Data Archiving and Networked Services

DANS

Preferred formats

September 2015, versie 3.0

>>>

Preferred formats

Data Archiving and Networked Services (DANS)
Postbus 93067 | 2509 AB Den Haag
Anna van Saksenlaan 51 | 2593 HW Den Haag
+31 70 349 44 50
info@dans.knaw.nl | dans.knaw.nl

Inhoudsopgave

1. Selecteren van bestandsformaten

Digitale gegevens worden opgeslagen in bestandsformaten. Vaak wordt gebruik gemaakt van standaardformaten van software. Meestal zal de keuze voor de software en daarmee het bestandsformaat afhankelijk zijn van het primaire doel van de gebruiker.

Zo zal bijvoorbeeld voor het maken van een tabel sneller gebruik gemaakt worden van software voor spreadsheets dan een tekstverwerkingsprogramma. Dit komt omdat een datatabel bepaalde eigenschappen vereist, die door de gespecialiseerde software beter wordt ondersteund. Daarbij kan gedacht worden aan de mogelijkheid om gegevens te sorteren, om formules te gebruiken, om een filter op de tabel te kunnen zetten, enzovoorts. Als dergelijke informatie wordt opgeslagen vanuit een spreadsheet-programma mag de gebruiker verwachten dat het bestandsformaat deze potentieel belangrijke eigenschappen ('significant characteristics') behoudt. Mocht de tabel in een tekstverwerkingsprogramma zijn gemaakt, dan is het minder waarschijnlijk dat
de software deze eigenschappen ondersteunt. Het tekstverwerkingsprogramma is daarentegen geschikt voor het opmaken van een artikel, met bijvoorbeeld gebruik van een functionele inhoudsopgave en paginanummers. Deze eigenschappen zal het spreadsheet-programma niet ondersteunen.

Wanneer informatie uit een toegepast programma wordt opgeslagen, gebeurt dit meestal in een standaard bestandsformaat van deze software. Dit geeft echter niet de garantie dat de inhoud van het bestand in de toekomst op dezelfde wijze gebruikt of weergegeven kan worden zoals bedoeld toen het bestand werd gemaakt. Formaten kunnen bijvoorbeeld afhankelijk zijn van ondersteuning door bepaalde software. Software kan buiten gebruik raken, of slechts bepaalde versies van formaten ondersteunen. Het is ook mogelijk dat specifieke eigenschappen van formaten alleen werken in de gebruikte software, of zelfs enkel in een bepaalde versie van deze software. Ook kunnen bestanden afhankelijk zijn van het gebruik van dure of exclusieve software waar niet iedereen zomaar toegang tot kan krijgen.

Om het risico van bestandsveroudering voor te zijn en om de toegankelijkheid en de duurzaamheid van de belangrijke eigenschappen van de bestanden te waarborgen, kan een aantal voorzorgsmaatregelen worden genomen. Een van die maatregelen is om bestandsformaten te gebruiken die een hoge kans hebben om vele jaren bruikbaar te blijven.

Als algemene richtlijn stelt DANS dat de bestandsformaten die het beste geschikt zijn voor duurzaamheid en toegankelijkheid op de lange termijn:

- Veel worden gebruikt
- Open specificaties hebben
- Onafhankelijk zijn van specifieke software, ontwikkelaars of leveranciers

In de praktijk blijkt het niet altijd mogelijk om formaten te gebruiken die voldoen aan al deze kenmerken.

2. Twee categorieën bestandsformaten: preferred en acceptable formats

DANS hanteert twee categorieen bestandsformaten:

- Preferred formats zijn de bestandsformaten waarvan DANS het vertrouwen heeft dat deze op de langere termijn de beste garanties bieden qua bruikbaarheid, toegankelijkheid en duurzaamheid. Het deponeren van onderzoeksdata in preferred formats zal zonder meer door DANS worden geaccepteerd.
- Acceptable formats zijn bestandsformaten die naast de preferred formats veel worden gebruikt; waar matige tot redelijke scores aan verbonden kunnen worden voor wat betreft de bruikbaarheid, toegankelijkheid en robuustheid op de lange termijn. De voorkeur van DANS ligt bij het gebruik van preferred formats, maar het gebruik van acceptable formats zal in de meeste gevallen ook worden toegestaan in het archief.

DANS beveelt deponeerders van data sterk aan om hun gegevens aan te leveren in het preferred format zoals bij het type data in de lijst is genoemd.

Indien uw data is opgeslagen in andere bestandsformaten dan in onderstaande lijst zijn vermeld, neem dan contact op met DANS via info@dans.knaw.nl.

2.1 Overzichtstabel

Deze tabel geeft een beknopt overzicht van de DANS preferred en acceptable formats. Raadpleeg de tekst onder de tabel voor een nadere uitleg per type data. Voor het overzicht zijn de extensies ingedeeld in verschillende typen van data, onderscheiden op basis van het primaire gebruik. Hier zit veel overlap in: zo kunnen de 'preferred formats' van de meeste typen data in feite gezien worden als platte tekstbestanden of opmaaktaal.

Achter in dit document staat een verklaring van alle gebruikte afkortingen.

Deze lijst kan mettertijd veranderen als gevolg van de ontwikkeling van nieuwe bestandsformaten en het in onbruik raken van andere formaten.

§	Type	Preferred format(s)	Acceptable format(s)
2.2	Tekst documenten	• PDF/A (.pdf)	• ODT (.odt) • MS Word (.doc, .docx) • RTF (.rtf) • PDF (.pdf)
2.3	Platte tekst	• Unicode text (.txt)	• Non-Unicode text (.txt)
2.4	Opmaaktaal	• XML (.xml) • HTML (.html; .xhtml) NB: indien valide en compleet (zie uitleg) indien benodigd: • gerelateerde bestanden: .css; .xslt; .js, .es (zie uitleg)	• SGML (.sgml)
2.5	Spreadsheets	• ODS (.ods) • CSV (.csv)	• MS Excel (.xls, .xlsx) • PDF/A (.pdf) • OOXML (.docx, .docm)

2.6	Databases	• SQL (.sql) • SIARD (.siard) • tabellen uit DB (.csv)	• MS Access (.mdb, .accdb), (versie 2000 of later) • dBase (.dbf) (versie 7 of later) • HDF5 (.hdf5, .he5, .h5)
2.7	Statistische data	• SPSS Portable (.por) • SPSS (.sav) • STATA (.dta) • DDI (.xml) • data (.csv) + setup (.txt)	• SAS (.7bdat; .sd2; .tpt) • R(*)
2.8	Afbeeldingen (raster)	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff) • PNG (.png) • JPEG 2000 (.jp2)	• DICOM (.dcm) (in overleg)
2.9	Afbeeldingen (vector)	• SVG (.svg)	• Illustrator (.ai) • EPS (.eps)
2.10	Audio	• WAVE; BWF (.wav) • FLAC (.flac)	• AIFF (.aif, .aiff) • MP3 (.mp3) • AAC (.aac, .m4a)
2.11	Video	• MPEG-2 (.mpg, .mpeg, …) • MPEG-4 H.264 (.mp4) • Lossless AVI (.avi) • QuickTime (.mov)	• MKV (.mkv)
2.12	Computer Aided Design (CAD)	• AutoCAD DXF versie R12 (.dxf)	• AutoCAD andere versies (.dwg, .dxf)
2.13	Geografische Informatie (GIS)	• GML (.gml) • MIF/MID (.mif/.mid)	• ESRI Shapefiles (.shp en bijbehorende bestanden) • MapInfo (.tab en bijbehorende bestanden) • KML (.kml)
2.14	Afbeeldingen (georeferentie)	• GeoTIFF (.tif, .tiff)	• TIFF World File (.tfw en .tif)
2.15	Raster GIS	• ASCII GRID (.asc, .txt)	• ESRI GRID (.grd en bijbehorende bestanden)
2.16	3D	• WaveFront Object (.obj) • X3D (.x3d)	• COLLADA (.dae) • Autodesk FBX (.fbx)
2.17	RDF	• W3C standaarden	
2.18	Computer Assisted Qualitative Data Analysis (CAQDAS)	Formaten gebruikt in de applicatie, behandeld naar type data per afzonderlijk bestand	• Export-formaten van de gebruikte applicatie (ATLAS.TI *copy bundle*; NVIVO *export project; …*) • QuDEX

(*) In onderzoek

In dit document wordt per type data een kort overzicht gegeven van de keuze voor het preferred format, van het gebruik van de data en van eventuele conversiemogelijkheden.

Dit document is bedoeld als leidraad voor deponeerders van data. Het is verre van de enige lijst van aanbevelingen over bestandsformaten in de wereld. Er bestaan diverse bronnen en wiki's over formaten en risico's. Bij DANS zijn meerdere bestaande documenten geëvalueerd aan de hand van eigen ervaringen omtrent de formaten waar DANS mee in aanraking is gekomen.
Bij de tot stand koming van dit document heeft DANS onder meer gebruik gemaakt van de volgende bronnen:

- http://guides.archaeologydataservice.ac.uk/g2gp

- http://www.digitalpreservation.gov/formats/index.shtml
- http://www.loc.gov/preservation/resources/rfs/index.html
- https://www.archivematica.org/wiki/Significant_characteristics

Dit document is dynamisch: een werkgroep binnen DANS draagt zorg voor de monitoring van bestandsformaten en actualiseert de adviezen op basis van nieuwe ontwikkelingen.

2.2 Tekst documenten

PDF, het 'Portable Document Format' dat wordt ontwikkeld door softwaregigant Adobe, kent het subtype PDF/A dat is ontworpen voor duurzaamheid op de lange termijn. PDF/A wordt internationaal aangehouden als de standaard voor (opgemaakte) tekst documenten. Een PDF/A is een op zichzelf staand document: alle lettertypen en afbeeldingen zijn in het bestand opgenomen, zodat het niet afhankelijk is van andere bestanden op de computer om de inhoud correct weer te geven.

PDF/A kent een aantal subtypen. Het subtype PDF/A-1a is aan te raden voor tekstdocumenten die volledig met de computer zijn gemaakt ('born-digital'). Voor gedigitaliseerde documenten is het subtype PDF/A-1b geschikt.

De 'Adobe Reader' is gratis te downloaden, maar op veel computers zal al software zijn geïnstalleerd waarmee PDF bestanden geopend kunnen worden. Adobe-software voor het aanmaken van PDF-bestanden is niet gratis, maar diverse gratis softwarepakketten als OpenOffice en IrfanView bieden ook PDF-ondersteuning. Ook bestaan er print-programma's waarmee documenten naar een PDF document kunnen worden 'geprint', bijvoorbeeld de gratis Bullzip PDF printer.

Voor het maken van een PDF-bestand moeten de standaardinstellingen worden aangepast om het juiste type PDF/A te genereren.

2.3 Platte tekst

Platte tekstbestanden hebben vaak de extensie TXT. Deze bestanden zijn gemakkelijk en met diverse software te openen. In tekstbestanden kunnen echter verschillende tekensets worden gebruikt, om bijvoorbeeld Latijnse letters, leestekens en andere bijzondere tekens te representeren. DANS vertrouwt er op dat de tekenset Unicode, gebruikmakend van 'Byte Order Mark' en UTF-codering, de zekerheid geeft dat alle karakters in alle computeromgevingen correct worden gerepresenteerd.

2.4 Opmaaktaal

Standardized General Markup Language (SGML) en Extensible Markup Language (XML) zijn opmaaktalen die gebruikt worden voor tekstdocumenten en datasets, zowel voor presentatie aan mensen als om uitwisseling van data tussen computers mogelijk te maken.

XML, SGML en HTML zijn vaak gebruikte opmaaktalen. Indien valide en compleet (zie hieronder) worden deze formaten gezien als geschikte, toegankelijke en

duurzame formaten.

Buiten deze formaten kunnen op XML of SGML gebaseerde formaten voorkomen die enkel door specifieke software kunnen worden gelezen. Dergelijke bestanden kunnen niet zonder meer worden geaccepteerd zonder verdere controle; neem hiervoor contact op met DANS.

XML is een vorm van SGML: alle XML-bestanden zijn SGML-bestanden. Doordat XML veel strikter is wat betreft syntax, is het gemakkelijker te valideren. HTML (HyperText Markup Language) is een andere vorm van SGML die vooral bedoeld is voor presentatie van tekst met opmaak (en layout) en hyperlinks naar andere documenten.

Naast "gewone" HTML bestaat ook XHTML. Dat is HTML volgens de striktere regels van XML.

SGML en XML worden niet of nauwelijks verder ontwikkeld. Van HTML is kort geleden de nieuwste versie 5 officieel tot W3C-standaard gemaakt. Omdat webtechnologie zich blijft ontwikkelen, is de verwachting dat HTML ook verder ontwikkeld blijft worden.

XML, HTML en SGML zijn veel gebruikte en ook geschikte formaten voor opmaaktaal, maar er moet wel goed op worden gelet dat de bestandsformaten *valide* en *compleet* zijn:

Validiteit
Valide Markup Language documenten zijn ‘well-formed’ én voldoen aan de regels die voor de bestandsformaten gelden.

Well-formed documenten vereisen dat de inhoud op een bepaalde wijze is gedefinieerd. Well-formed XML voldoet aan syntaxregels die onder meer stellen dat de gebruikte tekenset ook de aangegeven tekenset is; er geen verboden tekens in het bestand worden gebruikt; er sprake is van één ‘root-tag’ en elke ‘<tag>’ correct wordt afgesloten met een ‘</tag>’.

De regels voor de inhoud van een Markup-document staan beschreven in een DTD (Document Type Definition) of (XML) Schema bestand. Bovenaan XML- en HTML-documenten staat een verwijzing naar een het gebruikte DTD of schema. Deze verwijzing dient ook echt naar dit schema-bestand te leiden. Het liefst wordt dit schema meegeleverd, tenzij het bij een betrouwbare publieke dienstverlener staat.
Als gebruik wordt gemaakt van een Schema of DTD dat geen standaard is, moet de deponering van de data eerst nader besproken worden met DANS.
Middels schema's en DTD's kunnen hele nieuwe 'bestandsformaten' worden gedefinieerd, zoals SVG (Scalable Vector Graphics, voor vectorafbeeldingen), TEI (Text Encoding Initiative, gebruikt om tekst op te maken en te annoteren) en MathML (voor wiskundige formules).
Het World Wide Web Consortium (W3C) beheert de specificaties voor HTML en XML, en biedt een ‘Markup Validator’ die zowel XHTML als HTML kan valideren. Bovendien kan de validator enkele andere formaten valideren, zoals SMIL en MathML: http://validator.w3.org/

Compleetheid
Opmaaktaal kan berusten op het gebruik van andere bestandsformaten in aparte bestanden of binnen een bestand. Alle bestanden die bij een XML/HTML/SGML-bestand horen, moeten worden meegeleverd. Veel voorkomende, direct aan opmaaktaal gerelateerde bestanden zijn XLST stylesheets, CSS-definitiebestanden en JS/ES scripttalen:
Extensible Stylesheet Language Tranformations (XSLT)
XSLT is een XML-vocabulaire voor het transformeren van o.a. XML-bestanden. XSLT is een open standaard en is goed ondersteund. Mits gelinkte bestanden meegearchiveerd worden, kunnen we dit accepteren.
Cascading Style Sheets (CSS)
CSS is veelgebruikt op het web en wordt gebruikt om de opmaak van Markup Language documenten te definiëren. Er bestaan verschillende versies van CSS, Voor duurzame archivering moet duidelijk zijn bij welke bestanden de CSS hoort en welke versie CSS is gebruikt. Omdat browserspecifieke extensies voor kunnen komen moet ook bekend zijn wat de doelomgeving van de bestanden is, tenzij alleen de basiselementen worden gebruikt. Mocht de CSS referenties naar andere CSS-bestanden of externe bestanden bevatten, dan moeten deze links werken.
JavaScript / ECMAscript (.JS; .ES; …)
JavaScript en vergelijkbare scripttalen worden gebruikt voor allerlei zaken. Browsers lezen de scripts en voeren daarop commando's uit. De basis van JavaScript is goed ondersteund en kan gearchiveerd worden, maar bij script-bestanden moet wel opgelet worden dat er sprake kan zijn van afhankelijkheid van externe data.

2.5 Spreadsheets

Spreadsheets worden voornamelijk gebruikt voor omgang met tabulaire data: waarden in cellen, geordend in rijen en kolommen.

Een spreadsheet is echter vaak veel meer dan een platte tabel. Spreadsheets kunnen worden voorzien van nadere opmaak, denk bijvoorbeeld aan het gebruik van kleur in cellen of aan de weergave van de lijnen tussen de cellen. Ook kan de structuur van een spreadsheet van belang zijn. Cellen kunnen bijvoorbeeld berusten op berekeningen die worden gemaakt op basis van waarden in andere cellen. Daarom moet bij spreadsheets goed opgelet worden welke eigenschappen van belang zijn om te behouden; welke 'significant characteristics' in het bestand zitten.
Het formaat Open Document Spreadsheet (.ods) is een open, redelijk goed ondersteund en robuust spreadsheet-formaat dat is aan te bevelen als preferred format voor de duurzame opslag van spreadsheets met berekeningen en/of andere nadere (structuur)eigenschappen.

Kan een spreadsheet worden gezien of worden teruggebracht tot een platte tabel van rijen en kolommen? Dan kan ervoor worden gekozen om een CSV (Comma Separated Values) tekstbestand van de te tabel te maken. Zie het stuk 'CSV-bestanden' hieronder voor een nadere uitleg over de omgang met dit formaat. CSV-bestanden zijn enkel geschikt voor de opslag van platte tabellen. Een CSV behoudt geen opmaak (tekst noch cellen), formules, links naar externe bronnen . Is een directe visualisatie het primaire doel van de spreadsheet? Dan kan het bestand eventueel als een opgemaakt tekstbestand worden behandeld en als PDF/A worden aangeboden. Zie het onderdeel Preferred Formats – Opgemaakte

tekst voor nadere informatie.
PDF/A is primair geschikt voor de presentatie van opgemaakte tabellen. Het formaat biedt beperkte ondersteuning voor eigenschappen van spreadsheet als formules en links naar externe bronnen.

2.6 Databases

Databases bestaan in verschillende vormen, al is de bekendste vorm wellicht de relationele database. Databases worden beheerd door een Database Management System (DBMS). Naast zorgen voor consistentie in de data en het verwerken (lezen en schrijven) ervan, houden DBMS'en ook rollen en privileges voor gebruikers(groepen) bij en bieden ze een reeks functies aan om bewerkingen op de data uit te voeren.
Het bestandsformaat is meestal gekoppeld aan het DBMS, maar er bestaan onafhankelijke uitwisselformaten.
Veel DBMS'en ondersteunen de ISO-gestandaardiseerde versie van Structured Query Language (SQL): een taal om relationele databases te bevragen en te updaten. Samen met de Data Definition Language, om schema's te definiëren en aan te passen, kan de inhoud van een database als een verzameling schema- en datastatements worden opgeslagen.
De taal verandert zelden, maar de verschillende aanpassingen en uitbreidingen kunnen met de software meeveranderen. Indien uitbreidingen gebruikt wordt, moet uit de documentatie blijken welke SQL-versie gebruikt is.
Het is mogelijk om in SQL te verwijzen naar niet-bestaande en/of externe data, zonder dat het bestand ongeldig wordt. Als SQL gebruikt wordt voor uitwisseling van data, moeten daarom eventuele referenties worden meegeleverd, of moet elke referentie worden vervangen door de data waaraan gerefereerd werd.
Voor relationele databases wordt SIARD als geschikt duurzaam formaat gezien. SIARD: 'Software Independent Archiving of Relational Databases', is gemaakt voor het archiveren van relationele databases op een manier die zoveel mogelijk onafhankelijk is van het oorspronkelijk DBMS. Het formaat houdt rekening met de bekende te behouden eigenschappen ('significante karakteristieken') van databases.
SIARD is een open, gratis beschikbaar formaat en gebaseerd op heldere tekst-formaten: Unicode, XML, SQL(1999) waardoor het met diverse tools te openen is.
SIARD is een relatief jong formaat. Er bestaan tools voor het omzetten van databases naar SIARD alsmede voor het valideren van het formaat, maar de mogelijkheden zijn nog beperkt. Enkele conversietools zoals AccessToSiard en CSV2SIARD zijn te vinden via de volgende website: http://coptr.digipres.org/Category:File_Format_Migration. Voor het gebruik van deze tools is de SIARD Suite nodig. Uitleg over dit programma, met een link naar de website waar deze gratis aangevraagd kan worden, is te vinden op dezelfde website: http://coptr.digipres.org/SIARD_Suite
Databases kunnen gebruik maken van routines, die afhankelijk kunnen zijn van eigen scripttalen/programmeertalen uit het DBMS. Bij conversie naar SIARD is er een potentieel risico op het verlies van dergelijke routines, maar dit risico wordt niet heel groot geacht omdat dergelijke gebruikte talen naar verwachting niet in onbruik zullen raken.
dBase, HDF5 en Microsoft Access zijn alternatieve formaten voor databases die kunnen worden beschouwd als acceptabel, maar die beter geconverteerd kunnen worden naar meer duurzaam bestandsformaten.

Het dBase Table File Format (.dbf) is een propriëtair product. Het bedrijf achter dBase blijft wel oude versies van het formaat ondersteunen. Voor oudere versies dan versie 7 is geen officiële documentatie beschikbaar gesteld. dBase formaten zullen in diverse andere database applicaties kunnen worden ingelezen, waaronder LibreOffice/OpenOffice, MySQL, MS Access. Vanuit deze applicaties is het goed mogelijk om de dBase data naar andere formaten te exporteren.

Het Hierarchical Data Format (versie 5, niet compatibel met eerdere versies) is een algemeen datasetformaat met de mogelijkheid data in multidimensionele arrays op te slaan, gegroepeerd in collecties en/of hiërarchieën. Relaties tussen data in de arrays kunnen worden opgeslagen, maar het formaat biedt geen mogelijkheid voor het opslaan van gestructureerde (beschrijvende) metadata. Het formaat is open en kan ingelezen worden in diverse applicaties, maar het is moeilijk tot niet te verwerken zonder gebruik te maken van HDF5-software. Zie hiervoor de tools op http://www.hdfgroup.org/products/hdf5_tools/

In de praktijk wordt Microsoft Access veel gebruikt voor het maken van databases. De Access-formaten MDB en ACCDB worden buiten het commerciële Microsoft Access echter zeer slecht ondersteund. Door de verschillende versies van deze formaten kan het zelfs zo zijn dat Access zelf de bestanden niet altijd goed ondersteunt.
Vooralsnog heeft DANS voor veel databases gemaakt met Microsoft Access een duurzame en toegankelijke verwerking verzorgd door de tabellen uit de databases als losse CSV-tekstbestanden op te slaan. Zie het stuk 'CSV-bestanden' hieronder voor een nadere uitleg over de omgang met dit formaat. Opslag van de tabellen als CSV-bestanden behoudt enkel de tabulaire data uit een database. Eventuele overkoepelende documentatie wordt bij de CSV-bestanden in een apart document beschreven. In Microsoft Access-databases kan gebruik worden gemaakt van de functie 'Databasedocumentatie' voor het genereren van een document met kolombeschrijvingen en tabelrelaties: dit document kan conform opgemaakte tekst als PDF/A worden opgeslagen en met de tabellen van de database worden geleverd.
Daarnaast moet er op gelet worden dat alle gebruikte codes en variabelen verklaard kunnen worden, ook dit kan middels het voorzien van nadere beschrijvingen in een apart document ('codeboek').

CSV-bestanden
CSV, 'Comma Separated Values', is een wijze om tabulaire data in platte tekst te schrijven. Dit formaat biedt geen ondersteuning voor datatypen en metadata anders dan een kolomtitel. Het is de facto gebaseerd op de open standaard RFC4180, al bestaan er verschillende varianten (dialecten).
In een CSV-bestand worden de aparte waarden/cellen uit een tabel van elkaar gescheiden met een komma als scheidingsteken. CSV-bestanden kunnen in database-applicaties worden ingelezen, maar kunnen ook helder en snel als spreadsheet worden geopend, in bijvoorbeeld Microsoft Excel. Ook kunnen deze bestanden als tekstbestanden worden gelezen, bijvoorbeeld in Notepad.
Veel applicaties zullen CSV-bestanden zonder problemen kunnen openen. Afhankelijk van de standaardinstellingen op de computer voor het gebruik van scheidingstekens, kan het echter wel voorkomen dat een programma de kolommen niet automatisch van elkaar scheidt. In de applicatie kunnen kolommen nader worden gesplitst op basis van scheidingstekens; eventueel kan

de standaardinstelling op de computer worden aangepast. Bij Windows-systemen staat deze standaardinstelling onder 'Decimaal scheidingsteken'/'List separator' in het 'Land en Taal'/'Region and Language' configuratiescherm. Als hier een komma als scheidingsteken staat, zullen de CSV-bestanden in alle applicaties correct in gescheiden kolommen worden weergegeven.

2.7 Statische data

Er zijn verschillende software pakketten waarmee statistische analyses kunnen worden uitgevoerd. De meest gebruikte zijn SPSS, STATA, R en SAS.
Voor lange termijn archivering gebruikt DANS het SPSS portable formaat. Hoewel dit een gepatenteerd formaat is, is hiervoor gekozen omdat bij het opdelen van data en file-informatie er informatie verloren kan gaan.
De software SPSS wordt veelvuldig gebruikt waardoor verwacht mag worden dat het formaat in de toekomst toegankelijk blijft.
Bij software pakketten waarbij de data niet om te zetten zijn naar SPSS portable, archiveert DANS de data als data en setup.
Het SPSS portable bestand is weliswaar geschikt voor de duurzame archivering, maar in de praktijk zijn gebruikers over het algemeen niet bekend met dit formaat. Voor de toegankelijkheid wordt statistische data daarom ook beschikbaar gesteld in de standaard formaten van de meest gebruikte statistische software pakketten: SPSS .sav en STATA .dta.

2.8 Afbeeldingen (raster)

Voor raster afbeeldingen geeft DANS de aanbeveling om deze als ongecomprimeerde TIFF te archiveren én daarnaast als JPEG-bestanden te publiceren.

Apparaten zoals computerschermen, printers en dataprojectors kunnen digitale beelden verwerken. Dit doen ze door de beeldpunten of pixels waaruit een digitaal beeld bestaat te vertalen naar de specificaties van het apparaat. Het aantal pixels en de kleur van de pixels bepalen de verschijningsvorm van de digitale foto. De pixels vormen de snijpunten van een fijnmazig raster, vandaar dat deze beelden rasterimages worden genoemd.

De kwaliteit van een rasterimage wordt bepaald door de volgende factoren, die door de producent worden bepaald:

1. De resolutie. De pixeldimensie van een rasterimage bestaat uit het totaal aantal pixels in de horizontale en verticale dimensie. De fijnmazigheid of resolutie wordt uitgedrukt in het aantal pixels dat er per inch (2,54 centimeter) aanwezig is. De resolutie dient afgestemd te zijn op de details van het object dat gedigitaliseerd is. Dus niet te grof en niet te fijnmazig.
2. De dynamiek. In welke mate bevat het rasterimage alle kleuren van het origineel en hoe zijn deze kleuren gecodeerd; welke kleurruimte is toegepast? Accurate kleurweergave vereist kalibratie van de opnameapparatuur door een expert.
3. Compressie. Omdat rasterimages uit miljoenen pixels kunnen bestaan, kunnen compressietechnieken toegepast worden om de bestandgrootte te verkleinen.
4. Documentatie. Beschrijvende en technische/administratieve metadata. Deze kan zowel in het rasterimage worden opgenomen of apart worden gemaakt

(of een combinatie hiervan). Vele digitale camera's ondersteunen de EXIF standaard. Deze standaard bevat beschrijvingen zoals het tijdstip van de opname en camera-instellingen.
5. Het bestandsformaat. Het gekozen bestandsformaat dient bovenstaande kenmerken efficiënt en effectief te ondersteunen.

Met betrekking tot de archivering en duurzaamheid van rasterimages is het essentieel dat in de toekomst de rasterimages conform de intentie van de deponeerder gereproduceerd kunnen worden. Met gebruik van de formaten TIFF, JPEG en PNG kan redelijkerwijs worden aangenomen dat deze zonder problemen kunnen worden weergegeven en dat er standaard imageprocessing software beschikbaar is om de images te "renderen".

Ongecomprimeerde TIFF is het preferred format van DANS voor het behoud van raster afbeeldingen in maximale kwaliteit op de lange termijn.
TIFF-bestanden kunnen echter zeer omvangrijk zijn, wat ten koste kan gaan van de gebruiksvriendelijkheid. Daarom is het aan te bevelen om TIFF te gebruiken als archiveringsformaat en daarnaast de afbeeldingen voor gebruik beschikbaar te stellen in het breed ondersteunde formaat JPEG.
Het formaat PNG kan ook gekenmerkt worden als geschikt archiveringsformaat en is kleiner van omvang dan TIFF. Maar let op: PNG biedt beperkte mogelijkheden voor de opslag van technische/administratieve metadata in het bestand; het formaat biedt bijvoorbeeld geen ondersteuning van de hierboven genoemde EXIF standaard. Bij gebruik van PNG moet er dus op gelet worden of eventuele relevante metadata behouden blijft.
Er is de laatste jaren veel discussie over welk formaat geschikt is voor archief-doeleinden. De Koninklijke Bibliotheek beschouwt JPEG2000 inmiddels als haar archiefformaat, maar er zijn ook veel experts die een voorkeur hebben voor de andere formaten. Een probleem is dat de formaten (zowel JPEG als TIFF) verschillende compressiemethoden ondersteunen die niet allemaal als duurzaam worden gezien. Het JPEG2000 formaat dat de core coding ondersteunt wordt als meest duurzaam gezien.
De JPEG2000 images dienen te voldoen aan de "Part 1" van de JPEG2000 image compressie standaard (ISO/IEC 15444-1). Het programma "jpylyzer" kan gebruikt worden om het formaat te valideren en de technische kenmerken te extraheren: http://jpylyzer.openpreservation.org/
DANS is in het verleden een paar keer in aanraking gekomen met het formaat DICOM (Digital Imaging and Communications in Medicine); een standaard die beschrijft hoe medische beeldinformatie dient te worden opgeslagen, uitgewisseld en geprint.
De standaard definieert naast het bestandsformaat ook protocollen voor netwerken en applicaties en wordt gebruikt in zogenaamde PACS (Picture Archiving and Communication Systems): systemen die de beelden maken en beheren, bijvoorbeeld medische apparatuur.
In de standaard worden diverse soorten images ondersteund voor verschillende medische toepassingen, zowel stilstaand beeld als bewegend beeld. DICOM ondersteunt algemeen toegepaste compressiestandaarden, zoals JPEG en JPEG2000, of MPEG-2 voor videosequenties.
Het copyright op de standaard is in handen van het Amerikaanse National Electrical Manufacturers Association (NEMA). DICOM viewing software kan onderverdeel worden in twee groepen: (1) proprietary viewers die onderdeel zijn van de (medische) opnamesystemen en (2) DICOM viewing software voor PCs.

De populairste non-proprietary viewers die gratis beschikbaar zijn, zijn DicomWorks, Osiris en IrfanView (een veel gebruikte "all-format" viewer). Adobe heeft een plug-in ontwikkeld voor Photoshop die het mogelijk maakt om DICOM images en "header" (= metadata) informatie te bekijken of te exporteren naar andere formaten. Het programma IrfanView is ook in staat om images en/of animaties (sequentie van images) uit Dicom bestanden te halen.
Deze images kunnen in een "preferred format" als image of filmpje worden opgeslagen. In overleg met de depositor dient vastgesteld te worden of de contextinformatie in de DICOM bestanden relevant is voor archivering.

2.9 Afbeeldingen (vector)

SVG staat voor 'Scalable Vector Graphics'. Het is een robuust, op XML gebaseerd formaat voor statistische en dynamische vectorafbeeldingen. SVG is een open standaard en de ondersteuning van het formaat is over het verloop van tijd sterk toegenomen.

SVG vector afbeeldingen kunnen worden geopend in web-browsers als Firefox, Safari, Google Chrome en Explorer. Voor nadere bewerking kunnen vector image applicaties als Adobe Illustrator of Inkscape worden gebruikt. Inkscape is gratis te downloaden van de website: http://inkscape.org en werkt op Windows, Mac OS X en Linux.

Alle gangbare Vector Image formaten (EPS, AI, WMF, CDR) kunnen in Inkscape en Adobe Illustrator worden geopend en geconverteerd naar SVG.

2.10 Audio

De belangrijkste kenmerken van audio bestanden zijn:

1. De duur van het audio signaal.
2. Bitdiepte: het aantal bits waarmee het bemonsterde signaal wordt opgeslagen. Hoe meer bits er worden gebruikt, hoe nauwkeuriger we het originele signaal kunnen opslaan en dus kunnen reproduceren.
3. Bemonsteringsfrequentie: hoeveel bemonsteringen van het originele signaal we per seconden doen. Dit wordt meestal weergegeven in Hertz. Volgens de *nyquist-frequentie* geldt dat wanneer de bemonsteringfrequentie 2 maal zo groot wordt genomen als de hoogste frequentie in het signaal, het originele signaal exact gereproduceerd kan worden.
4. Het aantal kanalen: een waarde die het aantal unieke signalen in het audio object aangeeft. Bijvoorbeeld 2 (stereo).

Om opslagruimte en bandbreedte te besparen zijn er een aantal 'lossy' bestandsformaten bedacht. Deze formaten offeren een deel van de geluidskwaliteit op door op een slimme manier frequenties te verwijderen uit het geluidsspoor waardoor er minder data hoeft op te worden geslagen. Het gevolg hiervan zijn bestanden met een kleinere bestandsgrootte. Voor duurzame archivering is het wenselijk een 'lossless' bestand te leveren: een formaat met de beste kwaliteit, waar geen gegevens uit verloren zijn gegaan. Als gebruiksformaat kan het echter veel gebruiksvriendelijker zijn om een lossy export van de data aan te bieden. Het beste kan per geval worden bekeken of

het wenselijk is om naast de lossless originele data ook lossy formaten te leveren.

Het door Microsoft en IBM ontwikkelde WAVE formaat is het meest gebruikte formaat om ongecompliceerde audio op te slaan er zijn daarom veel tools beschikbaar om deze af te spelen of om te zetten. Er is een maximum bestandsgrootte van 4GB omdat het een 32-bit header gebruikt. Dat komt overeen met 6,8 uur aan audio (op CD Kwaliteit 44.1 kHz, 16-bit stereo). Voor bestanden die groter zijn is er door de Europese Radio-Unie het een uitbreiding op het WAVE formaat in het leven geroepen, namelijk het Broadcast Wave Formaat (BWF)

Het door Apple ontwikkelde AIFF formaat wordt ook gebruikt om ongecomprimeerde audio op te slaan. Het is een formaat vergelijkbaar met WAV maar het heeft niet zo een grote aanhang als WAV. Naast het standaard AIFF bestaat ook een variant genaamd AIFF-C/sowt; deze maakt gebruik van een andere volgorde van bytes waardoor de bestanden kleiner zullen zijn. De twee verschillende implementaties van AIFF zijn qua kwaliteit wel identiek en hebben beide de extensie '.aiff'. De verschillende opbouw van de twee implementaties kan problemen opleveren bij het afspelen van AIFF formaten, met name in oudere applicaties die niet op AIFF-C/sowt zijn berekend.

Het open formaat FLAC past lossless compressie toe, maar is minder bekend dan andere formaten en kent daardoor relatief geringe ondersteuning.

Qua 'lossy' audio is het meest bekende en meest gebruikte formaat MPEG1 – Audio Layer 3 (MP3) audio. Dit formaat kent een zeer brede ondersteuning. Een opvolger van het MP3 formaat is het Advanced Audio Codec (AAC). Deze is net als MP3 ook lossy maar biedt een soortgelijke geluidskwaliteit bij een lagere bitrate. Een gevolg hiervan is dat de bestandsgrootte van een AAC bestand kleiner is dan die van een MP3 bestand zonder dat dit merkbaar is in de kwaliteit.

2.11 Video

MPEG-2 is een open standaard ISO/IEC (13818) ontwikkeld voor DVD en digitale televisie. Het bevat een aantal profielen (Simple, Main, 4:2:2) welke verschillende standaard specificaties bevatten voor elementen als weergave, grootte en data rate.
MPEG-2 is niet geoptimaliseerd voor lage bitrates (< 1 Mbit/s) maar biedt een superieure kwaliteit bij hogere bitrates (> 3 Mbit/s) vergeleken met MPEG-1.
MPEG-4 is open standaard en een verdere ontwikkeling van de MPEG video standaard ISO/IEC (14496) en is deels gebaseerd op het Apple QuickTime (.mov). De uitbreidingen zijn vooral gedaan om het gebruik van AV materiaal op het internet beter te faciliteren door betere compressie middels de H.264 codec. MPEG-4 bestaat uit twee hoofdversies en bevat een grote hoeveelheid profielen welke geoptimaliseerd zijn voor verschillende doelen (web streaming, voice, HD video, etc).
Zoals aangeven is QuickTime (.mov) gebruikt als basis voor de MPEG-4 standaard. QuickTime bied weliswaar dezelfde (of meer) functionaliteit als MPEG-4; toch wordt het aangeraden om MPEG-4 te gebruiken wanneer mogelijk, omdat MPEG4 een vastgestelde standaard is.
Matroska (.mkv) is een open source en zeer flexibel alternatief voor bestaande container formaten zoals AVI, ASF, MOV, RM, MP4, MPG, etc. en kan vrijwel alle

codecs bevatten. MKV heeft echter als nadeel dat het nog een relatief geringe ondersteuning door videosoftware kent.

2.12 Computer Aided Design (CAD)

CAD: 'Computer Aided Design', is het gebruik van computers voor het maken van digitale tekeningen.
De ontwikkelaar Autodesk, met als voorname software AutoCAD, is absolute marktleider op het gebied van CAD. Hierdoor zijn de populaire, veelgebruikte CAD-formaten geen open formaten. Noch zijn open formaten ontwikkeld voor de uitwisseling van CAD-formaten.
De formaten van AutoCAD zijn DWG en DXF. Deze formaten worden ondersteund door vrijwel alle andere CAD-applicaties. DXF is specifiek ontworpen om data interoperabiliteit tussen AutoCAD en andere programmas te faciliteren. DXF versie R12 lijkt het beste ondersteund te worden voor succesvolle en correcte import in andere applicaties.
Een groot probleem met het gebruik van DXF is de ontwikkeling van het DWG-formaat. DWG biedt inmiddels mogelijkheden waarvan niet alle eigenschappen in DXF kunnen worden opgeslagen. Vooralsnog is DXF R12 echter wel de beste optie voor preservatie van CAD in een relatief open, breed ondersteund formaat. Wel moet altijd worden gecontroleerd of de export van DWG naar DXF niet tot verlies van data leidt; anders is het beter om het bij de DWG te houden.
Vanuit AutoCAD kunnen CAD-tekeningen gemakkelijk worden opgeslagen als DXF R12: File=>Save as=>Files of type: AutoCAD R12/LT2 DXF. Het is wenselijk om de CAD-tekening eerst in AutoCAD op te schonen door tijdelijke informatie uit het bestand te verwijderen met het commando 'purge' (purge all).

CAD-tekeningen kunnen worden opgemaakt in een 'Layout' met een afbeelding voor publicatie als doel. Dergelijke opgemaakte afbeeldingen kunnen goed vanuit AutoCAD naar PDF/A worden geprint (File=>Plot, gebruik de Adobe PDF printer, zet bij de 'properties' de settings op 'PDF/A-1b:2005(RGB)'). Dit behoudt het visuele doel van de afbeelding en is hiervoor een uitstekende oplossing, echter zal de digitale tekening niet meer in CAD te importeren zijn; de afbeelding verliest de verdere bewerkbare eigenschappen.

2.13 Geografische Informatie (GIS)

Met GIS, oftewel Geografische Informatie Systemen, worden digitale kaarten en afbeeldingen gemaakt. Het betreft veelal vector-afbeeldingen met een achterliggende datatabel als basis. Deze tabel is binnen de GIS-applicatie als tabulaire data te openen.

De voorname GIS-applicaties zijn ESRI ArcGIS en Pitney Bowes MapInfo Professional. ArcGIS slaat de data voornamelijk op als Shapefiles: een .shp met minstens twee bijbehorende bestanden .shx, .dbf, met optioneel tot 12 extra, aanvullende bestanden (.prj, .shp.xml, …).
MapInfo gebruikt TAB-bestanden; net als de Shapefiles bestaan de TAB-bestanden uit een collectie van bij elkaar behorende bestanden. Het hoofdbestand is een .tab-bestand, daarbij hoort een tabulair databestand: .dat .dbf of .xls, optionele bijbehorende bestanden zijn extensies .map, .id, .ind.
MapInfo TAB en ESRI Shapefiles worden veel gebruikt en kunnen indien gewenst als gebruiksformaat worden aangeboden. Maar voor de duurzaamheid op de

lange termijn zijn deze formaten niet geschikt. Beide formaten bestaan veelal uit binaire data, waarvan het niet gegarandeerd kan worden dat andere applicaties dan de applicaties waar ze mee zijn gemaakt de data foutloos kunnen openen.

Voor de lange termijn is het aan te bevelen om GIS-data op te slaan in een open, goed ondersteund en robuust tekstbestand. Twee formaten zijn hiervoor geschikt en gelden allebei als preferred formats voor GIS:
- GML is een XML ISO-standaard voor geografische data. Ondersteuning voor GML was voor de opname als ISO-standaard beperkt, maar de ondersteuning is sindsdien toegenomen en zal naar verwachting steeds beter worden.
-Het 'MapInfo Interchange Format' .mif, doorgaans verbonden met het bestand .mid, is het exportformaat van MapInfo, ontworpen met het oog op GIS-interoperabiliteit. Het is een helder, duidelijk gedocumenteerd, goed ondersteund en stabiel ASCII-tekstbestand.

GIS-applicaties bevatten standaard import-opties voor GML en MIF, alsmede opslag en –export opties naar GML en/of MIF. Voor betere export- en importmogelijkheden voor ArcGIS is eventueel de 'Data Interoperability extension' verkrijgbaar: hiermee kunnen bulk-conversies gemakkelijk worden uitgevoerd.

2.14 Afbeeldingen (georeferentie)

Gegeorefereerde afbeeldingen zijn raster afbeeldingen (TIFF, JPEG) voorzien van een middel om de afbeelding in Geografische Informatie Systemen (GIS) in te lezen. De afbeeldingen worden daarbij geprojecteerd en geschaald in een coördinatenstelsel.
GeoTIFF is een metadata-standaard voor het toevoegen van georeferentie aan een TIFF-afbeelding. Deze metadata wordt in het TIFF-bestand zelf opgenomen. Het is een open en goed ondersteund formaat.

2.15 Raster GIS

Geografische Informatie Systemen (GIS) worden voornamelijk gebruikt voor het maken van digitale vector-afbeeldingen (kaarten) met een achterliggende data-tabel. GIS kan echter ook worden gebruikt voor het maken van raster-afbeeldingen. Op basis van input in GIS kan bijvoorbeeld een hoogtekaart worden gegenereerd. Een raster-image gegenereerd in GIS kan nader worden opgemaakt met een kleurenschema.
Een dergelijke GIS raster-afbeelding wordt vaak een *grid* genoemd.

Grid-bestanden die direct aan commerciële pakketten gelieerd zijn zullen een lage mate van openheid, interoperabiliteit en robuustheid genieten. Een ESRI ArcInfo Grid (ook bekend als ArcGrid) kan bijvoorbeeld gebruik maken van diverse subdirectories met grotendeels binaire bestanden: .adf, .nit, .dir, log, …

Het is aan te bevelen om Grid-bestanden zoveel mogelijk om te zetten naar ASCII-tekst. Het mag van GIS-applicaties verwacht worden dat zij ASCII-grid bestanden correct kunnen importeren.
De ArcCatalog van ESRI ArcGIS biedt 'convert GRID to ASCII'-mogelijkheden; Surfer heeft een Grid=>Convert=>save to GS ASCII optie.
Let wel op: de conversiemogelijkheden zijn niet onbeperkt, noch probleemloos.

2.16 3D

Voor de opslag en de presentatie van 3D-afbeeldingen/modellen zijn geen bestandsformaten ontwikkeld die gemakkelijk gekenmerkt kunnen worden als 'preferred formats'. Het is een erkend probleem in de wereld van de digitale archivering: allerlei 3D-programma's hanteren eigen bestandsformaten, interoperabiliteit is beperkt en conversie naar andere formaten leidt snel tot verlies van functionaliteit of bepaalde eigenschappen van het bestand.
3D data is het beste in het oorspronkelijke formaat te behouden. Daarnaast kan gekeken worden of een export mogelijk is naar een open formaat. Voor het exportformaat gaat een primaire voorkeur uit naar X3D. Als X3D het 3D-model niet naar wens opslaat is COLLADA .dae de aanbevolen keuze. Controleer het exportformaat om te zien of de gewenste eigenschappen hierin worden opgeslagen, beschrijf elementen die ontbreken.

Voor enkel de geometrische objecten; zonder nadere aspecten als animaties en interactiviteit, is WaveFront OBJ het preferred format. OBJ is een zeer breed ondersteund open formaat voor de weergave van 3D geometrie. In een heldere, simpele structuur worden de ruimtelijke posities van elk punt van het object alsmede textuurcoördinaten geschreven.

Daarnaast kan nagedacht worden of het mogelijk is om onderdelen van de data op alternatieve wijze over te brengen. Zijn filmpjes (screencasts) of statische afbeeldingen geschikt voor het tonen van bepaalde informatie?
Hoewel er geen preferred format voor een interactief, dynamisch 3D-model bestaat zijn er mogelijk wel voorkeursformaten voor bepaalde elementen van het geheel.

2.17 RDF

RDF (Resource Description Framework) is een datamodel waarin kennis in grafen wordt uitgedrukt en met labels is geordend. Een aantal RDF-standaarden worden ondersteund door het World Wide Web Consortium (W3C). Naar verwachting zullen RDF applicaties altijd met deze W3C-standaarden om kunnen gaan:

- RDF/XML (.rdf)
- Trig (.trig)
- Turtle (.ttl)
- NTriples (.nt)
- JSON-LD

2.18 Computer Assisted Qualitative Data Analysis (CAQDAS)

In tegenstelling tot kwantitatief onderzoek, dat gebaseerd is op cijfermatige data waarmee berekeningen worden uitgevoerd, maakt kwalitatief onderzoek gebruik van niet-numerieke kwalitatieve gegevens, zoals teksten, afbeeldingen en films. Kwalitatieve gegevens kunnen opgeslagen worden in verschillende bestandsformaten waarvoor in veel gevallen al "preferred" en "acceptable" formats beschikbaar zijn (zoals TIFF of MP3).
CAQDAS (Computer Assisted Qualitative Data AnalysiS) programma's zijn programma's die verschillende onderzoeksmethoden en analysetechnieken faciliteren ter verrijking en analyse van kwalitatieve data. Voorbeelden van

CAQDAS programma's zijn ATLAS.TI, HyperRESEARCH, MAXqda, QDA Miner, NVivo, Qualrus en Transana.
In de loop der tijd zijn er verschillende versies van de programma's verschenen en zijn er versies voor verschillende besturingssystemen gekomen (bijv. Windows en/of MacOS). Vaak kunnen de bestanden afkomstig uit deze applicaties niet onderling uitgewisseld worden. CAQDAS programma's maken gebruik van gesloten "proprietary" formaten en ondersteunen geen open import of export formaten. Dit belemmert de duurzaamheid van de databestanden.
Een paar jaar geleden is een open standaard ontwikkeld om "richly encoded qualitative data" te representeren en te dienen als archiefformaat voor CAQDAS bestanden. De standaard is bekend onder de naam QuDEX (Qualitative Data Exchange Format) en onder andere beschikbaar in de vorm van een XML schema (zie http://data-archive.ac.uk/create-manage/projects/qudex). Het project heeft vooralsnog enkel een "proof of concept" opgeleverd, onder andere in de vorm van een QuDEX-aansluiting op het gesloten formaat van het programma ATLAS.TI, maar er zijn nog geen bruikbare archiveringsdiensten ontwikkeld op basis van het formaat. De standaard wordt nog niet ondersteund door de producenten van de verschillende CAQDAS programma's.
DANS is in de afgelopen jaren in aanraking gekomen met onderzoeksdata van twee soorten CAQDAS programma's: ATLAS.TI en NVIVO. Met betrekking tot de archivering van de bestanden die door deze programma's worden gemaakt dient gebruik te worden gemaakt van de exportfuncties van de programma's: ATLAS.TI kent de *copy bundle* functie. Deze functie is bedoeld om een samenhangende bundel te maken van alle bij elkaar horende bestanden. Deze bundel dient als back-up voor de bestanden en biedt de mogelijkheid om bestanden tussen computers te migreren.
NVIVO (versie 10) kent de optie *save project* waarmee alle onderdelen van een project (*sources, nodes, matrices, casebook*) kunnen worden gekopieerd. Deze bestanden samen vormen het archief.
Bij de documentatie van de exportbestanden dient de versie van het programma en het besturingssysteem opgenomen te worden.

Gebruikte afkortingen en acroniemen

.7bdat = SAS versie 7b data set (bestandsextensie gebruikt in SAS)
AAC = Advanced Audio Coding
.accdb = Access Database (bestandsextensie gebruikt in Microsoft Access 2007 en later)
.adf = ESRI ArcInfo Data File
.ai = Adobe Illustrator (bestandsextensie gebruikt in Adobe Illustrator)
AIFF = Audio Interchange File Format
ASCII = American Standard Code for Information Interchange
AVC = Advanced Video Coding
AVI = Audio Video Interleaved
BWF = Broadcast Wave Format
CAD = Computer-Aided Design
CAQDAS = Computer Assisted Qualitative Data Analysis
CDR = CorelDRAW file format
COLLADA = Collaborative Design Activity
CSS = Cascading Style Sheets
CSV = Comma Separated Values
.dae = Digital Asset Exchange (bestandsextensie gebruikt in COLLADA)
.dbf = dBase file (bestandsextensie gebruikt in dBase)
DDI = Data Documentation Initiative (metadata standaard voor statistische data, zie http://ww.ddialliance.org/)
DICOM = Digital Imaging and Communications in Medicine
.docx = Office Open XML document (bestandsextensie gebruikt in Microsoft Office)
.dta = data file (bestandsextensie gebruikt in STATA)
DTD = Document Type Definition
.dwg = Drawing (bestandsextensie gebruikt in AutoCAD)
.dxf = Drawing Interchange Format (bestandsextensie gebruikt in AutoCAD)
EPS = Encapsulated PostScript
ESRI = Environmental Systems Research Institute
ES = ECMAscript
.fbx = Filmbox (bestandsextensie gebruikt in software van Autodesk)
FLAC = Free Lossless Audio Codec
GIS = Geographic Information System
GML = Geography Markup Language
H.264 = codering voor video. Het is een standaard van de International Telecommunication Union (ITU. Ook bekend als MPEG-4 Part 10 of AVC.
HDF = Hierarchical Data Format
HTML = HyperText Markup Language
IEC = International Electrotechnical Commission
ISO = International Organisation for Standardization
JPEG = Joint Photographic Experts Group
JS = JavaScript
KML = Keyhole Markup Language
M4A = MPEG-4 Audio
MathML = Mathematical Markup Language
.mdb = Microsoft Access Database (bestandsextensie gebruikt in Microsoft Access 2003 en eerder)
MIF/MID = MapInfo Interchange Format / MapInfo Data File
MKV = Matroska Video

.mov = Apple QuickTime movie (bestandsextensie gebruikt in QuickTime)
MP2 = MPEG-1 Audio Layer II **MP3** =
MPEG-1/-2 Audio Layer III **MPEG** =
Moving Picture Experts Group
MPEG-2 = Moving Pictures Experts Group 2 (videoformaat)
MPEG-4 = Moving Pictures Experts Group 4
MS = Microsoft (bij software: MS Word; MS Excel, …)
.obj = WaveFront Object
ODS = OpenDocument Spreadsheet
ODT = OpenDocument Tekst.doc = document (bestandsextensie gebruikt in Microsoft Word)
OOXML = Office Open XML
PDF = Portable Document Format
PNG = Portable Network Graphics
.por = SPSS Portable (bestandsextensie van het uitwisselformaat van SPSS)
R = softwarepakket en programmeertaal, moderne implementatie van de programmeertaal S (Statistical programming language)
RDF = Resource Description Framework
RTF = Rich Text File
SAS = Statistical Analysis System (statistische data analyse software)
.sav = SPSS data file (bestandsextensie gebruikt in SPSS, de naam van de extensie is afgeleid van de opslag-functie 'save')
.sd2 = SAS data set (bestandsextensie gebruikt in SAS)
SGML = Standardized General Markup Language
.shp = Shapefile (bestandsextensie gebruikt in ESRI software)
SIARD = Software Independent Archiving of Relational Databases
SPSS = Statistical Package for the Social Sciences (dit softwarepakket wordt tegenwoordig ook buiten de sociale wetenschappen gebruikt)
SQL = Structured Query Language
SVG = Scalable Vector Graphics
.tab = Table File (bestandsextensie gebruikt in MapInfo)
TEI = Text Encoding Initiative
.tfw = TIFF World File
TIFF = Tagged Image File Format
.tpt = SAS Transport file (bestandsextensie gebruikt in SAS)
TXT = tekst
UTF = Unicode Transformation Formats
W3C = World Wide Web Consortium
WAVE = Waveform Audio File Format
WMF = Windows Metafile
X3D = XML-bestandsformaat voor representatie van 3D-computergraphics
XHTML = Extensible HyperText Markup Language
.xls = Excel Spreadsheet (bestandsextensie gebruikt in Microsoft Excel)
.xlsx = Office Open XML spreadsheet (bestandsextensie gebruikt in Microsoft Office)
XML = Extensible Markup Language
XSLT = Extensible Stylesheet Language Transformations

Voor u ligt het document Preferred formats. In dit document worden de 'preferred formats' of voorkeurs-bestandsformaten van DANS beschreven: de bestandsformaten waarvan DANS het vertrouwen heeft dat deze op de langere termijn de beste garanties bieden qua bruikbaarheid, toegankelijkheid en duurzaamheid. Per type data wordt een kort overzicht gegeven van de keuze voor het preferred format, van het gebruik van de data en van eventuele conversiemogelijkheden. Dit document is bedoeld als leidraad voor deponeerders van data. Neem voor meer informatie contact op met DANS.

Data Archiving and Networked Services (DANS)

DANS bevordert duurzame toegang tot digitale onderzoeksgegevens. Hiertoe stimuleert DANS dat wetenschappelijke onderzoekers gegevens duurzaam archiveren en hergebruiken, bijvoorbeeld via het online archiveringssysteem EASY (easy.dans.knaw.nl) en DataverseNL (dataverse.nl). Tevens biedt DANS met NARCIS (narcis.nl) toegang tot duizenden wetenschappelijke datasets, publicaties en andere onderzoeksinformatie in Nederland. Daarnaast verzorgt het instituut training en consultancy en doet het onderzoek naar duurzame toegang tot digitale informatie. Gedreven door data zorgt DANS er met zijn dienstverlening en deelname in (inter-)nationale projecten en netwerken voor dat de toegang tot digitale onderzoeksgegevens verder verbetert. Kijk op dans.knaw.nl voor meer informatie en contactgegevens.

Data Archiving and Networked Services (DANS)
Postbus 93067 | 2509 AB Den Haag
Anna van Saksenlaan 51 | 2593 HW Den Haag
+31 70 349 44 50
info@dans.knaw.nl | dans.knaw.nl

DANS is een instituut van KNAW en NWO

KNAW NWO

Door data gedreven