NVBS Excel naar ABCD

|  |  |
| --- | --- |
| Software | Excel 2016  Hulp programma in C  ABCD: 1.4 |
| Document historie | 2014-01-23 LR: Eerste versie (via Excel 2003, DB-III, db3iso.com op MS-DOS)  2018-04-20 FHo: Volledig vernieuwd door toepassen van cnv\_csv\_to\_iso  2018-05-02 FHo: Hoofdstuk Excel versies+ beter plaatje in inleiding  2018-05-10 FHo: CSV mag embedded linefeeds hebben. |
| File naam | NVBS Excel naar ABCD 2018-05-10.docx |
| Geassocieerde files | cnv\_csv\_to\_iso.c : Source code van het hulpprogramma  TestLettersEnVerschovenKolommen.xlsx: Test file |
| Leeswijzer | Weergave van een term of filenaam  Weergave van een kruisverwijzing  Weergave van een letterlijk detail in code of configuratie |

1 Inleiding 2

2 Werkvoorschrift Excel naar ABCD 3

2.1 Aanmaken Excel file 3

2.2 Converteren Excel spreadsheet naar een CSV file 4

2.3 Converteren CSV file naar een ISO file 5

2.4 Importeren van ISO file in ABCD 6

3 Achtergrond informatie 7

3.1 Teken sets 7

3.2 Lijstscheidingsteken 7

3.3 Excel versies 7

3.4 Ontwerp beslissingen 7

3.5 Portabiliteit / flexibiliteit van cnv\_csv\_to\_iso 8

# Inleiding

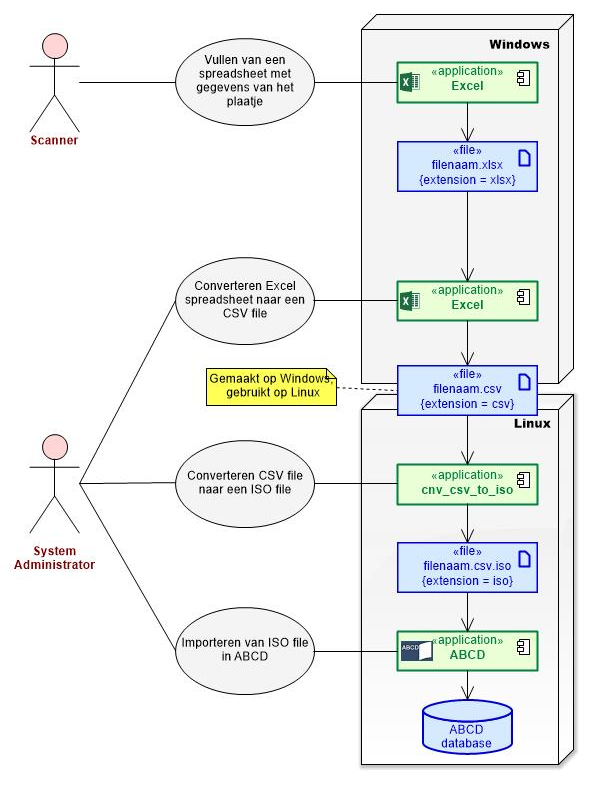
Dit document dient als werkvoorschrift voor het maken en inlezen van Excel files in ABCD.

Records kunnen direct in ABCD worden gemaakt. Records waarbij veel velden dezelfde, of nagenoeg dezelfde informatie bevatten, kunnen efficiënter met Excel worden gemaakt en daarna ingelezen worden in ABCD.

ABCD kan geen Excel files lezen, maar wel files in formaat ISO-2709. Dit document beschrijft de facetten van het proces om Excel files, via een hulpprogramma, om te zetten naar een ISO-2709 file en in te lezen in ABCD.

Er zijn veel werkwijzen mogelijk, afhankelijk van de gewenste procedure, de verdeling over de actoren en automatisering behoefte. Dit document beschrijft de huidige toestand van procedure en actoren met handmatige acties zoals van toepassing op de 3 SNR databases: SNR Beeldbank, SNR Railatlas, ODR Fotoarchief.

Hieronder een plaatje van het scenario in dit document. Andere scenario’s zijn mogelijk, maar worden hier niet uitgewerkt.



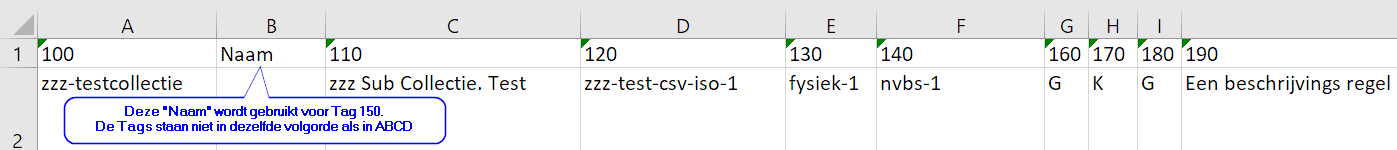
# Werkvoorschrift Excel naar ABCD

## Aanmaken Excel file

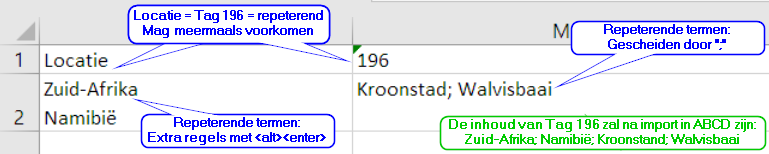
De Excel file heeft extensie .xls of .xlsx

Regel 1 bevat informatie over de ABCD velden (Tags) waarvoor informatie wordt aangeleverd. Voor regel 1 gelden de volgende regels:

1. Iedere cel van de eerste regel moet een bestaande Tag in ABCD zijn.  
   Een Tag is een 3-cijferig getal: 110, 120,…
2. De Tags mogen in willekeurige volgorde staan.
3. Het is niet nodig om alle Tags uit de ABCD database te vermelden. Als er geen gegevens voorhanden zijn kunnen ze weggelaten worden.
4. In plaats van een Tag mag ook de Title gegeven worden.  
   Voorbeeld: “Collectie” in plaats van “100”.
5. Een Tag van een *niet*-repeterend veld mag maar 1 maal voorkomen.  
   Een Tag van een repeterend veld *mag* vaker voorkomen.
6. Formattering van de cellen is mogelijk en wordt in de conversie weggelaten.

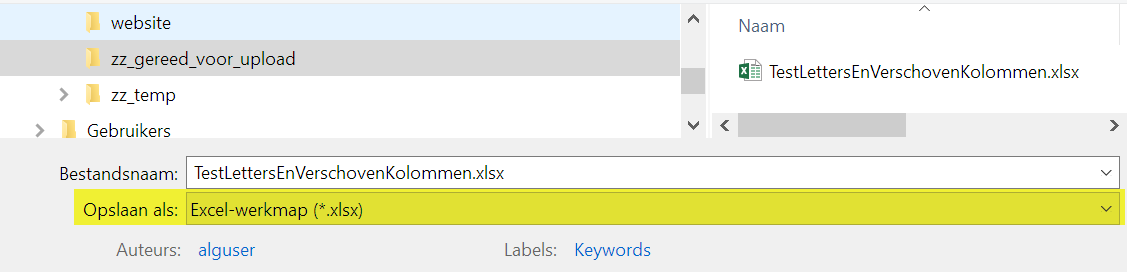
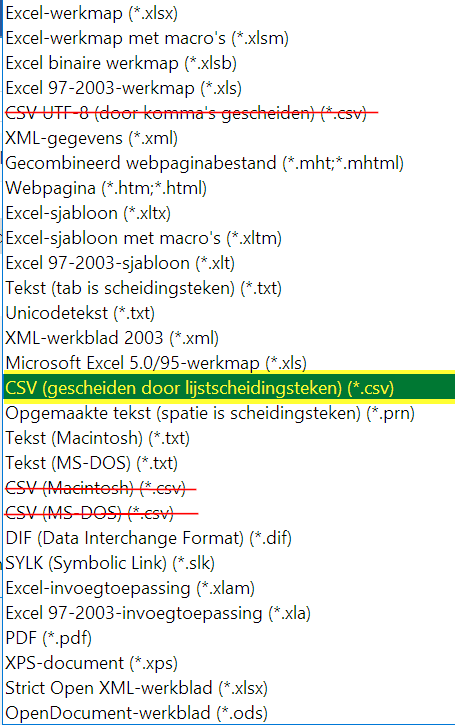
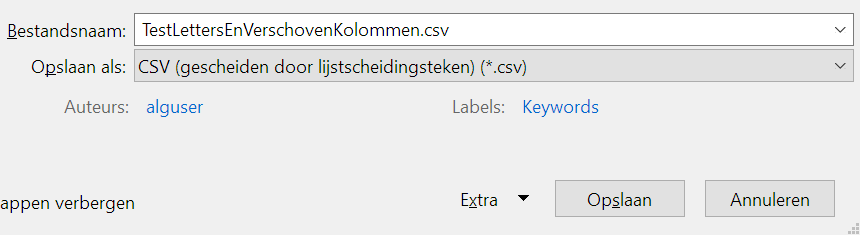
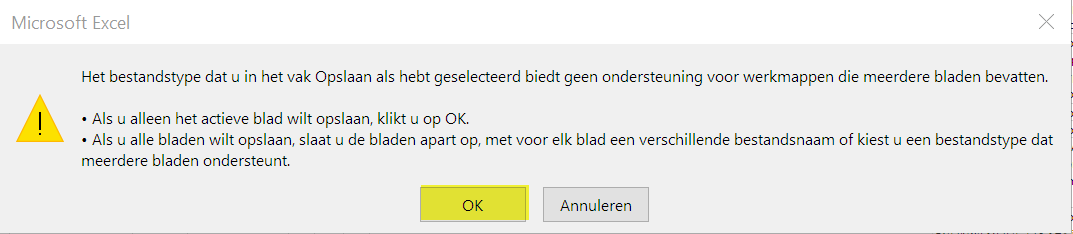
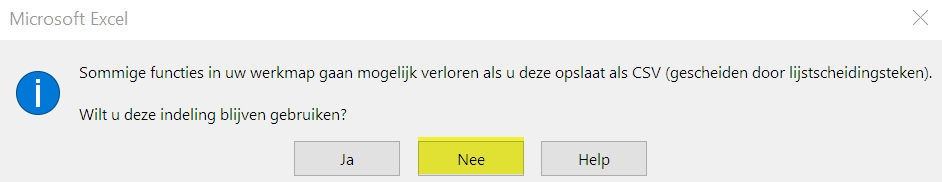
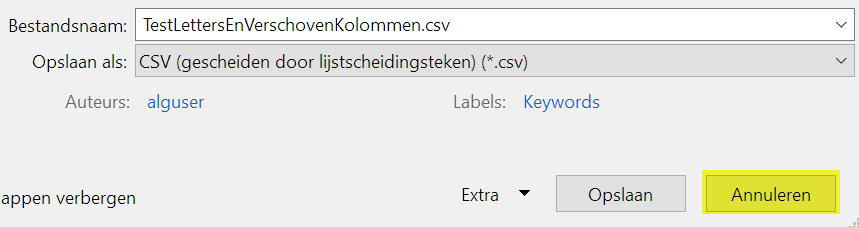


Ieder volgende regel beschrijft een nieuw record in ABCD. Voor deze regels geldt:

1. Een cel bevat informatie voor de betreffende Tag. Een cel mag leeg zijn.
2. Alle regels moeten hetzelfde aantal cellen hebben als de eerste regel.  
   Let op dat de laatste cel van een regel altijd informatie bevat. Excel exporteert lege cellen aan het einde niet.
3. Cellen waarvoor een picklist bestaat dienen de database waarde van ABCD te bevatten.  
   Voorbeeld: de picklist van Kwaliteit toont “Goed” en de database waarde hiervan is “2”.
4. De waarde van een cel wordt NIET getest op ABCD Input Type en/of lengte restricties.  
   🡺 Let op dat datums het juiste formaat hebben.  
   🡺 Let op dat velden met een picklist juiste waardes bevatten.
5. In repeterende velden kunnen meer waardes ingevoerd worden, gescheiden door een “;”.  
   Meer waardes in dezelfde cel kan ook door meer regels te gebruiken (Alt-Enter).  
   Als er meer kolommen zijn voor repeterende velden dan wordt de waarde van die cellen ook toegevoegd aan het repeterende veld.  
   
6. Alle leesbare letters uit ISO/IEC 8859-1 en Windows-1252 (is gelijk aan HTML ANSI) zijn technisch toegestaan. Zie 3.1 Teken sets.  
   Maar: De navolgende processen kunnen niet met al die tekens omgaan.  
   🡺 De tekens < en > worden door browsers geïnterpreteerd.  
   🡺 De dubbele quote “ geeft in combinatie met andere letters soms vreemde resultaten  
   🡺 De Oost-Europese letters worden bij afspraak niet gebruikt

## Converteren Excel spreadsheet naar een CSV file

Interactief kan dit door te op te slaan als CSV file.

1. Open de .xlsx file 🡪 Bestand 🡪 Opslaan Als 🡪 Kies een geschikte map
2. Klik de drop down list Opslaan Als  
   
3. Selecteer in de lijst “CSV (gescheiden door lijstscheidingsteken) (\*.csv)”.  
   Excel kiest de juiste code pagina: Windows-1252. Zie ook 3.2 Lijstscheidingsteken  
   Er zijn meer CSV opties, maar die geven in dit proces onjuiste resultaten.  
   🡺 
4. De dialoog toont nu een .csv extensie. Selecteer nu Opslaan.  
   En in de volgende pop-ups: OK 🡪 Nee 🡪 Annuleren  
     
     
   
5. Sluit Excel. Er is nu een CSV file.  
   Het is van belang dat Excel na deze actie gesloten wordt omdat bij een andere volgorde meer (verwarrende) dialogen verschijnen die ook tot onverwachte kopieën kunnen leiden.

## Converteren CSV file naar een ISO file

Interactief wordt deze actie gedaan in een bash-shell op Linux.  
Het conversie programma cnv\_csv\_to\_iso heeft geen interactie nodig en kan in scripts worden opgenomen. Zie de -h optie voor alle opties en varianten voor het aangeven van input/output redirection en environment variabelen.

Het volgende fragment toont een scenario.

$ export CNVXLS\_OPT\_8=y # Kan ook met commandline optie -8

$ ./cnv\_csv\_to\_iso -d odrfot.fdt -i TestLettersEnVerschovenKolommen.csv

-i: CSV to be converted will be read from : 'TestLettersEnVerschovenKolommen.csv'

-d: Field Definition Table will be read from: 'odrfot.fdt'

-o: Processing results will be written to : 'TestLettersEnVerschovenKolommen.csv.iso'

-c: Separator for elements in csv line : 'autodetect'

-r: Separator for repeats in csv elements : ';'

-8: ISO records are written as single line

-S: HTML brackets (<>) give a warning

Reading the File Definition Table: odrfot.fdt

Reading the File Definition Table: extracted 23 lines with Field information.

Processing csv file : TestLettersEnVerschovenKolommen.csv

Check first line of TestLettersEnVerschovenKolommen.csv for correct tags ...

- Best results in delimiter detection gives '|' (24 tokens).

- Column 2: value 'Naam' (= title in FDT) replaced by tag '150'

- Column 12: value 'Locatie' (= title in FDT) replaced by tag '196'

Check first line of TestLettersEnVerschovenKolommen.csv for correct tags: OK

Writing records to ISO file 'TestLettersEnVerschovenKolommen.csv.iso'

- Skipped record 6 (only empty tokens}

- Skipped record 14 (only empty tokens}

- Record 15: Detected 1 < signs, and 1 > signs.

Processed 16 records with 17 lines from CSV file 'TestLettersEnVerschovenKolommen.csv'

Written 13 records in 118 lines to ISO file 'TestLettersEnVerschovenKolommen.csv.iso'

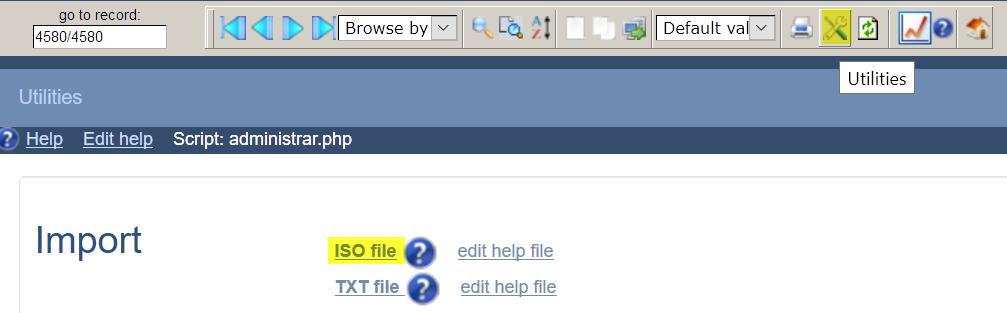
Successfull completion

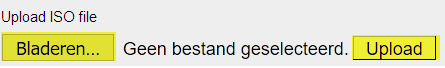
De status melding Successfull completion, met een exit status 0 geeft dat de ISO file correct gegenereerd is en geïmporteerd kan worden in ABCD.

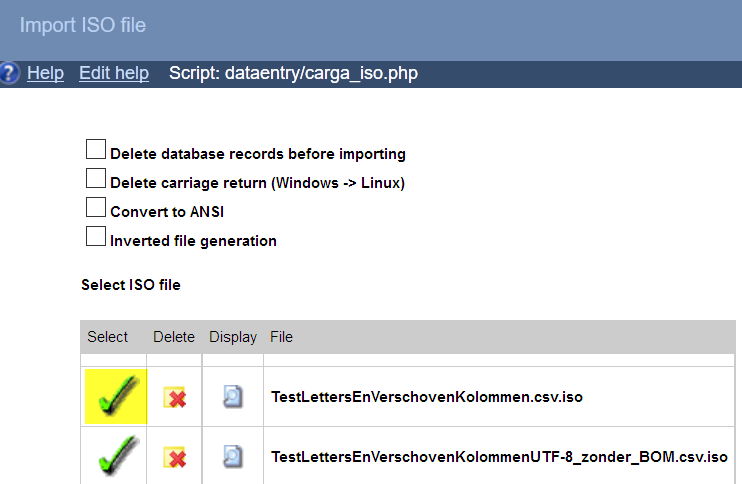
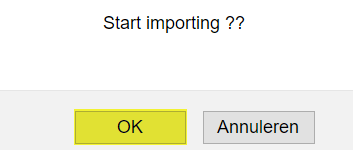
Bij fouten volgt een melding met prefix \*\*\*, status melding \*\*\* Program terminated with errors. \*\*\* en exit status 1. Een eventueel aanwezige ISO file is dan NIET bruikbaar.

De optie -S, voor het omzetten van de html tekens <> naar «» is experimenteel omdat de effecten voor de rest van het proces onduidelijk zijn.

## Importeren van ISO file in ABCD

1. Importeren in ABCD gaat met Utilities 🡺 Import ISO file  
   
2. De ISO file dient in de wrk folder (/opt/ABCD/www/bases/wrk) van ABCD te staan om geïmporteerd te kunnen worden. Dit kan op drie manieren:

* Door bij cnv\_csv\_to\_iso expliciet dit pad aan te geven als output.
* Op file niveau de file te verschuiven.
* In ABCD bij Import ISO file te Bladeren en Upload te kiezen.  
  

1. De selectie opties allemaal leeg laten !  
   Selecteer de te importeren file en daarna in de pop-up OK.  
    🡺 

Als terugkoppeling volgt een pagina met geïmporteerde records.

# Achtergrond informatie

## Teken sets

De moderne Excel werkt intern met UTF-8. ABCD werkt (nog) niet in UTF-8, maar de teken set voor ABCD is een subset van deze algemene set.

De volgende tekens (gescheiden door \_ om te zien of niets wegraakt), komen goed in de ABCD database terecht en worden daar ook correct getoond:  
032-063:\_ \_!\_""\_#\_$\_%\_&\_'\_(\_)\_\*\_+\_,\_-\_.\_/\_0\_1\_2\_3\_4\_5\_6\_7\_8\_9\_:\_;\_=\_?\_      ISO/IEC 8859-1, Windows-1252, HTML ANSI  
064-095:\_@\_A\_B\_C\_D\_E\_F\_G\_H\_I\_J\_K\_L\_M\_N\_O\_P\_Q\_R\_S\_T\_U\_V\_W\_X\_Y\_Z\_[\_\\_]\_^\_\_\_   ISO/IEC 8859-1, Windows-1252, HTML ANSI  
096-126:\_a\_b\_c\_d\_e\_f\_g\_h\_i\_j\_k\_l\_m\_n\_o\_p\_q\_r\_s\_t\_u\_v\_w\_x\_y\_z\_{\_|\_~\_       ISO/IEC 8859-1, Windows-1252, HTML ANSI  
128-159:\_€\_‚\_ƒ\_„\_…\_†\_‡\_ˆ\_‰\_Š\_‹\_Œ\_Ž\_‘\_’\_“\_”\_•\_–\_—\_˜\_™\_š\_›\_œ\_ž\_Ÿ\_             Windows-1252, HTML ANSI.  
160-191:\_¡\_¢\_£\_¤\_¥\_¦\_§\_¨\_©\_ª\_«\_¬\_®\_¯\_°\_±\_²\_³\_´\_µ\_¶\_·\_¸\_¹\_º\_»\_¼\_½\_¾\_¿\_       ISO/IEC 8859-1, Windows-1252, HTML ANSI  
192-223:\_À\_Á\_Â\_Ã\_Ä\_Å\_Æ\_Ç\_È\_É\_Ê\_Ë\_Ì\_Í\_Î\_Ï\_Ð\_Ñ\_Ò\_Ó\_Ô\_Õ\_Ö\_×\_Ø\_Ù\_Ú\_Û\_Ü\_Ý\_Þ\_ß\_   ISO/IEC 8859-1, Windows-1252, HTML ANSI  
224-255:\_à\_á\_â\_ã\_ä\_å\_æ\_ç\_è\_é\_ê\_ë\_ì\_í\_î\_ï\_ð\_ñ\_ò\_ó\_ô\_õ\_ø\_ù\_ú\_û\_ü\_ý\_þ\_ÿ\_       ISO/IEC 8859-1, Windows-1252, HTML ANSI

* ad 032-063: 60: < en 62: >. Komen beide wel in de ISO file en ABCD database.  
   Hier weggelaten omdat ze worden geïnterpreteerd door de browser.
* ad 128-159: Dit zijn in ISO 8859-1 controle tekens. Worden door ABCD wel goed weergegeven.
* ad 128-159: 129,141,143,144,157: Not Used. Weggelaten
* ad 160-191: 160: NBSP is een Non Breaking Space. Weggelaten.  
   173: SHY is een "Soft hyphen" en wordt alleen getoond als "-" als browser zelf  
   woorden af mag breken. Weggelaten.

## Lijstscheidingsteken

Het scheidingsteken is voor CSV files historisch gezien de komma (,).

In Excel is dit default ook de komma maar Excel wijkt daar zonder kennisgeving van af als er conflict is met het decimaal teken (een punt of een komma, afhankelijk van de ingestelde taal).

Configuratie (niet in Excel) via (Control panel 🡪 Region 🡪 Advanced 🡪 Set list separator).

Om eindeloos puzzelen te vermijden kan het CSV naar ISO programma automatisch het lijstscheidingsteken bepalen. Het test de volgende tekens ,;:|.~!@#$%^&-\_+=/\? en een tab.

## Excel versies

De procedure in dit document is gemaakt met Excel 2016. Formaten xlsx en xls zijn OK.

Windows kent ook een Excel App :   
Deze App niet gebruiken voor de conversie naar CSV (Gaat soms anders om met een “ en , in één regel + Maakt *altijd* UTF-8 + Gebruikt geen regio instellingen maar altijd een ,).

## Ontwerp beslissingen

In de historische werkwijze waren er restricties aan de Tags (alleen nummers, volgorde, niet dubbel). Door de re-design van de procedure zijn deze regels minder hard. Redenen om Titles en repeterende kolommen toe te staan zijn:

* Een naam voor een veld voor een gebruiker meer herkenbaar dan nummers.
* De interface naar Greenstone via een CSV file gebruikt namen voor kolommen. Door namen ook toe te staan kunnen deze files met minder moeite gebruikt worden te testen en om ook in te lezen in ABCD.
* Bij groepen gelijksoortige records kunnen repeterende velden soms een “standaard” waarde hebben. Door ook repeterende kolommen toe te staan kan dit gebruikt worden om de Excel sneller te vullen.
* De interface naar Greenstone via een CSV file gebruikt repeterende kolommen voor repeterende velden. Door repeterende kolommen toe te staan kunnen deze files met minder moeite gebruikt worden om te testen en in te lezen in ABCD.

Er zijn geen controles op de waardes (lengte, picklist, datum formaat,…) ingebouwd. Dit zou verdubbeling zijn van de code in ABCD. ABCD lijkt die restricties ook als aanwijzing te beschouwen en dwingt weinig af. De ISO interface heeft in ABCD 1.4 geen enkele restrictie.

Voor repeterende velden worden 3 mogelijkheden ondersteund. Die kunnen door elkaar en gelijktijdig gebruikt worden.

* In één cel gescheiden door een semicolon “;”.  
  Deze methode is ontstaan bij het scannen van plaatjes waarbij de CSV file door ABCD ingelezen wordt.
* Door voor iedere term een aparte kolom te nemen.  
  Deze methode is ontstaan bij de productie van CSV files door ABCD voor Greenstone. De methode wordt ondersteund om deze files te kunnen testen door in te lezen in ABCD.
* In één cel als verschillende regels. Dit levert embedded linefeeds in de CSV file. Deze methode wordt ondersteund omdat dit een de-facto standaard is (ondersteund door Excel). Zie:  
  <https://en.wikipedia.org/wiki/Comma-separated_values>  
  <http://www.creativyst.com/Doc/Articles/CSV/CSV01.htm>

## Portabiliteit / flexibiliteit van cnv\_csv\_to\_iso

De source code bevindt zich in 1 file: cnv\_csv\_to\_iso.c.

Afhankelijkheden van een OS en ontwikkel omgeving zijn bewust vermeden.

* Getest: Compilatie op Linux met standaard omgeving
* Getest: Compilatie op Windows met behulp van MinGW en gcc.  
  Deze versie kan op iedere PC met Windows draaien (dus zonder MinGW compilatie omgeving).
* Getest: Compilatie op Windows met behulp van cygwin en gcc.  
  Deze versie heeft de cygwin run-time bibliotheken nodig en draait alleen in een cygwin shell.
* Niet getest: Compilatie op Windows met Visual C

De CSV files mogen gemaakt worden op Linux (LF), Windows (CRLF) en Mac (CR).   
De geproduceerde ISO files hebben een regeleindes die afhankelijk zijn van het OS waar cnv\_csv\_to\_iso draait.