontwerp en inrichten van de workfow).

## Kwaliteitscontrole

Als controle op de volledigheid van ISO 9660 images, kan voor elk image de checksum worden vergeleken met de checksum van de fysieke drager (CD-ROMs en DVD's). Voor audio CD's is deze controle niet mogelijk. Eventueel zou voor audio CD's nog wel een controle kunnen worden uitgevoerd op afzondelijke WAV bestanden. Dit kan bijvoorbeeld met de JHOVE tool.

## Metadata

Voor elk image moet een set metadata worden aangemaakt. Een uitputtende lijst van noodzakelijke metadata valt buiten de scope van deze verkennende studie, en dit moet in een vervolgtraject nader worden uitgewerkt. Een mogelijk vertrekpunt hierbij is de lijst met metadata die is gebruikt in een vergelijkbaar project bij de British Library.

Omdat software op CD-ROMs vaak afhankelijk is van een specifieke technische omgeving (bijvoorbeeld Windows 3.11), is het belangrijk dat een gebruiker de beschikking heeft over de metadata waarin dit is vastgelegd. Meestal is dit te vinden in de annotatievelden in een catalogusrecord. Het is dan wel essentieel dat de link tussen de (metadata van de) disk images en de bijbehorende catalogusrecords altijd gehandhaafd blijft. Verder bevatten bij de dragers horende verpakkingen, hoesjes en boekjes ook vaak gebruikersdocumentatie. Dit is in feite ook metadata, maar om dit beschikbaar te maken voor een gebruiker zouden al deze materialen gedigitaliseerd moeten worden. Dit brengt wel een hoop extra werk met zich mee, en de vraag is of dit haalbaar is.

## Toegang en gebruik images

De meest voor de hand liggende toegangsstrategie voor CD-ROMs en DVD-ROMs is emulatie, waarbij de oorspronkelijke technische omgeving (bijvoorbeeld Windows 95) wordt nagebootst op een "moderne"computer. Disk images van Video DVD's kunnen in standaar mediaplayer software worden afgespeeld. Hetzelfde geldt voor WAV bestanden van een "geripte" audio CD's

## Aanbevelingen vervolgstappen behoud dragers

### Stap 1: opstellen beleid

Om te beginnen is het noodzakelijk dat Collectiebehoud beleid opstelt (*Preservation Policy*) ten aanzien van deze collectie dragers. In het document "Kavels en waarde" (augustus 2014) heeft Collectiebehoud aan het kavel waarbinnen deze collectie valt lage waarden toegekend[1]. De vraag is of dit terecht is, of dat de beoordeling vooral voortvloeit uit onbekendheid met deze collectie (die o.a. veel educatieve materialen omvat). De

[1] Link: <interne link verwijderd in externe versie>

Preservation Policy dient ook duidelijkheid te scheppen over *wat* behouden dient te blijven (bijvoorbeeld: alleen de inhoud van de dragers, of ook de verpakkingen, hoesjes en boekjes). Vervolgens moet een plan worden opgesteld (*Preservation Plan*) waarin beschreven staat hoe de dragers gered kunnen worden. Dit rapport is hierbij een eerste aanzet.

### Stap 2: vervolgonderzoek

Vervolgens moet de procedure voor het behoud van optische dragers in verder detail uitgewerkt worden. Uitgangspunt hierbij is dat van CD-ROMs en DVD's images worden gemaakt in ISO 9660 formaat, en dat audio CD's worden "geript" naar WAV bestanden (CD-I's blijven buiten de scope van de geautomatiseerde workflow[2]). Dit onderzoek zal zich specifiek moeten richten op de volgende aspecten:

- De hard- en softwareconfiguratie die nodig is voor een zoveel mogelijk geautomatiseerde workflow. Concreet: welke discrobot is in ons geval het beste, en in hoeverre voldoet de standaard software die daarop draait? Dit is vooral een kwestie van informeren bij de producenten van de Nimbie en Ripstation systemen. Daarnaast kunnen we hierbij gebruikmaken van de ervaring van collega-instellingen (met name de British Library en Library of Congress).
- Nader bepalen van de metadata die bij elk image worden opgeslagen, en de vorm waarin dat het beste kan (in samenspraak met Documentverwerking en Collectiebehoud). Uitgangspunt hierbij is dat de ingest van images en metadata in het DM later zo eenvoudig mogelijk wordt.

### Stap 3: verkennen discrobot

De volgende stap bestaat dan uit het verkennen en uittesten van de discrobot. Het gaat dan vooral om de volgende aspecten:

- Hoe werken de machine en de software globaal?
- Hoe zorg je ervoor dat de gemaakte images gekoppeld zijn aan de juiste metadata?
- Hoe verwerk je verschillende typen dragers binnen één batch?
- Wat gebeurt er als een drager fouten oplevert?
- Eventuele aanpassingen aan de softwareconfiguratie.

### Stap 4: opzetten operationele workflow

Vervolgens kan de operationele workflow worden opgezet. Dit brengt waarschijnlijk nog een beperkte hoeveelheid ontwikkelwerk met zich mee (bv scripts voor automatische kwaliteitscontroles, en het genereren van technische metadata zoals checksums). In principe is dit niet de taak van de afdeling Onderzoek, maar Onderzoek kan hierbij wel een begeleidende rol spelen, en adviseren over bijvoorbeeld tools voor de kwaliteitscontrole.

---

[2]Deze zouden eventueel in een afzonderlijk project handmatig kunnen worden verwerkt.

### Stap 5: operationele fase

Vóórdat de operationele workflow daadwerkelijk in gebruik wordt genomen, is het belangrijk dat de medewerker(s) die hierbij aan de knoppen zitten een training krijgen. Het gaat hierbij vooral om het tijdig herkennen van probleemgevallen, en het op de juiste manier daarop reageren. Deze training zou ook bij voorkeur door Onderzoek verzorgd moeten worden.

Te slotte is het belangrijk dat de voorbereidende en operationele fasen niet volledig van elkaar gescheiden kunnen worden. De ervaring van de British Library bij een soortgelijk project was dat ze tijdens het productiedraaien regelmatig onverwachte dingen tegenkwamen, waardoor de workflow aangepast moest worden. Het is dus een iteratief proces. Een consequentie hiervan is dat ook tijdens de operationele fase af en toe inzet nodig is vanuit Onderzoek.

### Randvoorwaarden

- Bereidheid bij Collectiebehoud tot het aanschaffen van de benodigde hard- en software. Het gaat hierbij voornamelijk om discrobots, en voldoende redundante opslag om de images veilig op te kunnen slaan. Een grove schatting hierbij: uitgaande van 15.000 dragers en een gemiddelde grootte van 600 MB per drager[3] kom je uit op bijna 9 TB.

- Beschikbare personele inzet bij Onderzoek (JvdK), Collectiebehoud, Documentverwerking en ICT Ontwikkeling bij vervolgonderzoek en ontwerp / testen van de workflow.

- Beschikbare personele inzet tijdens operationele fase (ophalen dragers uit magazijn, terugzetten, vullen, leeghalen en bedienen van de discrobots, handmatige invoer metadata, monitoren imagingproces, afhandelen problematische dragers).

## Aanbevelingen vervolgstappen toegang op lange termijn

Als we de disk images op termijn door middel van emulatie aan onze gebruikers in de leeszalen willen aanbieden, is hiervoor ook aanvullend werk nodig. De *Emulation as a Service* (EaaS) benadering van de Universiteit Freiburg lijkt hierbij een interessante optie. Een logische vervolgstap is om deze software eens uit te testen met een lokale installatie op de KB.

Voor emulatie is de beschikbaarheid van oude legacy software van essentieel belang. Het gaat hierbij om oude besturingssystemen (bijvoorbeeld MS DOS, Windows 95), hardwaredrivers en Office software. We hebben hiervan nog een aantal dozen met oude installatiediskettes en -CD-ROMs. Deze dragers zijn ook vergankelijk; daarnaast worden bijvoorbeeld diskettes niet meer door moderne computers ondersteund. Het verdient daarom de aanbeveling om van de belangrijkste software-installers die we hebben op korte

[3]De maximale opslagcapaciteit van een CD-ROM is ruim 700 MB; voor een single-layer DVD is dit ongeveer 8 GB.

termijn disk images te maken, en deze op een veilige plek duurzaam op te slaan. Onderzoek heeft hier al een beginnetje mee gemaakt, en het plan is om deze activiteit in 2016 verder uit te breiden.

# 2. Inleiding

## Achtergrond

DE KB beschikt over een diverse collectie van offline dragers. De door Collectiebehoud in 2012 opgestelde Kavelbeschrijvingen (Kavel D 16) noemen de volgende typen *digitale* dragers:

- CD-ROM
- DVD
- DVD-video
- DVD-ROM
- Compact Disc (audio)
- CD-i
- Diskette 3.5"
- Diskette 5.25"
- USB stick

Voor al deze typen offline dragers geldt dat het niet vanzelfsprekend is dat ze op de lange termijn toegankelijk blijven. Hierbij speelt een combinatie van de volgende factoren een rol:

1. De vergankelijkheid van het fysieke opslagmedium: bij floppy disks wordt het magnetische “signaal” na verloop van tijd steeds zwakker; optische dragers als CD’s en DVD’s gaan ook geleidelijk achteruit. Dit houdt het risico in dat de informatie op de drager op een bepaald moment gewoon (deels) verdwenen is.
2. Beschikbaarheid van hardware om de dragers uit te lezen. Zo is het nu al nagenoeg onmogelijk aan diskdrives voor 5.25" floppy’s te komen. Moderne PC’s bieden ook steeds minder ondersteuning voor CD’s en DVD’s.
3. Beschikbaarheid van de hardware- en softwareketen om de *inhoud* van de drager mee te renderen. Aan een CD-ROM met een installer voor een Windows 3.11 applicatie heb je ook weinig zonder de bijbehorende hard- en softwareomgeving. Dit kun je oplossen door emulatie van de oorspronkelijke omgeving.

## Doel en afbakening

Dit rapport is een verkennend onderzoek naar de beschikbare mogelijkheden om de toegankelijkheid van onze offline digitale dragers op de lange termijn optimaal kunnen waarborgen. Met het oog op de grote diversiteit aan offline dragers, beperkt dit rapport zich tot de categorie *optische* offline dragers: CD’s (inclusief CD-ROM, CD-i en audio CD) en DVD’s (inclusief DVD video en DVD ROM). Hierbij spelen twee aspecten een rol:

1. Het behoud van de informatie op de (vergankelijke) fysieke dragers door het maken van disk images.

2. De beschikbare mogelijkheden om de disk images te kunnen lezen / renderen.

De focus van dit onderzoek ligt op het eerste aspect, maar beide kunnen niet los van elkaar worden gezien: de disk images zullen immers in een vorm moeten worden opgeslagen waar emulatiesoftware later ook iets mee kan. Dit is belangrijk, omdat bij eerdere pogingen op de KB om CD's en DVD's te imagen, gebruik is gemaakt van ongebruikelijke formaten, waardoor we met die images nu eigenlijk niets meer kunnen. Het is belangrijk dat deze fout niet opnieuw wordt gemaakt.

## Onderzoeksvragen

Het onderzoek probeert een antwoord te geven op de volgende vragen:

- Om wat voor materialen gaat het precies, en om welke aantallen?
- In welke staat bevinden de dragers zich op dit moment?
- Wat zijn de meest geschikte formaten om de disk images op te slaan?
- Welke stappen moet je doorlopen voor het imagen van optische dragers?
- Welke tools, hardware en software gebruik je hiervoor? Hoe ziet de workflow er globaal uit?
- Wat is de haalbaarheid van het toepassen van de workflow op de volledige collectie offline optische dragers?
- Hoe ziet de beschikbaarstelling van de disk images aan onze gebruikers er globaal uit?

## Afwijking van oorspronkelijke projectvoorstel

In het projectvoorstel *Voorstel pilotproject behoud offline optische* uit april 2015 ging ik er vanuit dat het mogelijk zou zijn om voor dit verkennende onderzoek al een vrij gedetailleerde opzet te maken van de imaging workflow. De aantallen dragers in onze collectie blijken nu echter fors groter te zijn dan destijds voorzien (zie volgende hoofdstuk). Dit heeft als consequentie dat imagen ondoenlijk is zonder gebruik te maken speciaal hiervoor aan te schaffen hardware. Hierdoor zijn de workflowbeschrijvingen in dit rapport vrij summier gebleven, simpelweg omdat we in dit verkennende stadium (nog) niet over de hard- en software kunnen beschikken.

# 3. Optische dragers in de KB collectie: typen en aantallen

Noch de KB catalogus noch het GGC kennen expliciete codes toe aan optische dragers. Hierdoor is het niet eenvoudig om precieze aantallen te bepalen; daarnaast maakt dit een uitsplitsing naar het type drager (bv audio CD, CD-ROM, DVD-ROM) lastig. Indirect zijn wel globale schattingen te maken.

## Schatting op basis van GGC projectcodes

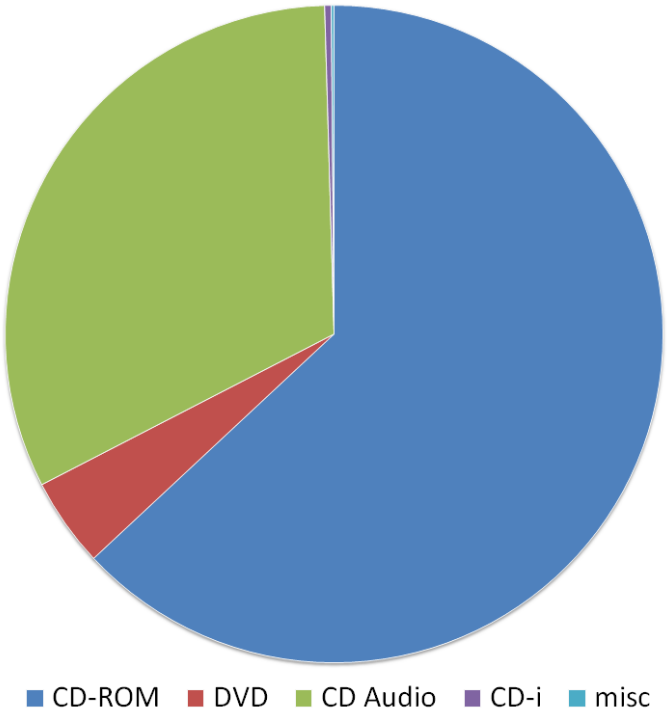
Yvonne van der Steen (Documentverwerking) heeft op basis van projectcodes in de GGC een schatting gemaakt, en kwam daarbij uit op een totaal van **15843** optische dragers. Omdat één van de gehanteerde projectcodes ook floppies en USB sticks omvat, is dit waarschijnlijk een overschatting van het daadwerkelijke aantal.

## Schatting op basis van KB catalogus

Hoewel je via de catalogus niet eenduidig op specifieke dragers kunt zoeken, is het wel mogelijk een schatting te maken op basis van het *extent* veld (weergegeven als "omvang" in de gebruikersinterface) in de catalogus. De invulling van dit veld is alleen niet consequent. Zo kom je voor CD-ROM entries varianten tegen als *cdrom*, *cd-rom*, *1 cd-rom*, *cdroms*, *1 optische schijf*, *1 PlayStation 3 schijf*, enzovoort. Voor audio CDs komen varianten als *cd*, *compact disc* en *audio-cd* voor. Met een aantal gerichte query's heb ik geprobeerd tot ruwe schattingen te komen voor de aantallen dragers in de subcategorieën CD/ DVD-rom, audio CD en CD-i. De volgende tabel laat de resultaten zien:

| Type drager | Query | Aantal |
|---|---|---|
| CD-ROM | extent any "cdrom* cd-rom*" | 7980 |
| DVD | extent any "dvd*" | 554 |
| Audio CD | type any "geluidsdrager" and extent any "cd* compact" | 4065 |
| CD-i | extent any "cdi" | 41 |
| Optische drager, ongespecifieerd | extent any "optisch* schijf" | 17 |

Dit levert een totaal op van bijna **13000** treffers in de catalogus. Hierbij moet opgemerkt worden dat één treffer vaak meerdere dragers omvat (bijvoorbeeld een boek met 8 CD-ROMs en een audio CD). Visueel zien de relatieve aandelen van de verschillende typen dragers er als volgt uit:




Opvallend is het grote aantal audio CD's, en het relatief kleine aantal DVD's. Overigens zijn niet alle items die in de catalogus als "geluidsdrager" staan aangeduid daadwerkelijk geluidsdragers. Zie bijvoorbeeld deze link naar het Srebrenica rapport van het NIOD; de bijbehorende cd-rom is (waarschijnlijk per ongeluk) ingevoerd als "geluidsdrager".

## Conclusies

Het totaal aantal dragers ligt waarschijnlijk ergens tussen de 13000 en 16000; mogelijk komt dit zelfs nog hoger uit. Ongeveer 30% hiervan zijn audio CD's; de resterende 60% bestaat vooral uit CD-ROMs. Het aandeel DVD's is beperkt. Ten slotte hebben we nog 41 CD-i schijfjes.

# 4. Verschillen tussen typen dragers

Het voert te ver om hier een volledige opsomming te geven van alle verschillen tussen de diverse type dragers. Ik beperk me hier tot een globaal overzicht, grotendeels gebaseerd op Alexander Duryee's artikel *An Introduction to Optical Media Preservation*[4].

## CD-ROM, DVD

Belangrijk is dat CD-ROMs, DVD-ROMs en video DVD's allemaal een *bestandssysteem* hebben dat fysieke locaties op de schijf expliciet koppelt aan bestanden en mappen. Hierdoor kan een gebruiker via het besturingssysteem (bijvoorbeeld Windows Explorer) door de inhoud bladeren, net als bij een harde schijf. Gangbare bestandssystemen zijn *ISO 9660*[5], *Joliet*[6] (een extensie van ISO 9660), *HFS*[7] (van Apple) en *UDF*[8](Universal Disk Format, de opvolger van ISO 9660).

### Behoud informatie op drager

Van dragers met een bestandssysteem als ISO 9660 of UDF kun je eenvoudig een imagebestand ("ISO image") maken. Zo'n ISO image is niets anders dan een tot op de byte identieke kopie van de gegevens op de fysieke drager. Voor het maken van een ISO image zijn talloze softwaretools beschikbaar; onder Linux kun je bijvoorbeeld de command-line *readom* tool[9] gebruiken; verder zijn er verschillende softwarepakketten met een gebruiksvriendelijke grafische interface, zoals bijvoorbeeld *IsoBuster*[10].

De ISO images zijn vervolgens op verschillende manieren benaderbaar. In een emulator kun je de image bestanden laden als "virtuele CD-ROM" (of DVD); eenmaal geladen kun je ze vervolgens op precies dezelfde manier gebruiken als de originele fysieke drager. Veel moderne besturingssystemen (bijvoorbeeld Windows 8 en Linux varianten als Ubuntu of Linux Mint) kunnen op een vergelijkbare manier ISO images laden. Verder kunnen ze gelezen worden met diverse data(de)compressietools, zoals *7-Zip*[11].

---

[4]Link: http://journal.code4lib.org/articles/9581

[5]Link: https://en.wikipedia.org/wiki/ISO_9660)

[6]Link: https://en.wikipedia.org/wiki/Joliet_%28file_system%29

[7]Link: https://en.wikipedia.org/wiki/Hierarchical_File_System

[8]Link: https://en.wikipedia.org/wiki/Universal_Disk_Format

[9]Link: http://linux.die.net/man/1/readom

[10]Link: http://www.isobuster.com/nl/

[11]Link: http://www.7-zip.org/

## Audio CD

Een audio CD heeft *geen* bestandssysteem, hierdoor is er ook geen sprake van afzonderlijke bestanden. De gegevens op een audio CD bestaan uit een onafgebroken stroom data, die tijdens het afspelen op een vaste, constante snelheid wordt uitgelezen. Elke audio CD bevat een inhoudsopgave (Table Of Contents, TOC) met daarin de startposities van afzonderlijke audiotracks. De startposities in de TOC zijn gedefinieerd als *tijdcodes* ten opzichte van het begin van de audiostroom (bijvoorbeeld: startpositie van track 5 op 20 minuten en 3.67 seconden). De TOC verwijst dus niet direct naar *bytepositie*s in de datastroom (wat bij een regulier bestandssysteem wel het geval zou zijn).

### Behoud informatie op drager

Het imagen van audio CD's is om een aantal redenen lastiger dan 'data' CD's en DVD's. Allereerst zijn door het ontbreken van een bestandssysteem de audiodata niet direct benaderbaar door het besturingssysteem van de machine waarop het image wordt gemaakt. Hierdoor is voor het uitlezen op softwareniveau een (interne) conversieslag nodig. Een hieraan gerelateerd probleem is, dat bij het uitlezen van de audiodata op een CD (meestal kleine) leesfouten optreden die op hardwareniveau niet goed te detecteren zijn. De betere audio-extractietools ondervangen dit door elke sector op een CD meerdere keren uit te lezen. Vanwege het ontbreken van een bestandssysteem, kunnen de geëxtraheerde audio-data ook niet worden opgeslagen als een ISO (of UDF) image. Het is gebruikelijk om de audio-tracks van een CD op te slaan als afzonderlijke audiobestanden. Meestal wordt hiervoor het WAV formaat gebruikt, dat in vrijwel alle media-players kan worden afgespeeld. Verder moet de afspeelvolgorde van de afzonderlijke bestanden ergens worden vastgelegd. Het is daarbij niet voldoende om dit alleen in de beschrijvende metadata te doen: je wilt namelijk ook dat een eindgebruiker de tracks in een mediaplayer in de juiste volgorde afspeelt. Hiervoor bestaan verschillende oplossingen:

1. Informatie over afspeelvolgorde vastleggen in de audiobestanden zelf. Dit kun je bereiken met een "Track Number" metadataveld als *RIFF Info Tag*[12] of *ID3 Tag*[13]. Voor het WAV formaat worden meestal *RIFF Info Tags* gebruikt, maar sommige mediaplayers herkennen alleen *ID3 Tags*.

2. De afspeelvolgorde vastleggen in een playlistbestand. Hiervoor bestaan verschillende formaten[14] zoals M3U[15] en het XSPF[16].

3. De afspeelvolgorde impliciet vastleggen via de bestandsnaam van elke track. Bijvoorbeeld:

---

[12]Link: http://www.sno.phy.queensu.ca/~phil/exiftool/TagNames/RIFF.html#Info

[13]Link: https://en.wikipedia.org/wiki/ID3

[14]Link: https://en.wikipedia.org/wiki/Playlist#Types_of_playlist_files

[15]Link: https://en.wikipedia.org/wiki/M3U

[16]Link: https://en.wikipedia.org/wiki/XML_Shareable_Playlist_Format

```
track-01.wav, track-02.wav, ...
```

Via het "Digital Preservation Q & A" forum heb ik geprobeerd te achterhalen wat collega-instellingen doen. Dit leverde uiteenlopende reacties op[17]. Waarschijnlijk is een combinatie van de hiervoor genoemde methoden de beste oplossing: bijvoorbeeld door de afspeelvolgorde in de naamgeving van de bestandsnamen vast te leggen, en daarnaast een playlist aan te maken. De playlist kan dan worden afgeleid uit de metadata (bijvoorbeeld via een XSLT transformatie die vanuit METS metadata een playlist in M3U formaat maakt).

## CD-i

In onze collectie hebben we ook nog 41 dragers in het CD-i formaat. Deze schijfjes hebben een indeling die de *Green Book* standaard [18] volgt.

### Behoud informatie op drager

Het maken van disk images van CD-i schijfjes is lastiger dan CD-ROMs en DVD's. Om te beginnen is het door de Green Book standaard gebruikte bestandssysteem niet compatible met ISO 9660 (hoewel er een CD-i Bridge format variant bestaat die dat wel is). Moderne besturingssystemen ondersteunen dit bestandssysteem ook niet: als je een CD-I in de CD-drive van een Windows PC doet, zijn de bestanden op de drager niet benaderbaar met bijvoorbeeld Windows Explorer. Verder zijn ze niet altijd goed uit te lezen met moderne hardware. Omdat het maar om een klein aantal schijfjes gaat, verdient het de aanbeveling om CD-i buiten de scope van de grootschalige workflow te houden, en later te kijken of ze handmatig verwerkt kunnen worden (bijvoorbeeld met de *ISOBuster*[19] software).

## Mixed mode CD

Ten slotte bestaan er nog CD's die zowel audio als data bevatten. Het gaat hierbij om "enhanced" CD's (bijvoorbeeld een audio CD waar ook een filmpje op staat) en "mixed mode" CDs (vooral toegepast bij computerspelletjes die audiotracks gebruiken voor geluid).

### Behoud informatie op drager

Mixed mode CD's bevatten zowel gegevens die onder een regulier bestandssysteem vallen alsook audio. Vooral in de jaren '90 werden zulke CD's vaak gemaakt voor spelletjes. Het imagen van dergelijke CD's is wel mogelijk, maar is complexer dan CD's met alleen data of audio. Eén van de problemen is dat er geen open, algemeen geaccepteerd formaat bestaat om de inhoud van zulke CD's op te slaan[20]. Hiervoor bestaan wel proprietary formaten,

---

[17]Link: http://qanda.digipres.org/1082/best-method-record-track-playing-order-for-ripped-audio-cds

[18]Link: https://en.wikipedia.org/wiki/Green_Book_%28CD_standard%29

[19]Link: http://www.isobuster.com/help/cd-i_and_vcd

[20]Link: http://anjackson.net/keeping-codes/practice/developing-a-robust-migration-workflow-for-preserving-and-curating-handheld-media.html

maar deze worden niet breed ondersteund, waardoor het onzeker is of zulke images op de lange termijn wel uitleesbaar blijven. Een artikel van Brown uit 2012 meldt dat de meeste emulatiesoftware images van dergelijke CD's niet goed kan inlezen[21]:

> It appears that none of the most widely used PC emulators (VMWare, Virtual Box, Xen, Qemu) support hybrid CD-ROMs.

Dappert et al. [22] beschrijven een nogal bewerkelijke procedure die bij de British Library is gevolgd om dergelijke dragers te imagen. Ik heb zelf ook enkele tests gedaan met de veelgebruikte *cdrdao* software[23]. Volgens de documentatie is de tool in staat om de inhoud van dergelijke CD's op te slaan in BIN/TOC formaat. In mijn tests lukte het me niet om van dergelijke CD's images te maken, zonder dat daarbij òf de audio òf de data ontbraken. Door de tool twee keer achtereenvolgend te draaien (voor 2 aparte sessies) lukte het uiteindelijk wel (maar dat levert dan 2 afzonderlijke images op). Het is ook niet duidelijk of de KB collectie zulke CD's bevat, en, zo ja, hoeveel.

[21]Link: http://www.ijdc.net/index.php/ijdc/article/view/216

[22]Zie: http://arxiv.org/pdf/1309.4932.pdf

[23]Link: http://cdrdao.sourceforge.net/

# 5. Images maken in productieomgeving

## Handmatige verwerking ondoenlijk

Het aantal optische dragers in de KB collectie is dermate groot, dat het ondoenlijk is om alle CD's en DVD's handmatig te verwerken. Als we voor het gemak uitgaan van 15000 dragers (het kunnen er ook meer zijn), en een gemiddelde tijd van 15 minuten voor het imagen van 1 drager, kom je uit op zo'n 3750 uur. Zelfs wanneer je telkens 3 CD's tegelijk op parallel draaiende machines zou verwerken (waarschijnlijk wel het maximaal haalbare dat één persoon aankan), betekent dat nog dat alleen het maken van de images al bijna een jaar in beslag gaat nemen. En dat is dan nog exclusief alle voorbereidingen en bijkomende handelingen: CD's uit het magazijn ophalen, de schijfjes uit de doosjes halen, handmatig metadata invoeren, etc. Daarnaast brengt een grotendeels op handwerk gebaseerde workflow ook een grote kans op fouten met zich mee, zeker bij de grote hoeveelheden waarmee we in onze collectie te maken hebben.

## CD / DVD robots

Voor een operationele productieomgeving is dus een belangrijke randvoorwaarde, dat voor een oplossing wordt gekozen die zoveel mogelijk geautomatiseerd is, met een minimum aan handmatig werk. Dit betekent dat voor het imagen een CD discrobot wordt gebruikt. Dat is een apparaat waar je meerdere dragers tegelijkertijd in kunt laden, waarna automatisch images van de schijfjes worden gemaakt.

### Acronova Nimbie

De British Library gebruikt hiervoor Nimbie robots van Acronova[24]. Hiermee kunnen tot maximaal 100 dragers tegelijk worden geladen; volgens een paper van de BL[25] kunnen ze ook continu draaien (d.w.z. dat de spindel *tijdens* het imagen bijgevuld kan worden met nieuwe dragers). Uit praktisch oogpunt kan dit een voordeel zijn. Hoewel de BL tevreden is met de Nimbie machines, vertelde BL's Peter May me dat ze nogal storingsgevoelig zijn. Verder is mij niet helemaal duidelijk wat de kwaliteit is van de met de machines meegeleverde (freeware) software. Het gaat me hierbij vooral om de software die gebruikt wordt voor audio CD's. Hoogwaardige extractietools zoals Exact Audio Copy[26] en Cdparanoia[27] bereiken een hoge kwaliteit door de audiodata op CD's meerdere keren uit te lexen. Het is onduidelijk hoe de met de Nimbie's meegeleverde software dit doet. Ook is in de BL workflow bij de verwerking van audio CD's een extra (post-processing) verwerkingsstap nodig voor het maken van WAV bestanden. Met een aanschafprijs van

[24]Link: http://www.acronova.com/brand/nimbie.html

[25]Zie: http://arxiv.org/pdf/1309.4932.pdf

[26]Link: http://www.exactaudiocopy.de/

[27]Link: https://www.xiph.org/paranoia/

rond de $600[28] zouden de Nimbie machines wel een goedkope oplossing zijn (zelfs als er tussentijds eens één vervangen moet worden).

### Acronova DupliQ

De DupliQ[29] is een discrobot met een laadcapaciteit van 25 dragers. Tests bij de BL brachten echter meerdere problemen aan het licht, waardoor deze machines niet geschikt zijn voor het imagen van grote hoeveelheden dragers.

### Ripstation

Een andere veelgebruikte discrobot is het *Ripstation* systeem van Formats Unlimited Inc.[30]. Deze robots hebben een laadcapaciteit van 300 dragers; afhankelijk van het model bevat de robot 2 of 4 DVD drives, waarmee volgens een in opdracht van Library of Congress gemaakt rapport[31] een verwerkingssnelheid te behalen valt van zo'n 30 tot 50 schijfjes per uur. De fabrikant zelf noemt vergelijkbare waarden voor audio CD's, maar lagere aantallen voor CD-ROMs (16 per uur) en DVD's (8 per uur). Afgaande op de documentatie op de website, is het niet mogelijk om tijdens het draaien van een batch nieuwe dragers te laden (dit in tegenstelling tot de Nimbie): je moet dus voor elke batch de machine vullen, en vervolgens wachten tot alle dragers zijn verwerkt. De robots worden geleverd met software waarmee het extractieproces aan de specieke wensen van de gebruiker kan worden aangepast. De fabrikant biedt afzonderlijke modellen aan die specifiek toegesneden zijn op respectievelijk data CD-ROMs, DVD-video en audio CD's. Hoewel de hardware in al deze gevallen identiek lijkt, is op grond van de website niet helemaal duidelijk of de bijgeleverde software een workflow ondersteunt waarbij tegelijkertijd CD-ROMs en audio CD's worden geladen. Dit is voor onze situatie wel een belangrijke vereiste[32], en vóór een eventuele aanschaf zal dit nagevraagd moeten worden. Verder is

---

[28]Link: http://store.acronova.com/dvd-autoloader-nimbie-usb-plus-nb21-dvd.html

[29]Link: http://www.acronova.com/brand/dupliq.html

[30]Link: http://www.mfdigital.com/ripstation.html

[31]De beveiliging is wel te omzeilen met de aanschaf van extra software, maar de de fabrikant van RipStation biedt hiervoor geen ondersteuning. Mogelijk is dit wettelijk ook niet toegestaan. Bron: http://www.digitizationguidelines.gov/audio-visual/documents/Preserve_DVDs_BloodReport_20140901.pdf

[32]De verwerking van audio CD's wijkt nogal af van de overige formaten. Op basis van de beschikbare documentatie is niet duidelijk of de RipStation software zodanig te configureren is dat bij van audio CD's automatisch een set WAV bestanden maakt (in plaats van een ISO image). Dit is vooral om praktische redenen van belang: bij het uit het magazijn halen van de dragers krijg je waarschijnlijk CD-ROMs, DVD's en audio CD's kriskras door elkaar heen. Als de dragers eerst op formaat gesorteerd moeten worden (of afzonderlijk uit het magazijn gehaald), levert dat extra werk op. Hierbij moet ook bedacht worden dat de informatie in de catalogus niet altijd juist is: een drager die in de catalogus beschreven wordt als audio CD, kan ik werkelijkheid een CD-ROM zijn.

mij, net als bij de Nimbie machines, niet duidelijk hoe goed de kwaliteit van de audio extractie is.

De Ripstation robots worden gebruikt bij Library of Congress[33] en de British Library[34]. Afhankelijk van het specifieke model kost zo'n machine (inclusief software) ongeveer $4000.

## Te behalen verwerkingssnelheid

Hoewel het verleidelijk is om af te gaan op door fabrikanten gepubliceerde verwerkingssnelheden, wijzen eerdere ervaringen bij de British Library uit dat juist zaken als handmatig laden van de robots en de invoer van metadata uiteindelijk de meest tijdrovende handelingen zijn bij het imagen van optische dragers[35]. Verder is het niet zeker dat in alle gevallen van de volledige laadcapaciteit van de CD-robot gebruik kan worden gemaakt: wanneer je 300 schijfjes tegelijk inlaadt, moet je ook precies vastleggen met welke titel/beschrijving elk schijfje overeenkomt. Bij zulke grote aantallen kunnen hierbij gemakkelijk fouten optreden. Voor de British Library was dit een reden om twee robots met een kleinere capaciteit te gebruiken[36]. Verder houd je altijd dragers over die fouten opleveren in de automatische workflow, en die moeten vervolgens handmatig worden verwerkt.

Bovenstaande redenen maken het lastig om een betrouwbare inschatting te maken van de benodigde hoeveelheid tijd. Het rapport van de British Library noemt een gemiddelde verwerkingssnelheid van 1050 schijfjes per maand, met 2 parallel draaiende Nimbie robots die elk met batches van enkele tientallen CD's tegelijk gevuld worden. Omdat de collectie optische dragers van de BL qua samenstelling nogal op de onze lijkt, lijkt dit voor ons ook wel een realistische schatting.

## Nog uit te zoeken

De volgende zaken moeten nog nader uitgezocht worden:

- Verwerking van mixed mode / multisesie CD's. De RipStation documentatie geeft aan dat multi-sessie DVD's niet naar een image file kunnen worden geëxtraheerd. Mogelijk betekent dit dat de verwerking van mixed mode CD's ook complicaties oplevert. De met de Nimbie robots meegeleverde software kan op zich wel overweg met multi-sessie/mixed mode CD's, maar kan het resultaat alleen wegschrijven naar een proprietary formaat waar een emulator niets mee kan.
- Kwaliteit van de bijgeleverde software voor audio-extractie.

---

[33]Link: http://blogs.loc.gov/digitalpreservation/2012/07/rescuing-the-tangible-from-the-intangible/

[34]Zie: http://arxiv.org/pdf/1309.4932.pdf

[35]Zie: http://arxiv.org/pdf/1309.4932.pdf

[36]Zie: http://arxiv.org/pdf/1309.4932.pdf

- DVD's (DVD-video) met kopieerbeveiling kunnen problemen opleveren. In de catalogus heb ik vluchtig door alle DVD-entries in gebladerd, en daaruit blijkt dat het bij veel van de video-DVD's in onze collectie om educatieve materialen gaat. Het ligt niet voor de hand dat zulke schijfjes beveiligd zijn, in tegenstelling tot DVD's die voor de consumentenmarkt zijn bedoeld (films, TV-series). Van deze laatste categorie hebben we juist heel weinig dragers in huis, dus waarschijnlijk zullen eventuele kopieerbeveiligingen ons weinig problemen opleveren[37].
- Ondersteuning van verschillende typen dragers (CD-ROMs, audio CD's) binnen één batch.
- De mate waarin de robots en de bijgeleverde software kwaliteitscontroles uitvoeren op de gemaakte images (bijvoorbeeld: checksums).

Mocht de KB serieus overwegen tot aanschaf, dan moet over bovenstaande punten eerst navraag worden gedaan bij de leveranciers. Verder is het dan een goed idee om contact op te nemen met onze collega's bij de British Library en Library of Congress.

[37]De beveiliging is wel te omzeilen met de aanschaf van extra software, maar de de fabrikant van RipStation biedt hiervoor geen ondersteuning. Mogelijk is dit wettelijk ook niet toegestaan. Bron: http://www.digitizationguidelines.gov/audio-visual/documents/Preserve_DVDs_BloodReport_20140901.pdf

# 6. Kwaliteitscontrole

In dit hoofdstuk geef ik een kort overzicht van enkele basale controles die nuttig kunnen zijn in de imaging workflow voor optische dragers, uitgesplitst naar type drager.

## CD-ROM, DVD

### Checksum

In het ideale geval zou van elke CD of DVD een checksum berekend moeten worden, en worden vergeleken met de checksum van het ISO image. De vraag is of dit haalbaar is: afgaande op enkele tests die ik op zelf mijn PC heb uitgevoerd drukt zo'n controle de performance nogal, omdat hiervoor de fysieke drager helemaal opnieuw uitgelezen moet worden. Het moet ook ondersteund worden door de discrobot.

### Controle op integriteit bestandssysteem

Verder kun je nog controleren op de integriteit van het ISO 9660 bestandssysteem. Hiervoor bestaat de *isovfy*[38] tool. De documentatie van isovfy maakt niet duidelijk op welke aspecten deze tool precies controleert. In een test die ik deed bleek *isovfy* niet in staat om enkele afgekapte images (waarbij een blok van ruim 50 MB ontbrak) te detecteren. Verder gaf *isovfy* vaak foutmeldingen die gerelateerd waren aan de fysieke drager zelf (blijkbaar omdat het bestandssysteem op bepaalde punten niet aan ISO 9660 voldeed). Hiermee lijkt de inzet van *isovfy* niet zinvol.

Voor het detecteren van afgekapte images zou je ook nog gebruik kunnen maken van informatie in de headers van het bestandssysteem zelf. Zo bevat de zgn *Primary Volume Descriptor* van een ISO 9660 bestandssysteem velden waarmee je in theorie de grootte van het hele ISO image kunt berekenen. Wanneer de werkelijke bestandsgrootte kleiner is, is dit een indicatie dat het ISO bestand niet compleet is. Als test heb ik een eenvoudig softwaretooltje geschreven dat zo'n controle uitvoert[39]. Voor eventuele inzet in een productieworkflow zou de tool op enkele details aangepast moeten worden (machine-leesbare output; ondersteuning van UDF en HSF bestandssystemen). Een dergelijke controle zou vooral nuttig zijn in geval van hardwarefalen (wat in sommige gevallen tot onvolledige images kan leiden).

Voor de duidelijkheid: bij gebruik van een checksumcontrole is het eigenlijk niet nodig om ook nog een controle op het bestandssysteem in te bouwen.

## Audio CD

De kwaliteitscontrole voor audio CD's is lastiger, omdat je hierbij geen gebruik kunt maken van checksums. Wannneer de audio als losse WAV bestanden wordt opgeslagen, zouden de

[38]Link: http://manpages.ubuntu.com/manpages/vivid/en/man1/isovfy.1.html

[39]Link: https://github.com/KBNLresearch/verifyISOSize

WAVs nog wel gevalideerd kunnen worden. Dit kan bijvoorbeeld met de WAV module van JHOVE[40].

[40]Link: http://jhove.openpreservation.org/

# 7. Metadata

Voor elk gemaakt image moet ook het nodige aan metadata worden aangemaakt en opgeslagen. Een uitputtende lijst van alle benodigde metadata valt buiten de scope van dit verkennende onderzoek. Een bruikbaar vertrekpunt is het rapport Van Dappert et. al (2011), waarin voor een vergelijkbaar project bij de British Library de volgende typen metadata[41] worden genoemd:

- **Unieke identifier voor elke drager** - de identifier koppelt een disk image aan de fysieke drager. Een mogelijke complicatie voor de situatie op de KB hierbij is onze catalogus ontsluit op het niveau van *publicaties* die dragers als bijlage hebben; de catalogus ontsluit niet op het niveau van *de individuele dragers zelf*. Een één op één koppeling met de catalogus is hierdoor niet mogelijk. Hiervoor is dus een oplossing nodig (bijvoorbeeld: uitbreiden van de PPN identifiers met een extensie, voor zover dit mogelijk is).
- **Fysieke lokatie van dragers tijdens imaging procedure en verantwoordelijke persoon**
- **Event-metadata over het imaging proces** - tijdstip waarop image is gemaakt; checksum; persoon die de image heeft gemaakt; gebruikte soft- en hardware; status van imaging proces (success / fail / partial success / etc.); namen van outputfiles (ISO / WAV).
- **Beschrijvende metadata van de fysieke drager** - kan ook via een link naar bestaand metadata register (bijvoorbeeld catalogus). Omdat de KB catalogus niet tot op dragerniveau ontsluit (zie eerdere opmerking bij unieke identifiers), is het waarschijnlijk onvermijdelijk dat deze metadata deels handmatig moeten worden ingevoerd.

Voor de situatie op de KB zal dit nog verder uitgewerkt moeten worden.

## Technische afhankelijkheden

Een eindgebruiker die een DVD of CD-ROM image in een geëmuleerde omgeving wil bekijken zal behoefte hebben aan (technische) metadata over de CD-ROM zelf (bijvoorbeeld: welk besturingssysteem is nodig om de inhoud te renderen). Deze informatie is soms (maar niet altijd!) te vinden in de één van de annotatievelden in de catalogusrecords. Een voorbeeld is het volgende catalogusrecord:

http://opc4.kb.nl/DB=1/PPN?PPN=177080930

Let op het veld “Annotatie”, met daarin info over het technische afhankelijkheden[42]:

---

[41]Zie: http://arxiv.org/pdf/1309.4932.pdf

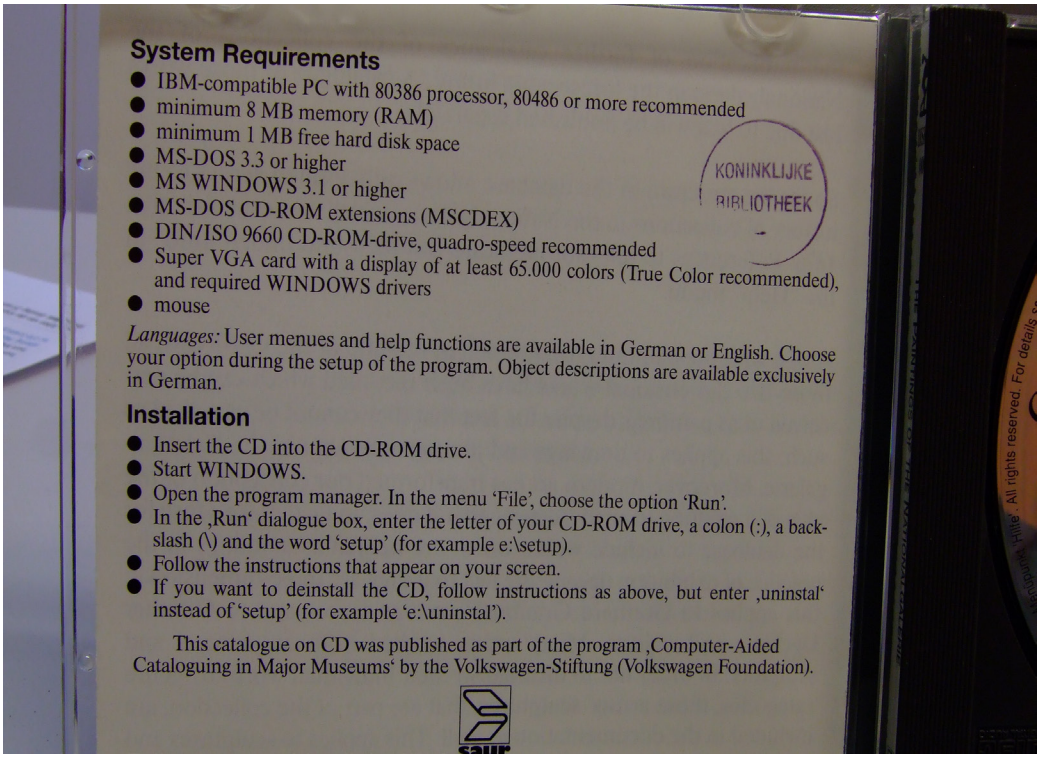
[42]Als je de onderliggende record via jrsu benadert (link: http://jsru.kb.nl/sru/sru?x-collection=GGC&operation=searchRetrieve&startRecord=1&maximumRecords=1&recordSchema=dcx&query=%22gordi%20virussen%22&xsl=http://www.kbresearch.nl/xportal/f

Systeemeisen: PC; 486 SX (Pentium aanbevolen); 8 Mb RAM; MS-Windows 3.11, 95 of 98; 1 Mb ruimte op de harde schijf; SVGA monitor (minimaal 256 kleuren); CD-ROM speler (2 speed); geluidskaart; muis

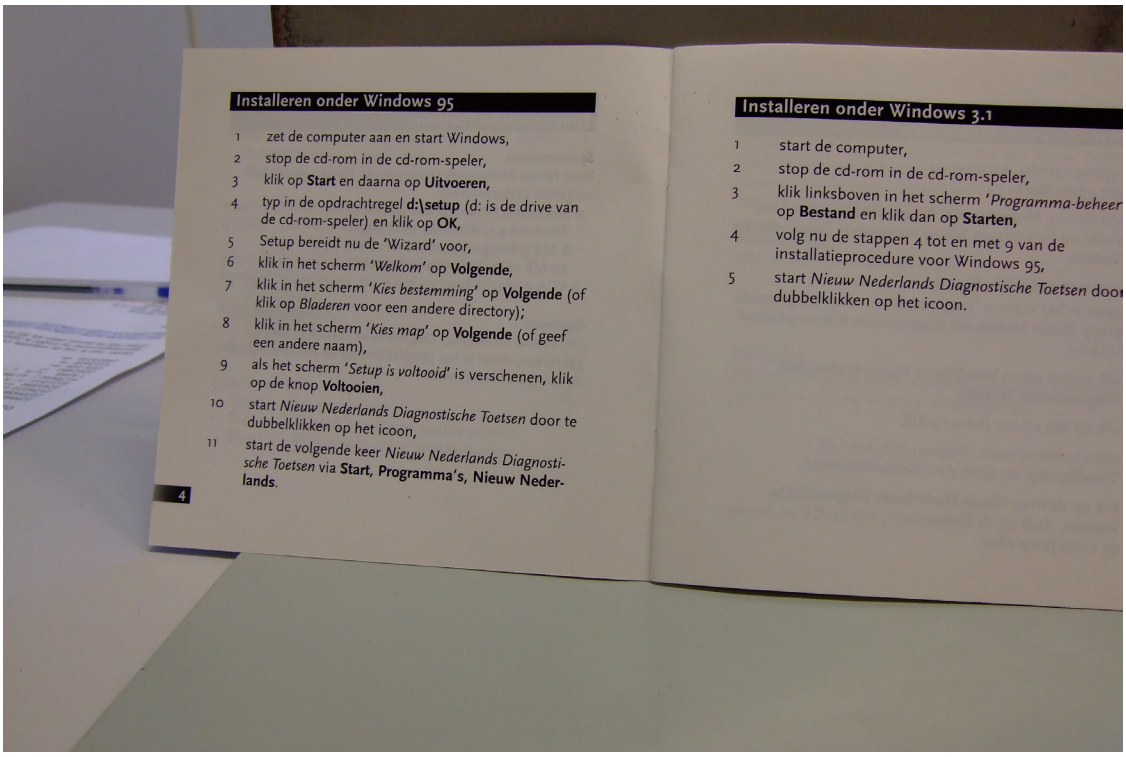
Het is dus belangrijk dat de link tussen de images en de bijbehorende catalogusrecord altijd gehandhaafd blijft.

## Verpakking / hoesjes / boekjes

In aanvulling op de vorige paragraaf: verpakkingen, hoesjes en boekjes bevatten vaak ook informatie die voor een gebruiker nuttig is. Het kan dan bijvoorbeeld gaan om gedetailleerde informatie over de vereiste technische omgeving, installatie-instructies en gebruikersdocumentatie. Een paar voorbeelden:




ull.xsl) is te zien dat deze informatie in het *dcx:annotation* veld te staan; maar er zijn meerdere van die velden. Niet alle cd-rom entries hebben dit veld.



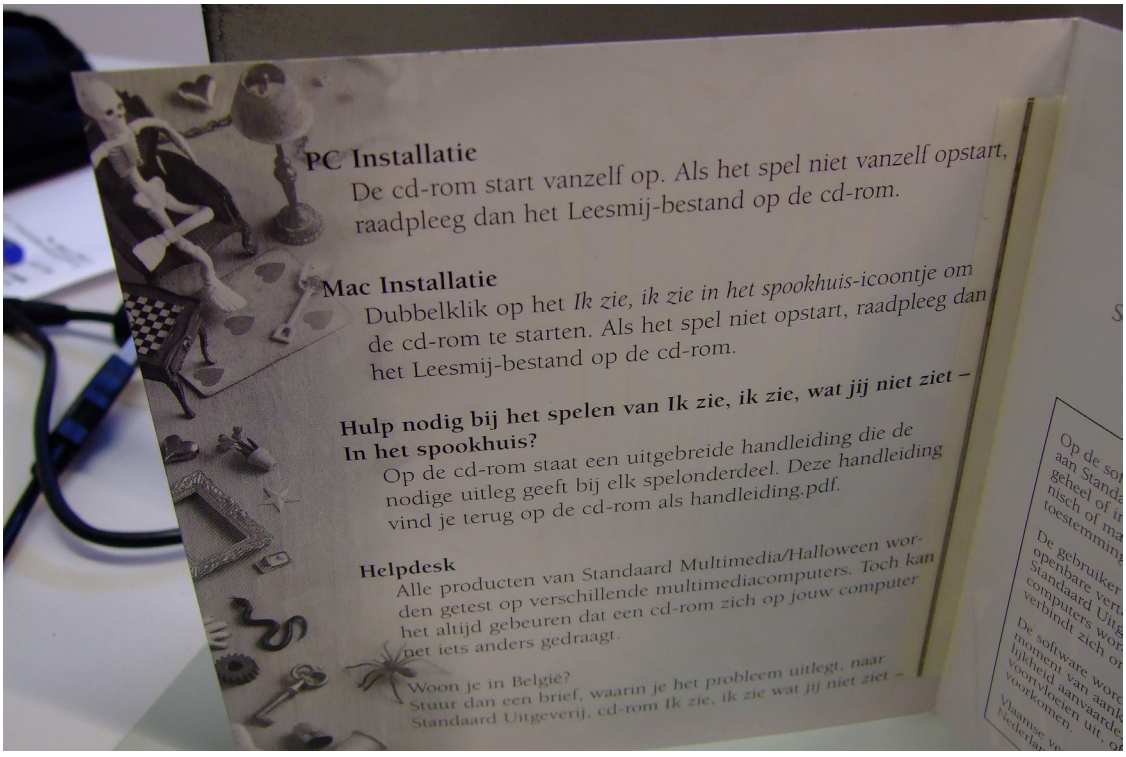
## Installeren onder Windows 95

1. zet de computer aan en start Windows,
2. stop de cd-rom in de cd-rom-speler,
3. klik op **Start** en daarna op **Uitvoeren**,
4. typ in de opdrachtregel **d:\setup** (d: is de drive van de cd-rom-speler) en klik op **OK**,
5. Setup bereidt nu de 'Wizard' voor,
6. klik in het scherm *'Welkom'* op **Volgende**,
7. klik in het scherm *'Kies bestemming'* op **Volgende** (of klik op *Bladeren* voor een andere directory);
8. klik in het scherm *'Kies map'* op **Volgende** (of geef een andere naam),
9. als het scherm *'Setup is voltooid'* is verschenen, klik op de knop **Voltooien**,
10. start *Nieuw Nederlands Diagnostische Toetsen* door te dubbelklikken op het icoon,
11. start de volgende keer *Nieuw Nederlands Diagnostische Toetsen* via **Start, Programma's, Nieuw Nederlands**.

4

## Installeren onder Windows 3.1

1. start de computer,
2. stop de cd-rom in de cd-rom-speler,
3. klik linksboven in het scherm *'Programma-beheer* op **Bestand** en klik dan op **Starten**,
4. volg nu de stappen 4 tot en met 9 van de installatieprocedure voor Windows 95,
5. start *Nieuw Nederlands Diagnostische Toetsen* doo dubbelklikken op het icoon.



## PC Installatie

De cd-rom start vanzelf op. Als het spel niet vanzelf opstart, raadpleeg dan het Leesmij-bestand op de cd-rom.

## Mac Installatie

Dubbelklik op het *Ik zie, ik zie in het spookhuis*-icoontje om de cd-rom te starten. Als het spel niet opstart, raadpleeg dan het Leesmij-bestand op de cd-rom.

**Hulp nodig bij het spelen van Ik zie, ik zie, wat jij niet ziet – In het spookhuis?**

Op de cd-rom staat een uitgebreide handleiding die de nodige uitleg geeft bij elk spelonderdeel. Deze handleiding vind je terug op de cd-rom als handleiding.pdf.

**Helpdesk**

Alle producten van Standaard Multimedia/Halloween worden getest op verschillende multimediacomputers. Toch kan het altijd gebeuren dat een cd-rom zich op jouw computer net iets anders gedraagt.

Woon je in België?
Stuur dan een brief, waarin je het probleem uitlegt, naar Standaard Uitgeverij, cd-rom Ik zie, ik zie wat jij niet ziet –

In feite zijn dit ook beschrijvende metadata, en idealiter zou dit voor een gebruiker beschikbaar moeten zijn. Maar om dit te realiseren moet je al deze materialen dan digitaliseren. Dat brengt een hoop extra werk met zich mee. Daar komt bij dat de bij de schijfjes behorende verpakkingen en boekjes nogal divers zijn (soms gaat het om uitgebreide gebruikershandleidingen), waardoor dit waarschijnlijk niet haalbaar zal zijn.

Via de "Digital Preservation Q & A" website heb ik geprobeerd erachter te komen hoe collega-instellingen hiermee omgaan. Dit leverde de volgende reacties op:

- Rhizome heeft voor een kleine collectie CD-ROMs foto's gemaakt van alle verpakkingen[43].
- Het Computer History Museum maakt ook scans van boekjes en verpakkingen, maar ook hier gaat het om (veel) kleinere aantallen dan we bij de KB hebben[44].
- De Stanford University Library heeft voor de omvangrijke "Cabinetry Collection" scans en foto's gemaakt van alle dozen, verpakkingen en boekjes[45].

## Invoer metadata en link met standaarden

In de BL workflow worden de metadata in de eerste instantie ingevoerd in een eenvoudige spreadsheet. Omdat de images uiteindelijk geïngest moeten worden in het DM, is het belangrijk dat zo'n spreadsheet zodanig wordt opgezet dat alle velden corresponderen met informatie-eenheden in het DM metadatamodel (lees: METS en PREMIS eenheden). Verder merken de auteurs van het BL rapport nog op dat van alle handelingen in de BL workflow de handmatige invoer van metadata uiteindelijk het meest tijdrovend was. De hoeveelheid werk die hiermee gepaard gaat moeten we dus niet onderschatten.

## Hybride collectie

Tot slot is het belangrijk te benadrukken dat het hier eigenlijk om een *hybride* collectie gaat met zowel een digitaal (de informatie op de drager)  alsook een fysiek aspect (verpakkingen en hoesjes, maar ook de papieren publicatie waarvan veel dragers in deze collectie bijlages zijn). De KB heeft met de verwerking van dergelijke collecties vooralsnog weinig tot geen ervaring, en ook geen eenduidige manier om een digitale component die bij een fysiek boek hoort te registeren.

---

[43]Link: http://qanda.digipres.org/1079/rom-dvd-imaging-customary-save-scans-booklets-covers-as-well; zie reactie Dragan Espenschied.

[44]Link: https://twitter.com/bitsgalore/status/646305395779207168; zie reactie Andrew Berger.

[45]Link: http://www.slideshare.net/charthai/what-the-hell-is-it-and-what-should-i-do-with-it-cataloging-challenging-collections; zie slide 12.

# 8. Toegang en gebruik images

De manier waarop de gemaakte images gebruikt kunnen worden is afhankelijk van de aard van de oorspronkelijke drager.

## CD-ROM / DVD-ROM: emulatie

Veel CD-ROMs en DVD-ROMs in de KB collectie bevatten software die bedoeld is voor gebruik binnen bepaalde hard- en softwareomgevingen. In veel gevallen gaat het dan om "oude" omgevingen die nu niet meer algemeen gebruikt worden, zoals MS DOS, Windows 95, enzovoort. De beste toegangsstrategie is dan emulatie. Hierbij wordt de oorspronkelijke omgeving (bijvoorbeeld Windows 95) nagebootst met emulatiesoftware die op een "moderne" machine draait. Sommige CD-ROMs bevatten alleen losse bestanden (bijvoorbeeld PDF's). In zo'n geval is emulatie niet nodig, maar het is dan wel belangrijk dat op de PC waarop een gebruiker een image benadert software aanwezig is waarmee ISO images geopend kunnen worden (bijvoorbeeld een datacompressietool als *7-Zip*[46]).

### Emulatie versus virtualisatie

In principe bootst een emulator de volledige hardware na waarop het oude besturingssysteem ontworpen is. Veel oude besturingssystemen (bijvoorbeeld MS-DOS en oude Windowsversies) kunnen ook op moderne hardware draaien, en dan is pure emulatie eigenlijk niet nodig. Voor zulke gevallen wordt vaak gebruikgemaakt van *virtualisatie*, waarbij maar een deel van de hardware nagebootst wordt. Veelgebruikte softwarepakketten hiervoor zijn Oracle VirtualBox[47] en VMWare[48]. Virtualisatie heeft als voordeel dat het sneller is dan volledige emulatie. In de praktijk worden beide termen vaak door elkaar heen gebruikt. Daar komt bij dat sommige emulatiesoftware ook in virtualisatiemodus kan draaien (bijvoorbeeld de populaire QEmu emulator[49]). Voor de leesbaarheid gebruik ik in wat nu volgt telkens de term "emulatie".

## DVD Video

DVD-Video images kunnen geopend worden in standaard mediaplayer software. Zo kan de open-source VLC Media PLayer [50] direct ISO images van DVD's openen; een image is hierbij op exact dezelfde manier te gebruiken als de fysieke drager.

## Audio CD

Audio CD's waarvan de individuele tracks als WAV bestand zijn opgeslagen kunnen door de meeste mediaplayers worden afgespeeld. Eventueel zou daarbij de inhoud van elke CD als

---

[46]Link: http://www.7-zip.org/

[47]Link: https://www.virtualbox.org/

[48]Link: http://www.vmware.com/

[49]Link: http://wiki.qemu.org/

[50]Link: http://www.videolan.org/vlc/

playlist kunnen worden opgeslagen (dit is een bestand waarin alle tracks die deel uitmaken van de CD, en de volgorde waarin ze worden afgespeeld zijn gedefinieerd). Een veelgebruikt formaat hiervoor is het XML Shareable Playlist formaat[51].

## CD-i

Voor het CD-i formaat bestaat een emulator, maar het gaat hierbij om closed-source software die sinds 2012 niet meer actief ontwikkeld lijkt te worden[52]. Verder heeft de emulator een image nodig van een systeem CD-i ROM chip. Op deze ROM chips rust copyright, waardoor de images niet verspreid mogen worden. Consequentie is dat om de emulator te kunnen gebruiken, je eerst een werkende (fysieke) CD-i speler nodighebt (bij de emulator zit software waarmee je een image van de chip kunt maken). Al met al betekent dit dat voor de emulatie van CD-i schijfjes nogal wat hobbels te nemen zijn. Vanwege het beperkte aantal CD-i's in de KB collectie heb ik dit in dit onderzoek niet verder uitgewerkt.

## Steekproef dragers depot

Als test heb ik een beperkt (27) aantal dragers uit de depotcollectie laten ophalen, en deze geprobeerd te imagen. Door op steekwoorden te zoeken heb ik geprobeerd om hierbij specifieke gevallen in de selectie te krijgen, zoals (zelfgebrande) recordable CD's, en CD's met software die onder oude DOS- of Windowsversies draait. Uiteindelijke resulteerde dit in de volgende selectie:

- 15 professioneel geproduceerde (glass-mastered) CD-ROMs
- 3 recordable (gebrande) CD-ROMs
- 4 DVD-ROMs
- 2 Video DVD's
- 1 audio CD
- 2 CD-i schijfjes

Hierbij moet de aantekening worden gemaakt dat een aanzienlijk deel van de dragers die in de catalogus staan vermeld, in de praktijk niet opvraagbaar zijn omdat ze geen aanvraagnummer hebben (voorbeeld: http://opc4.kb.nl/DB=1/PPN?PPN=154089893). Het is niet helemaal duidelijk wat de achterliggende reden hiervoor is. Opvallend is dat vooral dragers met spelletjes vaak niet opvraagbaar zijn. Omdat juist CD-ROMs met spelletjes vaak problematisch kunnen zijn (bijvoorbeeld door het gebruik van mixed-mode CD layouts) betekent dit wel dat deze subcategorie grotendeels buiten deze tests zijn gebleven.

[51]Link: https://en.wikipedia.org/wiki/XML_Shareable_Playlist_Format

[52]Link: http://www.cdiemu.org/

## Imaging procedure

Van alle dragers heb ik met behulp van veelgebruikte command-line softwaretools images geprobeerd te maken. Voor CD-ROMs en DVD's heb ik de *readom* tool gebruikt[53]; voor de audio CD *cdparanoia*[54]. Alle dragers zijn geimaged op een onder Linux draaiende ontwikkel-laptop met een externe (USB) DVD lezer. Dragers die op deze machine niet leesbaar bleken te zijn heb ik ook nog geprobeerd te imagen met mijn reguliere Windows PC. De tabel in Annex A geeft een overzicht van alle dragers in de steekproef, waarbij de meest rechtse kolom het resultaat van de imaging procedure laat zien.

Het meest in het oog springende resultaat is dat *alle* recordable CD-ROMs problemen opleverden: twee waren volledig onleesbaar, terwijl de derde alleen op de Windows machine uitleesbaar was (overigens ging het hier om en CD-ROM waar alleen één PDF bestand op stond). Vanwege het kleine aantal kan dit resultaat niet zonder meer als representatief worden verondersteld voor al onze recordable CD-ROMs. Desalniettemin valt te verwachten dat veel dragers binnen deze sub-categorie intussen onleesbaar zijn. Verder was één professioneel geproduceerde CD-ROM uit de testset onleesbaar. De *readom* en *cdparanoia* tools bleken beide niet overweg te kunnen met de CD-i schijfjes. Het bestandssysteem op deze dragers wordt ook niet herkend door moderne besturingssystemen, waardoor de individuele bestanden in bijvoorbeeld Windows Explorer of een Linux file manager niet benaderbaar zijn. Uiteindelijk lukte het me om deze dragers te imagen met de *cdrdao*[55] tool als BIN images[56]. Of deze images ook daadwerkelijk alle informatie van de drager bevatten is onduidelijk. Dit is een extra om de CD-i's buiten de bulkworkflow te houden.

## Proof-of-concept emulatie

Met behulp van de virtualisatiesoftware Oracle VitualBox[57] heb ik een aantal virtuele machines gemaakt met daarin "oude" toegangsomgevingen. Het gaat hierbij om:

- MS-DOS 6.2
- Windows 3.11
- Windows 95
- Windows ME

[53]Link: http://linux.die.net/man/1/readom

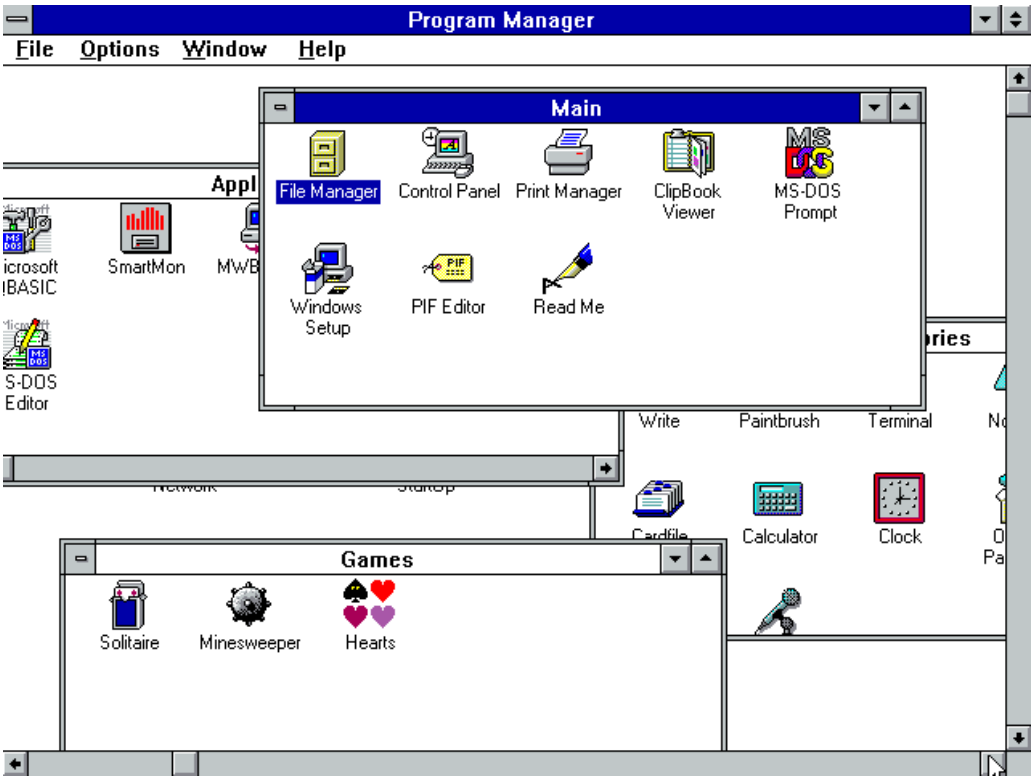
[54]Link: http://linux.die.net/man/1/cdparanoia

[55]Link: http://linux.die.net/man/1/cdrdao

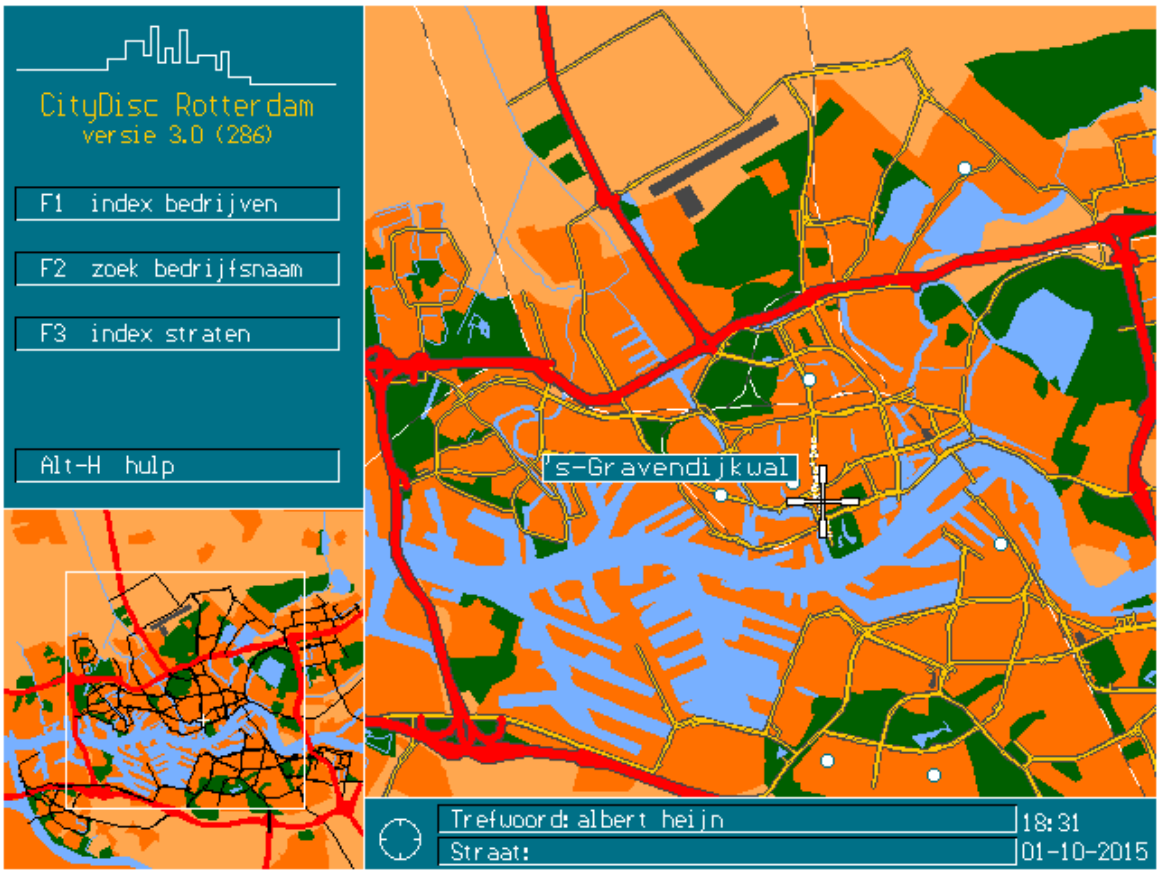
[56]Link: http://fileformats.archiveteam.org/wiki/CUE_and_BIN

[57]Link: https://www.virtualbox.org/

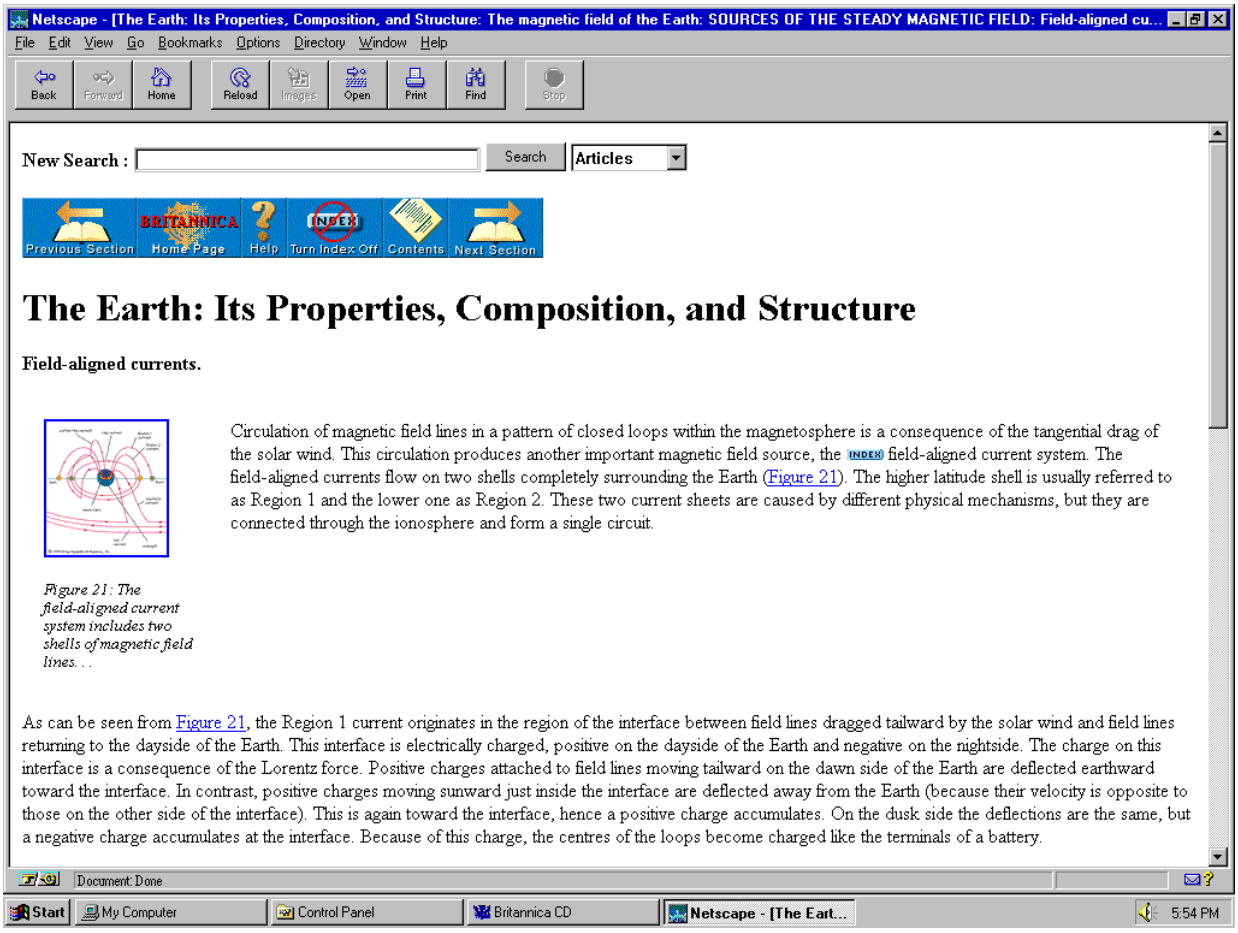
Als voorbeeld hieronder een screenshot van Windows 3.11, draaiend op een virtuele machine in VirtualBox:



Vervolgens heb ik deze virtuele machines gebruikt voor het renderen van oude CD-ROMs. Als voorbeeld is hier een screenshot van een interactieve stadsplattegrond van Rotterdam op CD-ROM uit 1995. Deze MS-DOS applicatie draait hier op een virtuele machine onder MS-DOS 6:



En hier een pagina uit de Encyclopædia Britannica CD-ROM uit 1997. Deze CD-ROM draait onder Windows 95, en gebruikt Netscape Navigator 2:




## Belang behoud oude software

Een randvoorwaarde voor het gebruik van emulatie is dat er kopieën beschikbaar zijn van originele besturingssystemen en software. Om bijvoorbeeld een machine met daarop Windows 3.11 te emuleren heb ik de originele installatiediskettes nodig (of images daarvan). Voor modernere Windows versies staan de installers op CD-ROM; in sommige gevallen (bijvoorbeeld Windows 95) is naast de installatie-CD ook nog een opstartdiskette nodig. Vaak zijn in aanvulling hierop ook hardwaredrivers nodig. Zo heb je om in MS-DOS CD-ROMs te kunnen lezen drivers nodig die niet bij het besturingssysteem zijn meegeleverd. De KB heeft tot op heden oude software nooit structureel bewaard (bijvoorbeeld door het imagen van installatiediskettes en -CD's). Als de KB serieus werk wil maken van emulatie, zal het behoud van software in de toekomst ook (veel) beter moeten worden geregeld. Enkele suggesties hierbij:

- Nieuwe installatie CD's en DVD's die bij IT Beheer binnenkomen na ontvangst direct imagen, en eventuele activeringscodes vastleggen als metadata.

- Van de de belangrijkste “oude” installatiediskettes en CD’s zo snel mogelijk images maken.
- Images vervolgens duurzaam opslaan (uiteindelijk in Digitaal Magazijn; op de kortere termijn bijvoorbeeld op de migratieserver).

Besturingssystemen, hardwaredrivers en Office software (tekstverwerkers, presentatiesoftware en spreadsheets) zouden hierbij de hoogste prioriteit moeten krijgen. belangrijke software die we nu niet meer hebben kan zo nodig tweedehands worden aangekocht (bijvoorbeeld via eBay).

## Emulatie in de leeszalen

Om onze gebruikers in de leeszalen toegang te kunnen bieden tot de (virtuele) CD-ROM collectie, zou de *Emulation as a Service* (EaaS) benadering van de Universiteit Freiburg een interessante optie kunnen zijn [58]. Tot dusver was *EaaS* vooral een *dienst* die op een server in Freiburg draaide, maar sinds kort is het ook mogelijk om de software lokaal te draaien (bijvoorbeeld op een server binnen de KB) [59],[60]. EaaS is op dit moment (oktober 2015) nog niet helemaal volwassen genoeg voor operationeel gebruik, maar de ontwikkeling van de software gaat snel, en het valt te verwachten dat gebruik in de leeszalen op korte termijn haalbaar wordt.

[58]Link:http://eaas.uni-freiburg.de/

[59]Link: http://openpreservation.org/event/from-the-toolbox-emulation-in-library-reading-rooms-tools-workflows-and-demo/

[60]Link: http://bw-fla.uni-freiburg.de/wordpress/?p=844

# 9. Aanbevelingen

## Vervolgstappen behoud dragers

### Stap 1: opstellen beleid

Om te beginnen is het noodzakelijk dat Collectiebehoud beleid opstelt (*Preservation Policy*) ten aanzien van deze collectie dragers. In het document "Kavels en waarde" (augustus 2014) heeft Collectiebehoud aan het kavel waarbinnen deze collectie valt lage waarden toegekend[61]. De vraag is of dit terecht is, of dat de beoordeling vooral voortvloeit uit onbekendheid met deze collectie (die o.a. veel educatieve materialen omvat). De Preservation Policy dient ook duidelijkheid te scheppen over *wat* behouden dient te blijven (bijvoorbeeld: alleen de inhoud van de dragers, of ook de verpakkingen, hoesjes en boekjes). Vervolgens moet een plan worden opgesteld (*Preservation Plan*) waarin beschreven staat hoe de dragers gered kunnen worden. Dit rapport is hierbij een eerste aanzet.

### Stap 2: vervolgonderzoek

Vervolgens moet de procedure voor het behoud van optische dragers in verder detail uitgewerkt worden. Uitgangspunt hierbij is dat van CD-ROMs en DVD's images worden gemaakt in ISO 9660 formaat, en dat audio CD's worden "geript" naar WAV bestanden (CD-I's blijven buiten de scope van de geautomatiseerde workflow[62]). Dit onderzoek zal zich specifiek richten op de volgende aspecten:

- De hard- en softwareconfiguratie die nodig is voor een zoveel mogelijk geautomatiseerde workflow. Concreet: welke discrobot is in ons geval het beste, en in hoeverre voldoet de standaard software die daarop draait? Dit is vooral een kwestie van informeren bij de producenten van de Nimbie en Ripstation systemen. Daarnaast kunnen we hierbij gebruikmaken van de ervaring van collega-instellingen (met name de British Library en Library of Congress).
- Nader bepalen van de metadata die bij elk image worden opgeslagen, en de vorm waarin dat het beste kan (in samenspraak met Documentverwerking en Collectiebehoud). Uitgangspunt hierbij is dat de ingest van images en metadata in het DM later zo eenvoudig mogelijk wordt.

### Stap 3: verkennen discrobot

De volgende stap bestaat dan uit het verkennen en uittesten van de discrobot. Het gaat dan vooral om de volgende aspecten:

- Hoe werken de machine en de software globaal?
- Hoe zorg je ervoor dat de gemaakte images gekoppeld zijn aan de juiste metadata?

---

[61] Link: <interne link verwijderd in externe versie>

[62] Deze zouden eventueel in een afzonderlijk project handmatig kunnen worden verwerkt.

- Hoe verwerk je verschillende typen dragers binnen één batch?
- Wat gebeurt er als een drager fouten oplevert?
- Eventuele aanpassingen aan de softwareconfiguratie.

### Stap 4: opzetten operationele workflow

Vervolgens kan de operationele workflow worden opgezet. Dit brengt waarschijnlijk nog een beperkte hoeveelheid ontwikkelwerk met zich mee (bv scripts voor automatische kwaliteitscontroles, en het genereren van technische metadata zoals checksums). In principe is dit niet de taak van de afdeling Onderzoek, maar Onderzoek kan hierbij wel een begeleidende rol spelen, en adviseren over bijvoorbeeld tools voor de kwaliteitscontrole.

### Stap 4: operationele fase

Vóórdat de operationele workflow daadwerkelijk in gebruik wordt genomen, is het belangrijk dat de medewerker(s) die hierbij aan de knoppen zitten een training krijgen. Het gaat hierbij vooral om het tijdig herkennen van probleemgevallen, en het op de juiste manier daarop reageren. Deze training zou ook bij voorkeur door Onderzoek verzorgd moeten worden.

Te slotte is het belangrijk dat de voorbereidende en operationele fasen niet volledig van elkaar gescheiden kunnen worden. De ervaring van de British Library bij een soortgelijk project was dat ze tijdens het productiedraaien regelmatig onverwachte dingen tegenkwamen, waardoor de workflow aangepast moest worden. Het is dus een iteratief proces. Een consequentie hiervan is dat ook tijdens de operationele fase af en toe inzet nodig is vanuit Onderzoek.

### Randvoorwaarden

- Bereidheid bij Collectiebehoud tot het aanschaffen van de benodigde hard- en software. Het gaat hierbij voornamelijk om discrobots, en voldoende redundante opslag om de images veilig op te kunnen slaan. Een grove schatting hierbij: uitgaande van 15.000 dragers en een gemiddelde grootte van 600 MB per drager[63] kom je dan uit op bijna 9 TB.

- Beschikbare personele inzet bij Onderzoek (JvdK), Collectiebehoud, Documentverwerking en ICT Ontwikkeling bij vervolgonderzoek en ontwerp / testen van de workflow.
- Beschikbare personele inzet tijdens operationele fase (ophalen dragers uit magazijn, terugzetten, vullen, leeghalen en bedienen van de discrobots, handmatige invoer metadata, monitoren imagingproces, afhandelen problematische dragers).

[63]De maximale opslagcapaciteit van een CD-ROM is ruim 700 MB; voor een single-layer DVD is dit ongeveer 8 GB.

## Vervolgstappen toegang op lange termijn

Als we de disc images op termijn door middel van emulatie aan onze gebruikers in de leeszalen willen aanbieden, is hiervoor ook aanvullend werk nodig. De *Emulation as a Service* (EaaS) benadering van de Universiteit Freiburg lijkt hierbij een interessante optie. Een logische vervolgstap is om deze software eens uit te testen met een lokale installatie op de KB.

Voor emulatie is de beschikbaarheid van oude legacy software van essentieel belang. Het gaat hierbij om oude besturingssystemen (bijvoorbeeld MS DOS, Windows 95), hardwaredrivers en Office software. We hebben hiervan nog een aantal dozen met oude installatiediskettes en -CD-ROMs. Deze dragers zijn ook vergankelijk; daarnaast worden bijvoorbeeld diskettes niet meer door moderne computers ondersteund. Het verdient daarom de aanbeveling om van de belangrijkste software-installers die we hebben op korte termijn disk images te maken, en deze op een veilige plek duurzaam op te slaan. Onderzoek heeft hier al een beginnetje mee gemaakt, en het plan is om deze activiteit in 2016 verder uit te breiden.

# Annex A: testdragers uit KB collectie

| Titel | Type | OS* | Imaging resultaat |
|---|---|---|---|
| Birds of tropical Asia : sounds and sights | CD-ROM | Windows 95, 98, ME, 2000, XP | OK |
| (Bijna) alles over bestandsformaten | CD-ROM | - | OK |
| FDC 1.2 : a simulink toolbox for flight dynamics and control analysis | CD-ROM (recordable) | - | Niet leesbaar (zowel op ontwikkel PC als Windows PC). Software + documentatie wel online nog toegankelijk via http://dutchroll.sourceforge.net/fdc.html. |
| Suske en Wiske stripmaker | CD-ROM | Windows 95 of hoger | OK |
| Ik zie, ik zie, wat jij niet ziet in het spookhuis | CD-ROM | Microsoft Windows 3.1x of Windows 95 | OK |
| Kommunikation mit Menschen einer nicht-schriftlichen Kultur | CD-ROM (recordable) | Windows | Niet leesbaar op ontwikkel PC; wel op Windows PC |
| De bewaarmachine | CD-ROM | Windows 95/3.1.; Apple Macintosh (minimaal 68030) systeem 7.1 of hoger | OK |
| Citydisc : 21 steden op CD-ROM | CD-ROM | MS-DOS 5.0 (of hoger) | OK |
| Die Gemälde der Nationalgalerie | CD-ROM | - | Niet leesbaar (zowel op ontwikkel PC als Windows PC) |

| Titel | Type | OS* | Imaging resultaat |
|---|---|---|---|
| Encyclopædia Britannica | CD-ROM | MS Windows 3.1, Windows for Workgroups v3.11, Windows 95; Windows NT | OK |
| De mening van Cor Galis | Audio CD | - | OK |
| Besturingssystemen | DVD-ROM | - | OK |
| Der totalitäre Staat | DVD | - | OK |
| Atlas van het gemeentelijk waterbeleid | CD-ROM (recordable) | - | Niet leesbaar (zowel op ontwikkel PC als Windows PC) |
| Fotoview Den Haag | CD-ROM | MS-Dos 5.0 en hoger | OK |
| CD recensierom : recensies over CD-ROMS | CD-ROM | Windows 3.1 of hoger | OK |
| Nieuw Nederlands. Bovenbouw. Diagnostische toetsen | CD-ROM | Windows 3.1, Windows 95 of hoger | OK |
| De feestcommissie in Marokko | DVD-video | - | OK |
| Algemene toxicologie 2 | CD-i | - | BIN file met cdrdao - OK |
| Omgaan met steriele medische hulpmiddelen | CD-i | - | BIN file met cdrdao - OK |

| Titel | Type | OS* | Imaging resultaat |
|---|---|---|---|
| Fotografische atlas van de praktische anatomie | DVD | - | OK |
| Sprekend verleden. SV-digitaal | DVD | Windows 98, ME, NT (+ Internet Explorer, Windows Media Player, Adobe Acrobat, DirectX 8, Microsoft Word) | OK |
| Klikbijbel | CD-ROM | Windows 95 of hoger | OK |
| De geologie van de provincie Utrecht | CD-ROM | - | OK |
| Bas gaat digi-taal | CD-ROM | Windows 95 | OK |
| De junior Bos(a)tlas | CD-ROM | Windows 3.1, 3.11, 95 | OK |
| Kamp Vught in de klas | DVD-Video | - | OK |