csv2siard v.1.8.7   
*Anwendungshandbuch*

**Inhalt**

[1 Programmbeschreibung 2](#_Toc343787082)

[2 cvs2siard installieren 3](#_Toc343787083)

[3 csv2siard konfigurieren 3](#_Toc343787084)

[4 Beispiel: GV-Daten in eine SIARD konvertieren 5](#_Toc343787085)

[5 Beliebige CSV-Dateien in eine SIARD-Datei konvertieren 6](#_Toc343787086)

[6 Präferenzen 7](#_Toc343787087)

[7 Konsolenausgabe 8](#_Toc343787088)

[8 Konvertierung von CSV zu Datenbankfeldern 9](#_Toc343787089)

[9 Unterstützte Datumformate 10](#_Toc343787090)

[10 CSV via ODBC 11](#_Toc343787091)

[11 Einfaches GUI zu csv2siard 19](#_Toc343787097)

[12 Installierte Dateien 20](#_Toc343787098)

# Programmbeschreibung

Das Tool **csv2siard** erlaubt die Konvertierung von CSV-Dateien in eine SIARD-Datei[[1]](#footnote-1). Der Vorteil einer solchen Konvertierung ist mehrfach. Erstens werden einzelne CSV-Dateien, die zusammen eine Sammlung bilden, in einer Datei zusammengefasst; zweitens werden die CSV-Dateien in ein standardisiertes Format gebracht und somit unterschiedliche CSV-Sammlungen bezüglich Zeichensatz, Datentrennzeichen, Zeilenstruktur etc. vereinheitlicht ; drittens steht mit SiardEdit[[2]](#footnote-2) ein frei erhältlicher Viewer für SIARD-Dateien zur Verfügung; und viertens ist auch bei grossen Datenmengen zur Datenanalyse ein Export in eine relationale Datenbank problemlos möglich.

**csv2siard** ist ein einfaches Kommandozeilen-Tool, das CSV-Dateien in Tabellen innerhalb einer SIARD-Datei umwandelt. Jede Datei wird zu einer Tabelle. Da bei CSV-Dateien keine Strukturinformationen im eigentlichen Sinne zur Verfügung stehen, generiert das Tool eine einfache Tabellenbeschreibung mit Feldnamen und Feldattribut für jede Datei in einem XML-Datenmodell. Das Datenmodell basiert auf dem Apache Torque 4.0 Standard[[3]](#footnote-3). Die Tabellen werden ohne relationale Abhängigkeiten und Feldeinschränkungen *(Constraints)* erzeugt. Das Datenmodell kann aber anschliessend manuell bearbeitet und mit zusätzlichen Datenbankinformationen aus externen Quellen (relationale Beziehungen, Feldeinschränkungen etc.) versehen werden In einem zweiten Durchgang kann dann dieses Datenmodell verwendet und damit zu den Tabellen in SIARD auch ein relationales Datenmodell gespeichert werden.[[4]](#footnote-4)

Steht für eine CSV-Datensammlung bereits ein Datenmodell zur Verfügung, z.B. weil die CSV-Dateien auf Grund einer solchen Spezifikation aus einer Datenbank exportiert worden sind, kann bei der Konvertierung diese Datenbankbeschreibung verwendet werden. **csv2siard** prüft in diesem Falle die Feldnamen, Feldattribute und Spaltenzahlen in den einzelnen Dateien vor der Konvertierung. Nicht geprüft werden in dieser Version spezifisch relationale Aspekte wie *Unique Constraints* und *Foreign Key Constraints*.

Zur Veranschaulichung sind aus dem KOST-*Projekt "Archivierung von Gebäudeversicherungsdaten"[[5]](#footnote-5)* das Datenmodell ***gv-model-v9.xml*** und eine kleine anonymisierte Testdatensammlung von CSV-Dateien im Ordner ***csvdata*** beigelegt. Zusätzlich ist auch eine Testsammlung zur Veranschaulichung von unterschiedlichen Datenfeldern mit dem Datenmodell ***datatype-model.xml*** und den Dateien in **datatype*[[6]](#footnote-6)*** beigelegt.

Der Vollständigkeit halber ist der Source Code in PHP ebenfalls beigelegt. Das ausführbare Programm ist mit Bamcompile[[7]](#footnote-7) kompiliert. **csv2siard** benötigt zusätzlich die Programme **7z.exe**, **file.exe** und **xmllint.exe**. Diese Programme sind Freeware, bitte beachten Sie jedoch die jeweiligen Urheberrechtsbestimmungen.

# cvs2siard installieren

|  |  |
| --- | --- |
| **2a** | csv2siard-Arbeitsverzeichnis erstellen  (zum Beispiel Ordner **csv2siard** im Verzeichnis **C:\Programme** oder **C:\Software**) |

|  |  |
| --- | --- |
| **2b** | **csv2siard.zip** herunterladen und in das Arbeitsverzeichnis  **C:\Software\csv2siard** entpacken.    Der Pfad zum ausführbaren Programm lautet anschliessend **C:\Software\csv2siard\bin\csv2siard.exe** |

# csv2siard konfigurieren

|  |  |
| --- | --- |
| **3a** | CSV-Dateien bereitstellen (z.B. im Verzeichnis **C:\Programme\csv2siard\csvdata**)    Die Dateinamen müssen den Einschränkungen der gewählten **FILE\_MASK** entsprechen; siehe Kapitel 6, Präferenzen. |
| **3b** | Ein Kommandozeilenfenster öffnen in Windows XP oder Windows 7: |

|  |  |
| --- | --- |
| **3c** | In das gewünschte Arbeitsverzeichnis wechseln, hier z.B. mit **CD /D C:\Programme\csv2siard** |

|  |  |
| --- | --- |
| **3d** | Tool starten und Usage / Help / Version anzeigen lassen. Der Pfad zum ausführbaren Programm lautet  **C:\Programme\csv2siard\bin\csv2siard.exe**  Besser lesbar:  **C:\Programme\csv2siard> bin\csv2siard.exe**  **Usage :: csv2siard.exe database csvpath siardfile [prefs]**  **database :: database description according to torque.v4**  **XML model or keyword :NO\_DB\_MODEL**  **csvpath :: path where to find the csv files or keyword :ODBC**  **siardfile :: SIARD file to be created**  **prefs :: configuration file (default preferences.prefs)**  **version :: 1.8** |
| **3e** | Präferenzen für die Konvertierung CSV -> SIARD festlegen. Die voreingestellten Werte werden in der Regel korrekt sein, siehe dazu unten Kapitel 6, Präferenzen.  **notepad C:\Programme\csv2siard\bin\preferences.prefs[[8]](#footnote-8)**    Wichtig sind vor allem die korrekten Einstellungen für diese fünf Werte:  **DELIMITED=;**  **QUOTE="**  **COLUMN\_NAMES=TRUE  CHARSET=ISO-8859-1**  **FILE\_MASK=\*.dat** |

# Beispiel: GV-Daten in eine SIARD konvertieren

|  |  |
| --- | --- |
| **4** | **csv2siard** erwartet als Argumente eine Datei mit der Datenbankbeschreibung in XML, den Pfad zu den CSV-Dateien und einen Namen für die neu anzulegende SIARD-Datei, optional kann eine andere Präferenzdatei gewählt werden.  Die Datenbankbeschreibung für GV-CSV-Dateien **gv-model-v9.xml** wird beim Installieren des Tool gleich angelegt:  **bin\csv2siard.exe gv-model-v9.xml csvdata new.siard** |

# Beliebige CSV-Dateien in eine SIARD-Datei konvertieren

|  |  |
| --- | --- |
| **5** | **csv2siard** kann auch ohne Datenmodell ein Set von CSV-Dateien in eine SIARD-Datei konvertieren. Mit der Option **:NO\_DB\_MODEL** wird ein einfaches Datenmodell **no\_db\_model.xml** für die mit der Option **FILE\_MASK** in der Präferenzdatei ausgewählten CSV-Dateien angelegt.  Die SQL-Namenskonvention muss bei der Vergabe der Dateinamen und bei den Spaltennamen beachtet werden.[[9]](#footnote-9) Im Fehlerfall werden Spaltennamen automatisch in Namen vom Typ **column…** konvertiert.  Die Option **CHECK\_COLUMN=FALSE** in der Präferenzdatei erlaubt auch die Konvertierung von durch MS-Excel erzeugten CSV-Dateien mit unterschiedlicher Spaltenzahl:  **bin\csv2siard.exe :NO\_DB\_MODEL csvdata new.siard** |

Bei mit **MS-Excel** erstellten CSV-Dateien kann es vorkommen, dass die Zeilen eine unterschiedliche Spaltenanzahl haben. Um diese Dateien trotzdem konvertieren zu können, muss in der Präferenzdatei die Option **CHECK\_COLUMN=FALSE** definiert sein.

# Präferenzen

|  |  |
| --- | --- |
| **6** | ***# Default*** ***preferences***  CHARSET (default 'ISO-8859-1') [[10]](#footnote-10) # character-set (US-ASCII, ASCII,   # OEM, ANSI, ISO-8859-1 and UTF-8)  COLUMN\_NAMES (default true) # First row contains column names  DELIMITED (default ';') # CSV column separator  QUOTE (default '"') [[11]](#footnote-11) # Optional field quotation  FILE\_MASK (default '\*.dat') # Wild card is replaced with table name  # or is converted to tablename  CHECK\_COLUMN (default true) [[12]](#footnote-12) # Check column count,   # not applicable with MS-Excel CSV  CHECK\_NAMES (default true) [[13]](#footnote-13) # Check column names in first row  CHECK\_DATABASE\_INTEGRITY (default false) # Not implemented yet  DATE\_FORMAT (default settings) # Special date format string  PI\_COUNT (default '100') # Progress indicator per line processed  TMPDIR (default System tempdir) # default temp dir  UNICODE\_EXTENDED (default false)[[14]](#footnote-14) # Convert non UNICODE character  VERBOSITY (default false) # Display additional messages  ***# Optional content settings***) [[15]](#footnote-15)  ARCHIVED\_BY (default empty) # Database archived by  CONTACT (default empty) # Archivist's contact details  DB\_TYPE (default 'CSV') # Type of Database or database product  DESCRIPTION (default empty) [[16]](#footnote-16) # Database description  OWNER (default '(...)') # Data owner prior to archiving  SIARD\_SCHEMA (default 'schema0') # default schema  SIARD\_USER (default 'admin') [[17]](#footnote-17) # default user  TIMESPAN (default '(...)') # Data creation time span  ***# ODBC settings***  ODBC\_DSN # Database source name for the connection  ODBC\_USER # Database user name  ODBC\_PASSWORD # Database password |

**Achtung:** es findet keine Zeichensatzkonvertierung statt, wenn ein falscher Zeichensatz mit **CHARSET** spezifiziert wird – der vermutete Zeichensatz wird aber angezeigt.

# Konsolenausgabe

|  |  |
| --- | --- |
| **7** | Die Konsolenausgabe zeigt zuerst den Copyright-Hinweis und mit der Option **VERBOSITY** die für diese Konvertierung gesetzten Präferenzen.  Mit der Option **:NO\_DB\_MODEL** wird anschliessend eine Kurzfassung des erstellten Datenmodells angezeigt.  Die eigentliche Konvertierung wird für jede CSV-Datei zusammen mit dem ermittelten Zeichensatz gesondert angezeigt.  Mit **VERBOSITY** wird am Schluss der eigentliche Aufbau der SIARD-Datei als ZIP-Datei angezeigt. |

Die **encoding** Angaben sind eine Vermutung, die bei der Option **:NO\_DB\_MODEL** durch eine Analyse der Tabellen ermittelt wird; es kann hier aber zu Fehlern kommen. Deshalb erfolgt die Konvertierung von CSV-Daten zu SIARD einzig aufgrund der Präferenz **CHARSET** (default ISO-8859-1). Gewisse Zeichensatzkonvertierungen sind implizit, z.B. US-ASCII zu ISO-8859-1, siehe die Fussnote zu **CHARSET** weiter oben.

# Konvertierung von CSV zu Datenbankfeldern

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **CSV Sample Daten** | **Typenprüfung & Konvertierung** | **Torque 4.0** | **XML** | **SQL-99** |
| 127 | ctype\_digit | TINYINT | xs:integer | INTEGER |
| -232767 | ctype\_digit | SMALLINT | xs:integer | INTEGER |
| -2147483647 | ctype\_digit | INTEGER | xs:integer | INTEGER |
| 2147483647 | ctype\_digit | BIGINT | xs:integer | INTEGER |
| 345.6789 | is\_numeric | FLOAT | xs:float | FLOAT |
| 1.23457E+15 | is\_numeric | REAL | xs:float | FLOAT |
| 1.23457E+22 | is\_numeric | DOUBLE | xs:float | FLOAT |
| 1234567891 | is\_numeric | NUMERIC | xs:decimal | NUMERIC |
| 12345678.25 | is\_numeric | DECIMAL | xs:decimal | NUMERIC |
| A | xml\_encode | CHAR | xs:string | CHARACTER VARYING |
| ABV | xml\_encode | VARCHAR | xs:string | CHARACTER VARYING |
| Victor jagt zwölf Boxkämpfer quer über den Sylter Deich | xml\_encode | LONGVARCHAR | xs:string | CHARACTER VARYING |
| 2003-12-31 | convert2XMLdate | DATE | xs:date | DATE |
| 01:02:03 | convert2XMLdate | TIME | xs:time | TIME |
| 2003-12-31T01:02:03 | convert2XMLdate | TIMESTAMP | xs:dateTime | TIMESTAMP |
| 00011011[[18]](#footnote-18) | bit->hex | BIT | xs:hexBinary | BIT |
| PK□□[[19]](#footnote-19) | bin->hex | BINARY | xs:hexBinary | BIT VARYING |
| VGhpcyBpcyBhbiBlbmNvZGVkIHN0cmluZw==[[20]](#footnote-20) | base64->hex | VARBINARY | xs:hexBinary | BIT VARYING |
| R0lGODlhDAAKAJEAAP///3N1B1FRUQAAACwAAAAADAAKAAACGpSPB8ttDcELNE5Ac5ACVww+ESOOnLkkqlEAADs= | base64->hex | LONGVARBINARY | xs:hexBinary | BIT VARYING |
| R0lGODlhDAAKAJEAAP///3N1B1FRUQAAACwAAAAADAAKAAACGpSPB8ttDcELNE5Ac5ACVww+ESOOnLkkqlEAADs= | base64->hex | BLOB | xs:hexBinary | BLOB |
| The quick brown fox jumps over the lazy dog | xml\_encode | CLOB | xs:string | CHARACTER VARYING |
| http://ch.php.net/manual/en/function.base64-decode.php |  | REF | xs:string | CHARACTER VARYING(255) |
| TRUE | to\_bool | BOOLEANINT | xs:boolean | BOOLEAN |
| FALSE | to\_bool | BOOLEANCHAR | xs:boolean | BOOLEAN |

**Achtung:** Nicht alle Torque- und SQL99-Datentypen werden erkannt und unterstützt.  
Die Option **:NO\_DB\_MODEL** kann keine CSV-Dateien mit binären Feldern erkennen und bearbeiten. Felder mit binärem Inhalt müssen manuell im Datenmodell eingetragen werden. Uncodierte binäre Datenfelder vom Type BINARY dürfen keine CSV-Delimiter-Zeichen und „neue Zeile“-Zeichen enthalten.

# Unterstützte Datumformate

|  |  |  |
| --- | --- | --- |
| **DATE Standard** | **Datumformat** | **Beispiel** |
| Preference: DATE\_FORMAT | Format string nach PHP strftime() [[21]](#footnote-21) |  |
| non-standard | YY MM DD hh ii ss | "20080701223807" |
| XMLRPC (Compact) | YY MM DD "t" hh ii ss | "20080701t223807" or "20080701T093807" |
| XMLRPC | YY MM DD "T" hh ":" ii ":" ss | "20080701T22:38:07" "20080701T9:38:07" |
| EXIF | YY ":" MM ":" DD " " hh ":" ii ":" ss | "2008:08:07 18:11:31" |
| MySQL | YY "-" MM "-" DD " " hh ":" ii ":" ss | "2008-08-07 18:11:31" |
| WDDX | YY "-" MM "-" dd "T" hh ":" ii ":" ss | "2008-7-1T9:3:37" |
| ISO 8601/SOAP | YY "-" MM "-" DD "T" hh ":" ii ":" ss | "2008-07-01T22:35:17.02" "2008-07-01T22:35:17.03+08:00" |
| Common Log Format | D "/" M "/" YY : hh ":" ii ":" ss " "   tz correction | "10/Oct/2000:13:55:36 -0700" |
| MS-Excel non standard (DE) | DD "." MM "." YY " " hh ":" ii ":" ss | "01.07.2008 09:03:37" |
| UNIX date format |  | "Sat Nov 04 12:02:33 EST 1989"  "now" "epoche" |

# CSV via ODBC

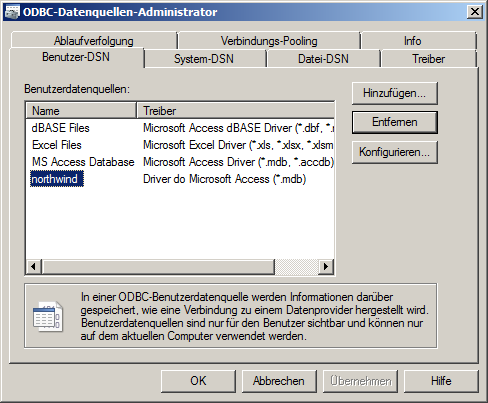
CSV-Dateien können auch via ODBC[[22]](#footnote-22) angesprochen werden. Eine Microsoft ODBC-Datenquelle wird in Form einer DSN *(Datasource Name)* via Systemsteuerung > Verwaltung > Datenquellen (ODBC) als Benutzer-DSN oder als System-DSN eingerichtet. Alternativ ist auch die direkte Angabe eines *ODBC Connection Strings* möglich. Neben Text-Tabellen können natürlich auch andere ODBC-Quellen (z.B. Excel oder MS-Access) angesprochen werden.

Da ODBC *(Open Database Connectivity)* als standardisierte Datenschnittstelle SQL als Abfragesprache verwendet, steht die volle Mächtigkeit dieser Sprache bei der Datenprüfung, Datenkonvertierung und Datenmodellierung zur Verfügung. Da ODBC inzwischen auch ausserhalb der Microsoft-Welt ein Standard ist und einen entfernten *(remote)* Datenzugriff erlaubt, können auch Daten von Datenbankservern in SIARD-Format umgewandelt werden.

Zum Testen sind drei ODBC-Datenquellen beigelegt, die CSV-Quellen im Ordner ***odbcdata,*** die MS-Excel-Mappe ***demo.xls*** und dieMS-Access-Datenbank ***demo.mdb.*** Es sind dies die gleichen anonymisierten Testdaten aus dem KOST-Projekt *„Archivierung von Gebäudeversicherungsdaten“*, wie sie weiter oben schon Verwendung finden.

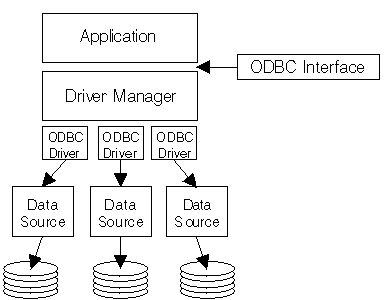
## SIARD-Konvertierung via ODBC

Drei zusätzliche Parameter (**ODBC\_DSN**, **ODBC\_USER** und **ODBC\_PASSWORD)** in der Präferenzdatei sind für die Konfigurierung einer ODBC-Verbindung notwendig.

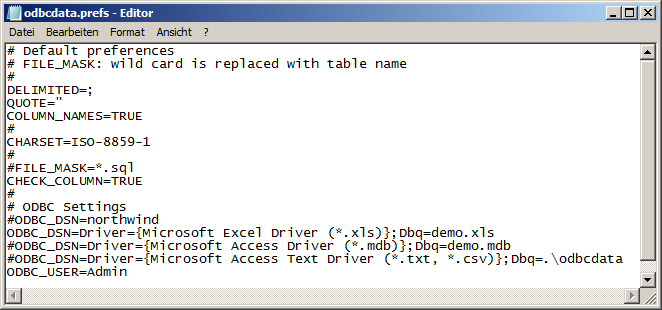
Der Parameter **ODBC\_DSN** kann entweder einen DSN *(Datasource Name)* oder einen *ODCB Connection String* enthalten; **ODBC\_USER** und **ODBC\_PASSWORD** sind selbsterklärend und bei ODBC Text- und Excel-Quellen nicht notwendig.

Ein DSN *(Datasource Name)* wird mit dem   
**ODBC-Datenquellen-Administrator** Tool, das sich bei Windows XP / Windows 7 in der Systemsteuerung > Verwaltung > Datenquellen (ODBC) befindet, eingerichtet. Je nach Berechtigungslevel können Benutzer-DSN oder System-DSN eingerichtet werden.

Beispiel für ein Benutzer-DSN: **ODBC\_DSN=northwind**

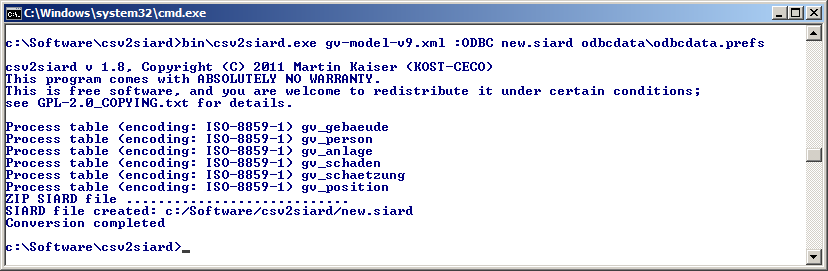
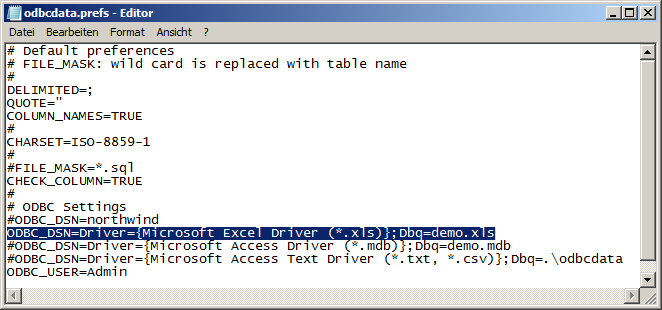


Verbindung mit einem *ODCB Connection String:* **ODBC\_DSN=Driver={Microsoft Access Driver (\*.mdb)};Dbq=demo.mdb [[23]](#footnote-23)  
 ODBC\_DSN=Driver={Microsoft Access Text Driver (\*.txt, \*.csv)};Dbq=C:.\odbcdata\  
 ODBC\_DSN=Driver={Microsoft Text Driver (\*.txt; \*.csv)};Dbq=.\odbcdata  
 ODBC\_DSN=Driver={Microsoft Excel Driver (\*.xls)};Dbq=demo.xls**



In **odbcdata/odbcdata.prefs** sind die entsprechenden Parameter bereits eingetragen.

Die Auswahl der in der SIARD-Datei zu übernehmenden Tabellen und Felder erfolgt über das XML-Datenmodell. Wird beim Ausführen von **csv2siard.exe** statt des Laufwerkpfads **csvpath** das Schlüsselwort **:ODBC** gewählt, wird für jede Tabelle im Datenmodell die folgende SQL Query **SELECT \* FROM TABLENAME** ausgeführt. **DELIMITED** und **QUOTE** sind ohne Bedeutung, hingegen bestimmt **COLUMN\_NAMES=TRUE**, dass die Spaltennamen der ODBC-Quelle mit dem Datenmodell übereinstimmen müssen, andernfalls wird nur die Spaltenreihenfolge beachtet. Da bei einer ODCB-Datenquelle der Zeichensatz nicht via Datenverbindung ermittelt werden kann, muss **CHARSET** ebenfalls richtig gesetzt werden.



In diesem Beispiel konvertieren wir die Tabellen in der Excel-Mappe **demo.xls** in eine SIARD-Datei.

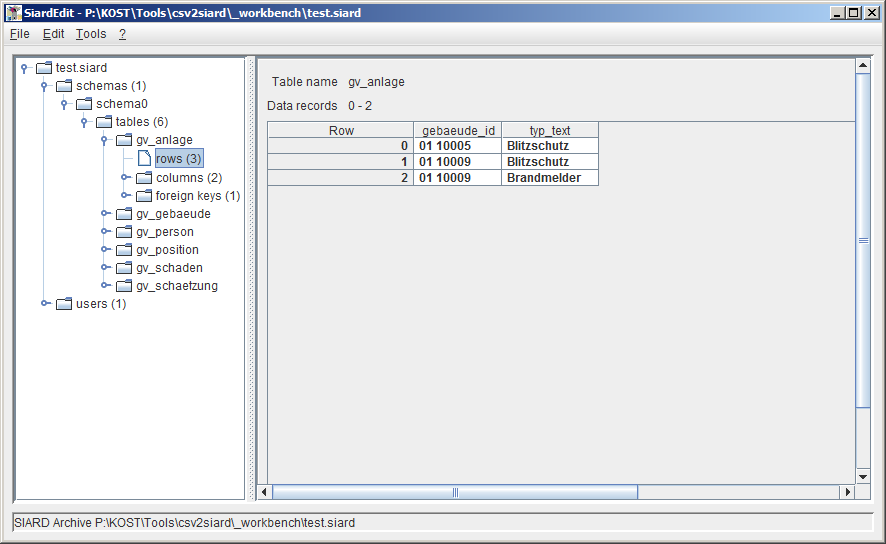
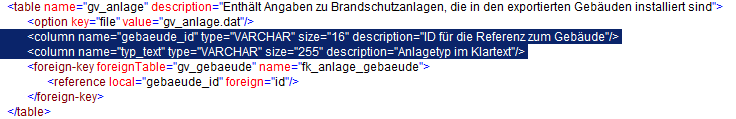
## Ausgewählte Spalten übernehmen

Wird eine ODBC-Datenquelle verwendet, können mit Hilfe des Datenmodells auch einzelne Spalten aus den Ursprungstabellen ausgewählt und in die neue SIARD-Datei übertragen werden. Das funktioniert mit der Präferenzeinstellung **COLUMN\_NAMES=TRUE** und einem entsprechenden Datenmodell.

Im Beispiel sollen aus der Tabelle/Datei **gv\_anlage.csv** nur die Spalten **gebaeude\_id** und **typ\_text** übernommen:

gv\_anlage.csv





gv-model.xml

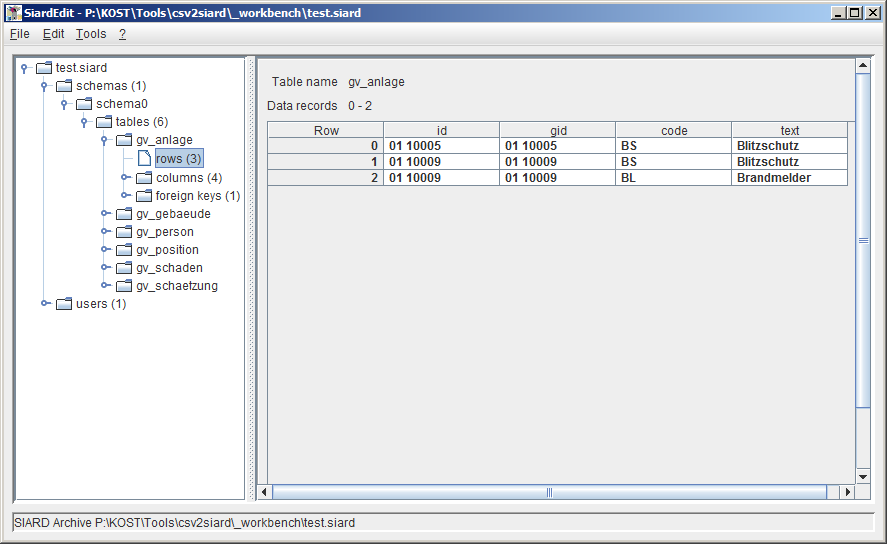
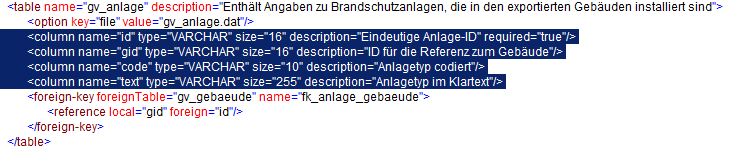
## Spalten umbenennen

Wird eine ODBC-Datenquelle und die Präferenzeinstellung **COLUMN\_NAMES=FALSE** verwendet, werden die Spalten der CSV-Tabelle/Datei von links nach rechts an die Datenfelder im Datenmodell gebunden, eine Feldnamenprüfung findet nicht statt. Damit ist es möglich, den Feldern via Datenmodell neue Feldnamen zu zuweisen.

Im Beispiel werden die Spalten in der Tabelle/Datei **gv\_anlage.csv** in **id,** **gid,** **code** und **text** geändert.

gv\_anlage.csv



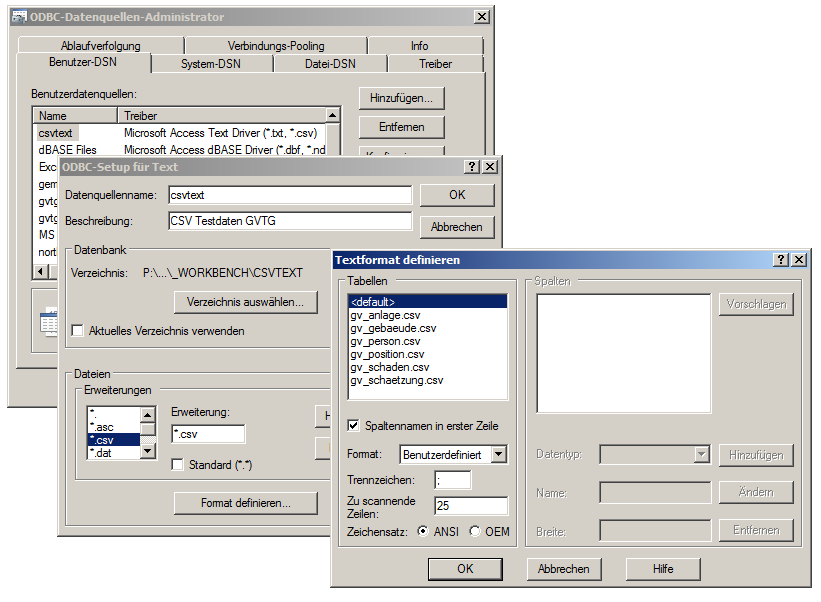
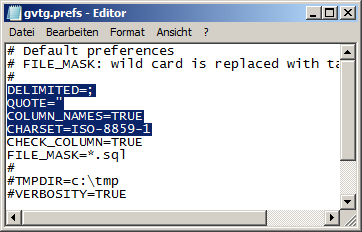


gv\_model.xml

## ODBC-Text-Datenquelle

Mit dem *Microsoft Access Text Treiber* ist es auch möglich, CSV-Dateien via **ODBC** anzusprechen und damit die volle Mächtigkeit der SQL-Abfragesprache bei der Umformung oder Auswahl der Daten zu nutzen.

Einige Punkte sind zu beachten beim Anlegen einer solchen Datenquelle:  
Alle CSV-Dateien müssen im gleichen Verzeichnis sein und zwingend die Endung **.txt** oder **.CSV**[[24]](#footnote-24) haben.   
Wichtig ist auch, dass beim Anlegen einer ODBC-Text-Datenquelle mit dem   
**ODBC-Datenquellen-Administrator** Tool Trennzeichen und Zeichensatz[[25]](#footnote-25) richtig und so wie in der csv2siard Präferenzdatei definiert gesetzt werden   
(Zeichensatz **ANSI** ist gleichbedeutend mit **ISO-8859-1** und **OEM** gleichbedeutend mit **extended ASCII**).



Nach dem Anlegen einer Text DSN *(Datasource Name)* liegt im gewählten Verzeichnis eine Datei**schema.ini[[26]](#footnote-26)**, dort sind die einzelnen Dateien/Tabellen und Felder beschrieben:

**. . .**

**[gv\_anlage.csv]**

ColNameHeader=True

**Achtung:** Fehlerhafte Feldtypenbeschreibung in der schema.ini Datei kann zu fehlerhafter Datenübernahme von ODBC Text-Datenquellen führen, siehe dazu weiter unten

Format=Delimited(;)

MaxScanRows=25

CharacterSet=ANSI

Col1=ID Integer

Col2=GEBAEUDE\_ID Char Width 16

Col3=TYP\_CODE Char Width 10

Col4=TYP\_TEXT Char Width 255

**[gv\_gebaeude.csv]**

ColNameHeader=True

**. . .**

Im Prinzip kann diese Datei auch mit einem Texteditor angelegt werden. Wird keine Feldbeschreibung angegeben, wird der Feldtyp vom ODBC Treiber selbst ermittelt.

Ist das Verzeichnis **odbcdata** ist schon mit einer **schema.ini** Datei entsprechen konfiguriert, können wir auch ohne DSN mit einem entsprechenden *ODCB Connection String*

**ODBC\_DSN=Driver={Microsoft Access Text Driver (\*.txt, \*.csv)};Dbq=C:.\odbcdata\**oder  
 **ODBC\_DSN=Driver={Microsoft Text Driver (\*.txt, \*.csv)};Dbq=C:.\odbcdata\**auf die CSV-Dateien im Verzeichnis **odbcdata** zugreifen

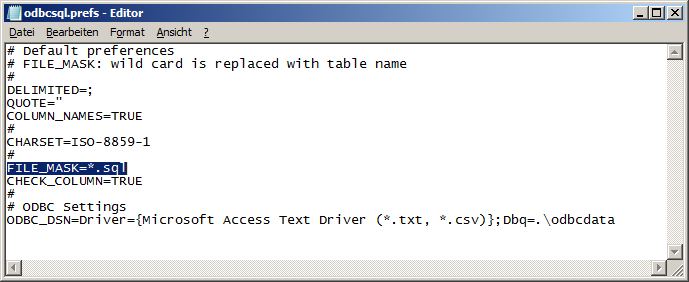
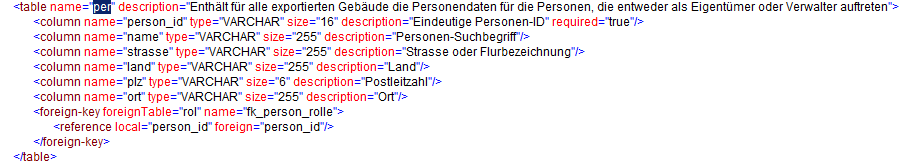
## Erweiterte ODBC-Unterstützung

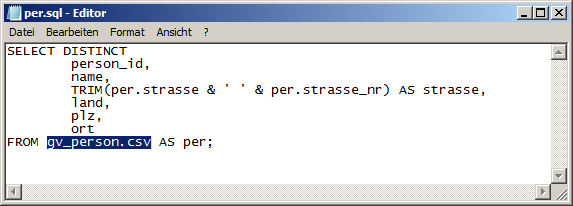
Im Gegensatz zur direkten Konvertierung von CSV-Dateien besteht bei der Konvertierung über eine ODBC-Verbindung mit Hilfe der Abfragesprache SQL aber eine noch weit grössere Freiheit bei der Umformung oder Auswahl der Daten.

Wird anstelle des Schlüsselwortes **ODBC** für **csvpath** ein Verzeichnis gewählt, werden in diesem Verzeichnis alle Dateien nach den im Datenmodell angegebene Tabellenamen mit der Präferenzeinstellung **FILE\_MASK** ausgewählt (wie bei der Auswahl von CSV-Dateien) und der in diesen Dateien gefundene SQL-Befehl auf der ODBC-Datenquelle ausgeführt. Der so erzeugte ODBC-Datenstrom wird in die entsprechende SIARD-Tabelle eingefügt. Damit ist es möglich, beliebige, komplexe Abfragen und die daraus generierten Tabellen in SIARD zu speichern.

*Ein Beispiel:*Die CSV-Tabellen im Verzeichnis **odbcdata** werden normalisiert, d.h. weil jede Person in **gv\_person** auch sowohl Verwalter wie auch Eigentümer eines Gebäudes in **gv\_gebaeude** sein kann (M:N-Beziehung), wird **gv\_person** via die neue Zwischentabelle **rol** verknüpft. Im gleichen Zug werden auch noch Vereinfachungen am Datenmodell vorgenommen, d.h. es werden die Codewert-Spalten entfernt und in Person die Felder **strasse** und **strasse\_nr** zusammengeführt. Das beigelegte Datenmodell **gv-model-nf.xml** ist die Grundlage dieser Transformation, die einzelnen SQL Abfragen für die neuen Tabellen befinden sich im Verzeichnis **odbcsql**. Wir sehen, dass dort auch eine Datei **gv\_rolle.sql** für die neue Tabelle **gv\_rolle** vorhanden sein muss. Zu Demonstrationszwecken werden alle Tabellennamen auf drei Buchstaben reduziert, also **gv\_person** zu **per**.

gv\_model-nf.xml

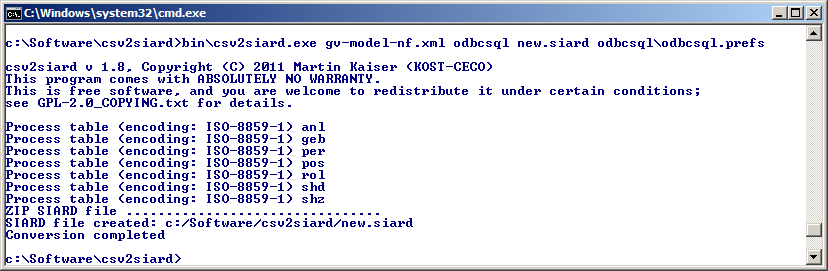
**

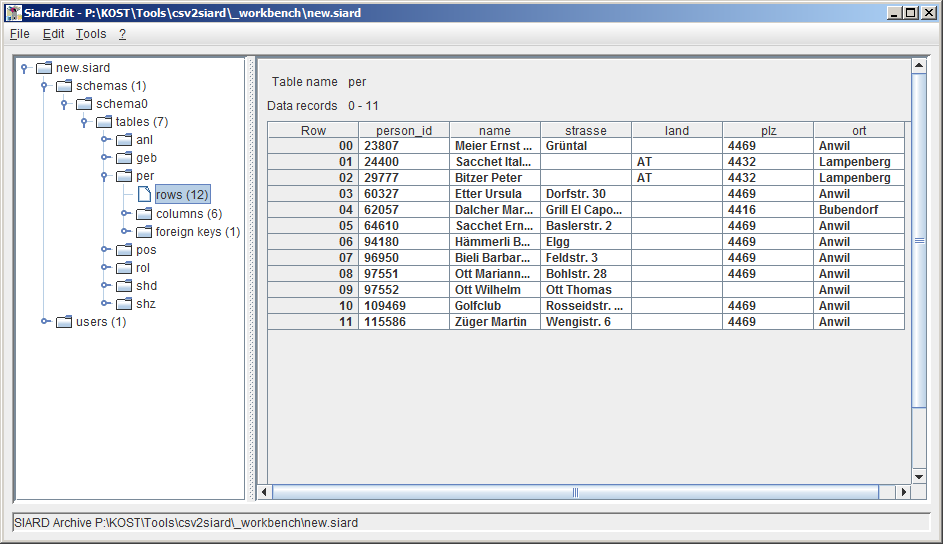


***Achtung:*** Tabellen in einer ODBC-Text-Quelle haben als Namen den vollständigen Dateinamen mit Datei-Extension, also im Beispiel **gv\_anlage.csv**.  
In einer **ODBC**-Excel-Quelle muss ein **$**-Zeichen zum Mappennamen hinzugefügt werden: **gv\_anlage$**

Wir starten die Konvertierung im Ordner **C:\software\csv2siard** wie folgt:

**bin\csv2siard.exe gv-model-nf.xml odbcsql new.siard odbcsql\odbcsql.prefs**





## schema.ini für ODBC Textquellen

Leider ermittelt das **ODBC-Datenquellen-Administrator** Tool die Feldtypen nicht immer richtig bei einer **schema.ini** Datei ohne Feldbeschreibung bzw. die vorgeschlagenen Feldtypen stimmen nicht wirklich, auch nicht innerhalb der angegebenen Scantiefe.

Im Beispiel enthält die CSV Datei **gv\_gebaeude.csv** im Feld   
**strasse\_nr** in zwei Zeilen keine Zahl sondern Text.

**Feldtypeninkonsistenz zwischen schema.ini und Datentabellen führt dazu, dass Datensätze fehlerhaft übernommen werden.**

* Deshalb müssen entweder die vorgeschlagenen Feldtypen sehr genau manuell überprüft werden.
* Oder es wird für alle Felder der allgemeine Typ **CHAR**[[27]](#footnote-27) verwendet und die Typenumwandlung erst durch ein entsprechendes XML Datenmodell bewirkt. **csv2siard** überprüft dabei die entsprechenden Formatkonvertierung für jede Feld.
* Drittens, und das ist wohl der sicherste Weg, wird das beigelegte Tool   
  **csvschema.exe** verwendet. Das Tool analysiert jede Spalte zeilenweise und ermittelt den bestmöglichen Datentyp in der Reihenfolge **INTEGER, FLOAT, DATETIME, CHAR**. Siehe anschliessend

## csvschema.exe

Das Hilfsprogramm generiert eine CSV Textquellen **schema.ini** Datei für die ausgewählten CSV Dateien im Ordner **csvpath**. Die Auswahl und Spezifizierung der CSV Dateien (Dateiendung, Zeichensatz, Trennzeichen etc.) erfolgt in der Präferenzdatei **prefs**. Die **schema.ini** Datei wird in den Ordner **csvpath** geschrieben

**bin\csvschema.exe**

**Usage :: csvschema.exe csvpath [prefs]**

**csvpath :: path where to find csv files**

**prefs :: configuration file (default) preferences.prefs**

Das Hilfsprogramm wird also ganz ähnlich wie **csv2siard** aufgerufen:

**bin\csvschema.exe csvdata csvdata\csvdata.prefs**

**......**

**New XML database model written: C:/Users/U80789~1/AppData/Local/Temp/no\_db\_model.xml**

**[gv\_anlage] => P:/KOST/Tools/csv2siard/\_workbench/csvdata/gv\_anlage.dat**

**[gv\_gebaeude] => P:/KOST/Tools/csv2siard/\_workbench/csvdata/gv\_gebaeude.dat**

**[gv\_person] => P:/KOST/Tools/csv2siard/\_workbench/csvdata/gv\_person.dat**

**[gv\_position] => P:/KOST/Tools/csv2siard/\_workbench/csvdata/gv\_position.dat**

**[gv\_schaden] => P:/KOST/Tools/csv2siard/\_workbench/csvdata/gv\_schaden.dat**

**[gv\_schaetzung] => P:/KOST/Tools/csv2siard/\_workbench/csvdata/gv\_schaetzung.dat**

**New CSV schema.ini written: csvdata/schema.ini**

## odbcheck.exe

Das beigelegte Tool **odbcheck.exe** erlaubt es, auf einfachem Weg eine ODBC Datenquelle zu testen und zu befragen.

**bin\odbcheck.exe  
 Usage :: odbcheck.exe sqlfile [prefs]  
 sqlfile :: sql select statement or file or keyword :TABLES  
 prefs :: configuration file (default) preferences.prefs**

Die in der Präferenzdatei spezifizierte ODBC Datenquelle wird ausgelesen. Ein Aufruf mit der Option **:TABLES** zeigt die Tabellen und Feldstruktur der Datenquelle.

**bin\odbcheck.exe :TABLES odbcdata\odbcdata.prefs**

**ODBC Text Driver: Driver={Microsoft Text Driver (\*.txt; \*.csv)};Dbq=.\odbcdata**

**Array**

**(**

**[gv\_anlage.csv] => Array**

**(**

**[ID] => CHAR (255)**

**[GEBAEUDE\_ID] => CHAR (255)**

**[TYP\_CODE] => CHAR (255)**

**[TYP\_TEXT] => CHAR (255)**

**)**

**[gv\_gebaeude.csv] => Array**

**(**

**[ID] => CHAR (255)**

**[GRUNDSTUECK\_ID] => INTEGER (10)**

**[POLICE\_NR] => CHAR (255)**

**[SUCHBEGRIFF] => CHAR (255)**

**[STATUS\_CODE] => CHAR (255)**

**[STATUS\_TEXT] => CHAR (255)**

**[ZWECK\_CODE] => INTEGER (10)**

**[ZWECK\_TEXT] => CHAR (255)**

**[BAUJAHR] => INTEGER (10)**

**[BAUART\_CODE] => CHAR (255)**

**[BAUART\_TEXT] => CHAR (255)**

**[LAGE\_CODE] => CHAR (255)**

**[LAGE\_TEXT] => CHAR (255)**

**[KANTON] => CHAR (255)**

**[BEZIRK] => INTEGER (10)**

**[GEMEINDE\_BFS] => INTEGER (10)**

**[GEMEINDE] => CHAR (255)**

**[STRASSE] => CHAR (255)**

**[STRASSE\_NR] => INTEGER (10)**

**[WOHNUNG\_NR] => CHAR (255)**

**[PLZ] => CHAR (255)**

**[PLZ\_ZUSATZ] => CHAR (255)**

**[ORT] => CHAR (255)**

**[ORT\_ZUSATZ] => CHAR (255)**

**)**

Durch ein entsprechendes SQL Select Statement kann die Datenquelle auch mit der SQL Syntax befragt werden:

**bin\odbcheck.exe "SELECT \* FROM gv\_anlage.csv" odbcdata\odbcdata.prefs**

**ODBC Text Driver: Driver={Microsoft Text Driver (\*.txt; \*.csv)};Dbq=.\odbcdata**

**Array**

**(**

**[ID] => CHAR (255)**

**[GEBAEUDE\_ID] => CHAR (255)**

**[TYP\_CODE] => CHAR (255)**

**[TYP\_TEXT] => CHAR (255)**

**)**

**ID;GEBAEUDE\_ID;TYP\_CODE;TYP\_TEXT**

**01 10005;01 10005;BS;Blitzschutz**

**01 10009;01 10009;BS;Blitzschutz**

**01 10009;01 10009;BL;Brandmelder**

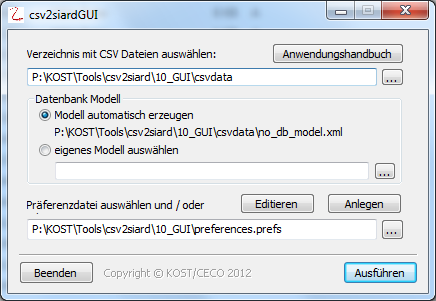
**Result row count: 3**

# Einfaches GUI zu csv2siard

Ein einfaches Windows Programm **c2sGUI.exe** erlaubt es csv2siard nicht nur über die Kommandozeile, sondern auch über eine grafisches Benutzerschnittstelle (GUI) zu bedienen. Das Programm ist weitgehend selbsterklärend, bzw. übernimmt Syntax und Logik vollumfänglich von der Kommandozeilen Version **csv2siard.exe**.







Das Einstiegsformular dient der unterstützen Erfassung aller Programmparameter. Eine Präferenzdatei kann ausgewählt, neu angelegt oder editiert werden. Die SIARD Datei, wie auch ein automatisch erzeugtes Datenmodell **(:NO\_DB\_MODELL)**, wird der Einfachheit halber direkt in das CSV Verzeichnis geschrieben. Bestehen kein Schreibrechte in diesem Verzeichnis, wird stattdessen der Desktop gewählt.

Mit dem „Ausführen“ Knopf wird die Konvertierung gestartet und anschliessend eine LOG Datei mit Notepad angezeigt. Die einmal gewählten Werte werden beim Ausführen für den nächsten Programmlauf gespeichert.

„Beenden“ schliesst das Programm.

# Installierte Dateien

|  |  |
| --- | --- |
| **11** | Folgende Dateistruktur wird beim Installieren von **csv2siard** angelegt:  └───Programme  └───csv2siard  │ Anwendungshandbuch\_v1.7.pdf  │ database-torque-4-0.xsd  │ datatype-model.xml  │ demo.mdb  │ demo.xls  │ gv-model-nf.xml  │ gv-model-v9.xml  │  ├───bin  │ crc32sum.exe  │ csv2siard.exe  │ csvschema.exe  │ expat.dll  │ file.exe  │ GPL-2.0\_COPYING.txt  │ iconv.dll  │ libxml2.dll  │ magic.mgc  │ magic1.dll  │ odbcheck.exe  │ preferences.prefs  │ regex2.dll  │ sablot.dll  │ xmllint.exe  │ zlib1.dll  │  ├───csvdata  │ gv\_anlage.dat  │ gv\_gebaeude.dat  │ gv\_person.dat  │ gv\_position.dat  │ gv\_schaden.dat  │ gv\_schaetzung.dat  │  ├───datatype  │ ascii.csv  │ datatype.prefs  │ datatype\_binary.csv  │ datatype\_date.csv  │ datatype\_int.csv  │ datatype\_numeric.csv  │ datatype\_real.csv  │ datatype\_string.csv  │ datatype\_utf8.csv  │  ├───odbcdata  │ gv\_anlage.csv  │ gv\_gebaeude.csv  │ gv\_person.csv  │ gv\_position.csv  │ gv\_schaden.csv  │ gv\_schaetzung.csv  │ odbcdata.prefs  │ schema.ini  │  ├───odbcsql  │ anl.sql  │ geb.sql  │ odbcsql.prefs  │ per.sql  │ pos.sql  │ rol.sql  │ shd.sql  │ shz.sql  │  └───source  c2odbc.php  c2schema.php  c2sconfig.php  c2sconvert.php  c2screate.php  c2sfunction.php  c2snodbmodel.php  c2stimedate.php  c2sxml.php  csv2siard.bcp  csv2siard.php  zip.php |

1. SIARD ist die Archivierungslösung für relationale Datenbanken des Schweizerischen Bundesarchives: <http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=de>. [↑](#footnote-ref-1)
2. SiardEdit ist Teil der SIARD Suite und wird vom Schweizerischen Bundesarchiv unentgeltlich zu Verfügung gestellt. [↑](#footnote-ref-2)
3. Siehe dazu das Apache DB Project <http://db.apache.org/torque/releases/torque-4.0/index.html>. [↑](#footnote-ref-3)
4. D.h. in einem ersten Durchgang wird mit **csv2siard** ein Datenmodell generiert, das danach manuell ergänzt wird. In einem zweiten Durchgang wird mit den gleichen CSV-Dateien und diesem ergänzten Datenmodell die gewünschte SIARD-Datei erzeugt. [↑](#footnote-ref-4)
5. Transferprojekt Gebäudeversicherung: <http://kost-ceco.ch/cms/index.php?transferprojekt_de>. [↑](#footnote-ref-5)
6. Achtung, die Dateien im Ordner **datatype** haben die Dateiendung **.csv**, die Preference-Datei **preferences.prefs** muss in diesem Fall geändert oder **datatype/datatype.prefs** verwendet werden. [↑](#footnote-ref-6)
7. Bambalam PHP EXE Compiler/Embedder: <http://www.bambalam.se/bamcompile/>. [↑](#footnote-ref-7)
8. Die *default* Präferenzdatei „preferences.prefs“ wird erst im aktuellen Arbeitsverzeichnis und dann im Installationsverzeichnis csv2siard\bin\ gesucht. [↑](#footnote-ref-8)
9. Durch DBMS *(Database Management System)* gegebene Namenseinschränkung für Tabellen und Spalten: Nur Buchstaben aus dem US-ASCII Zeichensatz, Zahlen und der Unterstrich sind erlaubt, das erste Zeichen muss ein Buchstabe sein; keine Unterscheidung zwischen Gross- und Kleinschreibung, maximale Namenslänge ist 30 Zeichen. [↑](#footnote-ref-9)
10. Gewisse Zeichensätze schliessen andere Zeichensätze ein; so ist zum Beispiel US-ASCII in ANSI und ISO-8859-1 enthalten, ASCII aber nicht in ANSI und ISO-8859-1. Dieser Umstand kann zu irreführenden Fehlermeldungen bei der Analyse der CSV-Dateien mit der Option :NO\_DB\_MODEL führen.  
    (Extended ASCII und OEM sind identische Zeichensätze, ISO-8859-1 ist ein *Subset* von ANSI) [↑](#footnote-ref-10)
11. Das Einfassen der Felder in ein Zitatzeichen *(Quotation Mark)* ist in CSV nicht obligatorisch und macht nur in dem Falle Sinn, wo ein Feldtrennzeichen *(Column Separator)* Teil des Feldinhaltes ist. [↑](#footnote-ref-11)
12. MS-Excel CSV-Dateien können unterschiedliche Spaltenzahlen pro Zeile haben. Die Überprüfung der Anzahl Spalten auf Grund der Vorgabe im Datenbankschema oder der Vergleich mit der Spaltenzahl der ersten Spalte (Feldnamen) schlägt hier in der Regel fehl. [↑](#footnote-ref-12)
13. In gewissen Fällen kann es notwendig sein, die Überprüfung der Spaltennamen in der ersten Zeile auszuschalten. Dann nämlich, wenn diese Spaltennamen nicht den SQL-Namensvorgaben entsprechen und im Datenbankschema durch Dummy-Namen ersetzt worden sind. [↑](#footnote-ref-13)
14. Gewisse Steuerzeichen sind nicht Teil des UNICODE-Zeichensatzes und auch als XML-Entities nicht in einer XML Datei erlaubt, siehe <http://www.w3.org/TR/2000/REC-xml-20001006#charsets>. Mit dieser Einstellung wird diese Einschränkung aufgehoben und die Zeichen in \u00xx Notation dargestellt   
    *(escaped Unicode encodings).* [↑](#footnote-ref-14)
15. Werden Sonderzeichen oder Umlaute in den *optional content settings* verwendet, muss die Preference-Datei UTF-8 codiert gespeichert werden. [↑](#footnote-ref-15)
16. Empty String. DESCRIPTION, ARCHIVED\_BY und CONTACT sind nicht datenbankbezogene Informationsfelder, sie können leer gelassen und mit SiardEdit bearbeitet werden. OWNER und TIMESPAN sind ebenfalls archivische Informationsfelder, müssen aber Text enthalten. [↑](#footnote-ref-16)
17. SIARD\_USER und SIARD\_SCHEMA sind datenbankrelevante Felder. Bei einem Export einer SIARD-Datei in eine Datenbank wird ein Schema oder Datenbank mit dem SIARD\_SCHEMA Namen angelegt und ein Datenbankuser mit dem Namen SIARD\_USER erhält die Admin-Rechte in diesem Schema. [↑](#footnote-ref-17)
18. 8-Bit codiert 0x1B bzw. ESC. [↑](#footnote-ref-18)
19. Uncodierte Signatur einer ZIP Datei (vier Byte 0x504B0304). [↑](#footnote-ref-19)
20. Base64 codiert „This is an encoded string“. [↑](#footnote-ref-20)
21. Folgende Formatbezeichner werden unterstützt „%S, %M, %H, %d, %m, %Y“  
     %Y - Jahr als 4-stellige-Zahl inklusive des Jahrhunderts  
     %m - Monat als Zahl (Bereich 01 bis 12)   
     %d - Tag des Monats als Zahl (Bereich 01 bis 31)   
     %H - Stunde als Zahl im 24-Stunden-Format (Bereich 00 bis 23)   
     %M - Minute als Dezimal-Wert   
     %S - Sekunden als Dezimal-Wert   
    Zum Beispiel erkennt DATE\_FORMAT=%Y/%m/%d folgendes Datum: „2008/07/01“ [↑](#footnote-ref-21)
22. Unter ODBC *(open database connectivity)* versteht man eine von der Firma Microsoft 1992 entwickelte Software-Schnittstelle (API), die den Zugriff von Anwendungsprogrammen auf unterschiedliche Datenbanken gewährleisten soll. Der Vorteil besteht in der Unabhängigkeit der Anwendungsprogrammierung von der zugrunde liegenden Datenbankimplementierung. Seit Windows 2000 ist ODBC integraler Bestandteil des Betriebssystems. ODBC ist inzwischen aber auch in der UNIX Welt verfügbar, das Pendant in der JAVA Welt ist JDBC. Auf die verschiedenen Datenbanken wird mit einem jeweils speziellen ODBC-Treiber zugegriffen, solche Treiber existieren für alle gängigen Datenbanken (Oracle, DB2, SQL-Server, Access, Informix, MySQL, um nur einige zu nennen).

    Die ODBC Schnittstelle ist als API in unterschiedlichen Programmiersprachen verfügbar und unterstützt SQL basierte Abfragen. [↑](#footnote-ref-22)
23. Der Dateinamen für **Dbq** unterliegt einigen Einschränkungen, so darf er keine Leerzeichen enthalten und Ordner und Dateinamen dürfen nicht mit Zahlen beginnen. Relative Dateipfade sind aber möglich, z.B **Dbq=.\csvtext\** [↑](#footnote-ref-23)
24. Andere Dateiendungen wie zum Beispiel **.dat** führen zu Problemen. [↑](#footnote-ref-24)
25. Bei einer ODBC-Datenquelle kann der Zeichensatz nicht via Datenverbindung ermittelt werden. Die Zeichensätze ANSI und OEM sind programmtechnisch nicht zu unterscheiden, sodass eine manuelle Prüfung (Stichproben) sinnvoll ist. [↑](#footnote-ref-25)
26. Siehe dazu MSDN, Microsoft Developer Network: „Schema.ini File (Text File Driver)“ <http://msdn.microsoft.com/en-us/library/windows/desktop/ms709353(v=vs.85).aspx> [↑](#footnote-ref-26)
27. Es gibt fünf Text File Data Types: CHAR, DATETIME, FLOAT, INTEGER, LONGCHAR  
    <http://msdn.microsoft.com/en-us/library/windows/desktop/ms715429(v=vs.85).aspx> [↑](#footnote-ref-27)