

csv2siard v.1.6 **Anwendungshandbuch**

Inhalt

1	Programmbeschreibung.....	2
2	csv2siard installieren.....	3
3	csv2siard konfigurieren	3
4	Gebäudeversicherungsdateien in eine SIARD-Datei konvertieren	5
5	Beliebige CSV-Dateien in eine SIARD-Datei konvertieren	6
6	Präferenzen	7
7	Konsoleausgabe	8
8	Installierte Dateien	9

1 Programmbeschreibung

Das Tool **csv2siard** erlaubt die Konvertierung von CSV-Dateien in eine SIARD-Datei¹. 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 SiardEdit² 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 Standard³. 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

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 spezifische relationale Aspekte wie *Unique-Constraints* und *Foreign Key Constraints*.

Zur Veranschaulichung sind aus dem KOST-Projekt "Archivierung von Gebäudeversicherungsdaten"⁵ das Datenmodell *gv-model-v8.xml* und eine kleine anonymisierte Testdatensammlung von CSV-Dateien beigelegt.

Der Vollständigkeit halber ist der Source Code in PHP ebenfalls beigelegt. Das ausführbare Programm ist mit Bamcompile⁶ kompiliert. **csv2siard** benötigt zusätzlich die Programme *7z.exe*, *file.exe* und *xmllint.exe*. Diese Programme sind Freeware, bitte beachten Sie jedoch die jeweiligen Urheberrechtsbestimmungen.

¹ SIARD ist die Archivierungslösung für relationale Datenbanken des Schweizerischen Bundesarchives: <http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=de>.

² SiardEdit ist Teil der SIARD Suite und wird vom Schweizerischen Bundesarchiv unentgeltlich zu Verfügung gestellt.

³ Siehe dazu das Apache DB Project <http://db.apache.org/torque/releases/torque-4.0/index.html>.

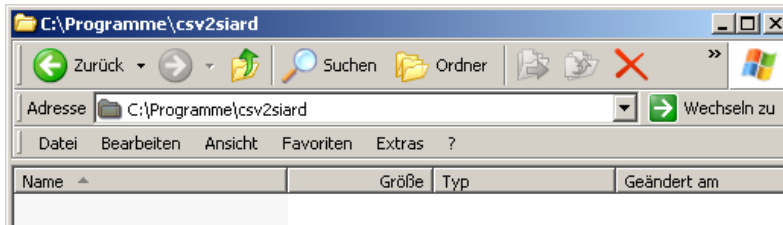
⁴ D.h. in einem ersten Durchgang wird mit **csv2siard** ein Datenmodell generiert, das danach manuell ergänzt wird. In einem zweiten Durchgang wird mit den gleichen CSV-Dateien und diesem ergänzten Datenmodell die gewünschte SIARD Datei erzeugt.

⁵ Transferprojekt Gebäudeversicherung: http://kost-ceco.ch/cms/index.php?transferprojekt_de.

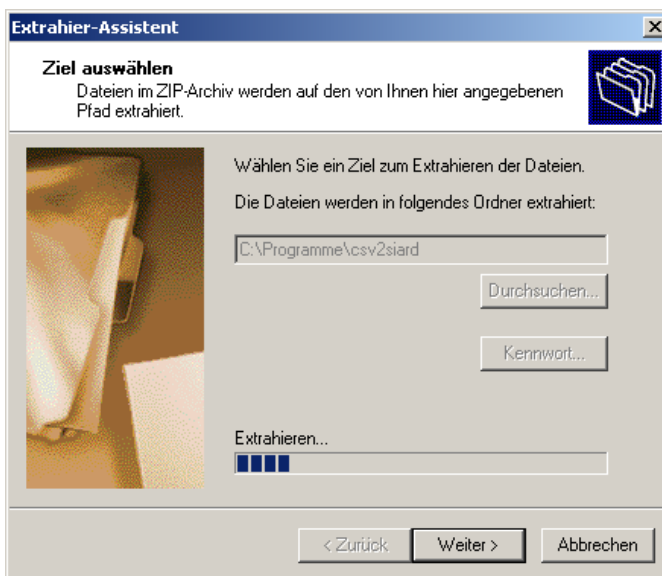
⁶ Bambalam PHP EXE Compiler/Embedder: <http://www.bambalam.se/bamcompile/>.

2 csv2siard installieren

- 1 csv2siard Arbeitsverzeichnis erstellen.
Zum Beispiel Ordner **csv2siard** im Verzeichnis **C:\Programme** anlegen



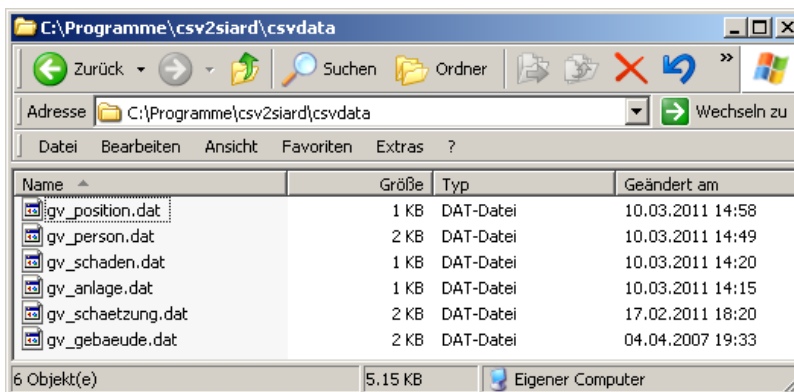
- 2 **csv2siard.zip** herunterladen und in das Verzeichnis **C:\Programme\csv2siard** entpacken.



Der Pfad zum ausführbaren Programm lautet anschließend **C:\Programme\csv2siard\bin\csv2siard.exe**

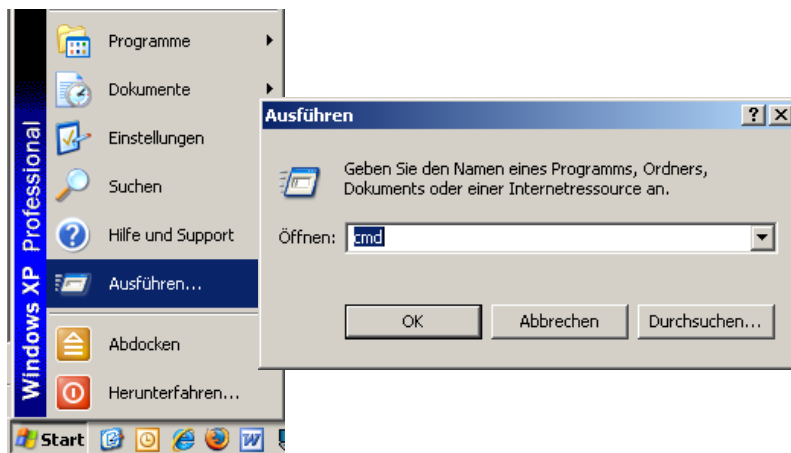
3 csv2siard konfigurieren

- 3 CSV Dateien bereitstellen, z.B. im folgenden Verzeichnis:
C:\Programme\csv2siard\csvdata

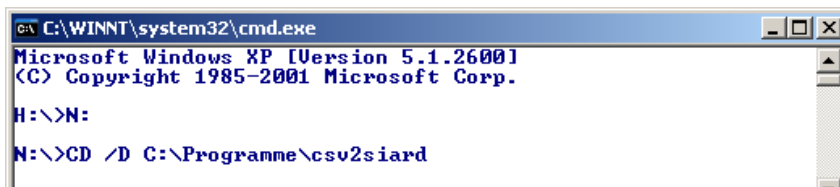


Die Dateinamen müssen den Einschränkungen der gewählten **FILE_MASK** entsprechen, siehe Kapitel 6. Präferenzen.

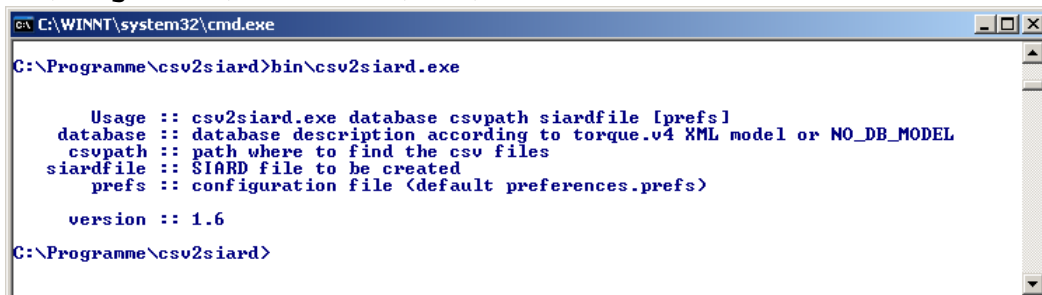
4 Ein Kommandozeilenfenster öffnen:



5 In das gewünschte Arbeitsverzeichnis wechseln, hier z.B. mit
`CD /D C:\Programme\csv2siard`



6 Tool starten und Usage / Help / Version anzeigen lassen.
Der Pfad zum ausführbaren Programm lautet
`C:\Programme\csv2siard\bin\csv2siard.exe`



Besser lesbar:

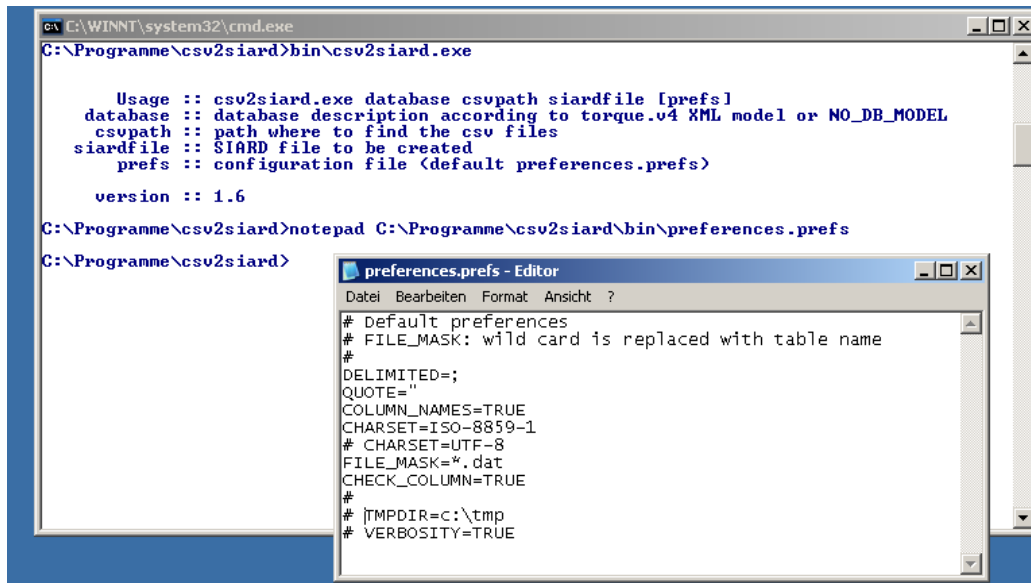
```
C:\Programme\csv2siard> bin\csv2siard.exe

Usage :: csv2siard.exe database csvpath siardfile [prefs]
database :: database description according to torque.v4
           XML model or NO_DB_MODEL
csvpath  :: path where to find the csv files
siardfile :: SIARD file to be created
prefs    :: configuration file (default preferences.prefs)

version :: 1.6
```

- 7 Präferenzen für die Konvertierung CSV -> SIARD festlegen. Die voreingestellten Werte werden in der Regel korrekt sein, siehe dazu Kapitel 6. Präferenzen, weiter unten.

notepad C:\Programme\csv2siard\bin\preferences.prefs



The screenshot shows a Windows command prompt window with the following text:

```
C:\WINNT\system32\cmd.exe
C:\Programme\csv2siard>bin\csv2siard.exe

Usage :: csv2siard.exe database csvpath siardfile [prefs]
database :: database description according to torque.v4 XML model or NO_DB_MODEL
csvpath :: path where to find the csv files
siardfile :: SIARD file to be created
prefs :: configuration file <default preferences.prefs>

version :: 1.6

C:\Programme\csv2siard>notepad C:\Programme\csv2siard\bin\preferences.prefs
C:\Programme\csv2siard>
```

The Notepad window, titled "preferences.prefs - Editor", shows the following content:

```
# Default preferences
# FILE_MASK: wild card is replaced with table name
#
DELIMITED=;
QUOTE="
COLUMN_NAMES=TRUE
CHARSET=ISO-8859-1
# CHARSET=UTF-8
FILE_MASK=*.dat
CHECK_COLUMN=TRUE
#
# TMPDIR=C:\tmp
# VERBOSITY=TRUE
```

Wichtig sind vor allem die korrekten Einstellungen für diese fünf Werte:

```
DELIMITED=;
QUOTE="
COLUMN_NAMES=TRUE
CHARSET=ISO-8859-1
FILE_MASK=*.dat
```

4 Gebäudeversicherungsdateien in eine SIARD-Datei konvertieren

- 8 **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-v8.xml** wird beim Installieren des Tool gleich angelegt:

bin\csv2siard.exe gv-model-v8.xml csvdata new.siard



The screenshot shows a Windows command prompt window with the following text:

```
C:\WINNT\system32\cmd.exe
C:\Programme\csv2siard>bin\csv2siard.exe gv-model-v8.xml csvdata new.siard

csv2siard v 1.6, Copyright (C) 2011 Martin Kaiser (KOST-CECO)
This program comes with ABSOLUTELY NO WARRANTY.
This is free software, and you are welcome to redistribute it under certain conditions;
see GPL-2.0_COPYING.txt for details.

Process table (encoding: iso-8859-1) gv_gebaeude
Process table (encoding: us-ascii) gv_person
Process table (encoding: us-ascii) gv_anlage
Process table (encoding: us-ascii) gv_schaden
Process table (encoding: us-ascii) gv_schaetzung
Process table (encoding: us-ascii) gv_position
ZIP SIARD file .....
SIARD file created: C:/Programme/csv2siard/new.siard
Conversion completed

C:\Programme\csv2siard>
```

5 Beliebige CSV-Dateien in eine SIARD-Datei konvertieren

- 9 **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.⁷ 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



```
C:\WINNT\system32\cmd.exe
C:\Programme\csv2siard>bin\csv2siard.exe NO_DB_MODEL csvdata new.siard

csv2siard v 1.6, Copyright (C) 2011 Martin Kaiser (KOST-CECO)
This program comes with ABSOLUTELY NO WARRANTY.
This is free software, and you are welcome to redistribute it under certain conditions;
see GPL-2.0_COPYING.txt for details.

CSV file C:/Programme/csv2siard/csvdata/gv_anlage.dat does not conform to ISO-8859-1 encoding
CSV file C:/Programme/csv2siard/csvdata/gv_person.dat does not conform to ISO-8859-1 encoding
CSV file C:/Programme/csv2siard/csvdata/gv_position.dat does not conform to ISO-8859-1 encoding
CSV file C:/Programme/csv2siard/csvdata/gv_schaden.dat does not conform to ISO-8859-1 encoding
CSV file C:/Programme/csv2siard/csvdata/gv_schaetzung.dat does not conform to ISO-8859-1 encoding
.....
New XML database model written: C:\Programme\csv2siard\no_db_model.xml
[gv_anlage] => C:/Programme/csv2siard/csvdata/gv_anlage.dat
[gv_gebaeude] => C:/Programme/csv2siard/csvdata/gv_gebaeude.dat
[gv_person] => C:/Programme/csv2siard/csvdata/gv_person.dat
[gv_position] => C:/Programme/csv2siard/csvdata/gv_position.dat
[gv_schaden] => C:/Programme/csv2siard/csvdata/gv_schaden.dat
[gv_schaetzung] => C:/Programme/csv2siard/csvdata/gv_schaetzung.dat

Process table (encoding: us-ascii) gv_anlage
Process table (encoding: iso-8859-1) gv_gebaeude
Process table (encoding: us-ascii) gv_person
Process table (encoding: us-ascii) gv_position
Process table (encoding: us-ascii) gv_schaden
Process table (encoding: us-ascii) gv_schaetzung
ZIP SIARD file .....
SIARD file created: C:/Programme/csv2siard/new.siard
Conversion completed
```

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.

⁷ 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.

6 Präferenzen

```
10  # Default preferences
    DELIMITED (default ';')          # CSV column separator
    QUOTE (default '"')8           # Optional field quotation
    COLUMN_NAMES (default true)      # First row contains column names
    CHARSET (default 'ISO-8859-1')9 # character-set (US-ASCII, ASCII,
                                     # OEM, ANSI, ISO-8859-1 and UTF-8)
    FILE_MASK (default '*.dat')      # Wild card is replaced with table name
                                     # or is converted to tablename
    CHECK_COLUMN (default true)10   # Check column count,
                                     # not applicable with MS-Excel CSV
    CHECK_NAMES (default true)11    # Check column names in first row
    CHECK_DATABASE_INTEGRITY (default false) # Not implemented yet
    TMPDIR (default System tempdir)  # default temp dir
    PI_COUNT (default '100')         # Progress indicator per line processed
    VERBOSITY (default false)        # Display additional messages

    # Optional content settings
    DESCRIPTION (default empty)12  # Database description
    ARCHIVED_BY (default empty)      # Database archived by
    CONTACT (default empty)          # Archivist's contact details
    OWNER (default '(...)')          # Data owner prior to archiving
    TIMESPAN (default '(...)')       # Data creation time span
    DB_TYPE (default 'CSV')           # Type of Database or database product
    SIARD_USER (default 'admin')13  # default user
    SIARD_SCHEMA (default 'schema0') # default schema
```

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

⁸ 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.

⁹ 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.

¹⁰ MS-Excel CSV Dateien können unterschiedliche Spaltenzahl pro Zeile haben. Die Überprüfung der Anzahl Spalten auf Grund der Vorgabe im Datenbank Schema oder der Vergleich mit der Spaltenzahl der ersten Spalte (Feldnamen) schlägt hier in der Regel fehl.

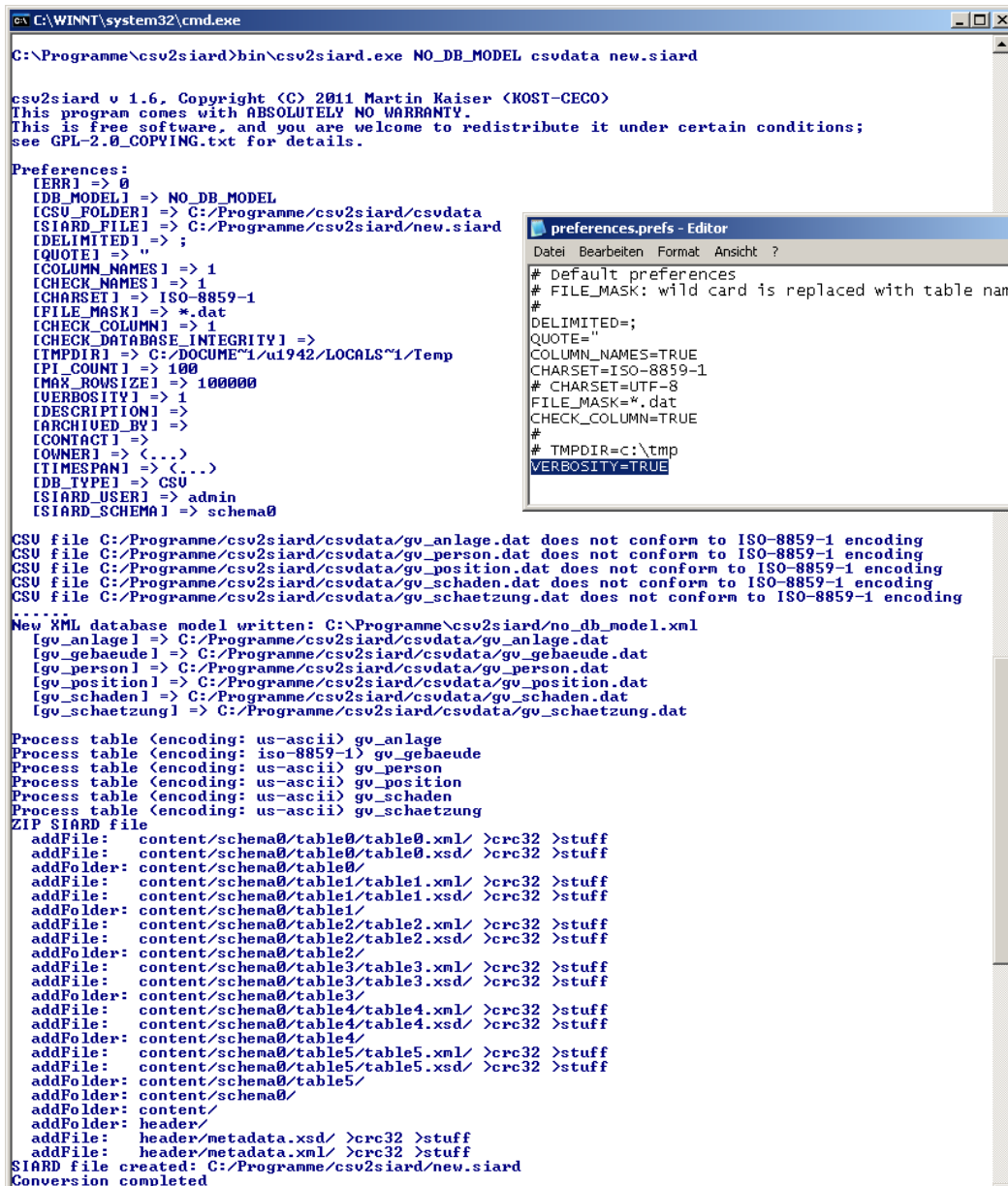
¹¹ In gewissen Fällen kann es notwendig sein die Überprüfung der Spaltennamen in der ersten Zeile auszuschalten. Dann nämlich, wenn diese Spaltennamen nicht den SQL Namensvorgaben entsprechen und im Datenbank Schema durch Dummy-Namen ersetzt worden sind.

¹² Empty String. DESCRIPTION, ARCHIVED_BY und CONTACT sind nicht Datenbank bezogene Informationsfelder, sie können leer gelassen werden und mit SiardEdit bearbeitet werden. OWNER und TIMESPAN sind ebenfalls archivische Informationsfelder, müssen aber Text enthalten.

¹³ SIARD_USER und SIARD_SCHEMA sind Datenbank relevante Felder. Bei einem Export einer SIARD Datei in eine Datenbank wird ein Schema oder Datenbank mit dem SIARD_SCHEMA Namen angelegt und ein Datenbank User mit dem Namen SIARD_USER erhält die Admin Rechte in diesem Schema.

7 Konsoleausgabe

- 11 Die Konsoleausgabe zeigt zuerst den Copyright-Hinweis, und mit der Option **VERBOSEITY** 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 **VERBOSEITY** wird am Schluss der eigentliche Aufbau der SIARD Datei als ZIP Datei angezeigt.



```
C:\Programme\csv2siard>bin\csv2siard.exe NO_DB_MODEL csvdata new.siard

csv2siard v 1.6, Copyright (C) 2011 Martin Kaiser (KOST-CECO)
This program comes with ABSOLUTELY NO WARRANTY.
This is free software, and you are welcome to redistribute it under certain conditions;
see GPL-2.0_COPYING.txt for details.

Preferences:
[ERR] => 0
[DB_MODEL] => NO_DB_MODEL
[CSV_FOLDER] => C:/Programme/csv2siard/csvdata
[SIARD_FILE] => C:/Programme/csv2siard/new.siard
[DELIMITED] => ;
[QUOTE] => "
[COLUMN_NAMES] => 1
[CHECK_NAMES] => 1
[CHARSET] => ISO-8859-1
[FILE_MASK] => *.dat
[CHECK_COLUMN] => 1
[CHECK_DATABASE_INTEGRITY] =>
[TMPDIR] => C:/DOCUMENT1/u1942/LOCALS~1/Temp
[PI_COUNT] => 100
[MAX_ROW_SIZE] => 100000
[VERBOSEITY] => 1
[DESCRIPTION] =>
[ARCHIVED_BY] =>
[CONTACT] =>
[OWNER] => <...>
[TIMESPAN] => <...>
[DB_TYPE] => CSV
[SIARD_USER] => admin
[SIARD_SCHEMA] => schema0

CSV file C:/Programme/csv2siard/csvdata/gv_anlage.dat does not conform to ISO-8859-1 encoding
CSV file C:/Programme/csv2siard/csvdata/gv_person.dat does not conform to ISO-8859-1 encoding
CSV file C:/Programme/csv2siard/csvdata/gv_position.dat does not conform to ISO-8859-1 encoding
CSV file C:/Programme/csv2siard/csvdata/gv_schaden.dat does not conform to ISO-8859-1 encoding
CSV file C:/Programme/csv2siard/csvdata/gv_schaetzung.dat does not conform to ISO-8859-1 encoding
.....
New XML database model written: C:\Programme\csv2siard\no_db_model.xml
[gv_anlage] => C:/Programme/csv2siard/csvdata/gv_anlage.dat
[gv_gebaeude] => C:/Programme/csv2siard/csvdata/gv_gebaeude.dat
[gv_person] => C:/Programme/csv2siard/csvdata/gv_person.dat
[gv_position] => C:/Programme/csv2siard/csvdata/gv_position.dat
[gv_schaden] => C:/Programme/csv2siard/csvdata/gv_schaden.dat
[gv_schaetzung] => C:/Programme/csv2siard/csvdata/gv_schaetzung.dat

Process table (encoding: us-ascii) gv_anlage
Process table (encoding: iso-8859-1) gv_gebaeude
Process table (encoding: us-ascii) gv_person
Process table (encoding: us-ascii) gv_position
Process table (encoding: us-ascii) gv_schaden
Process table (encoding: us-ascii) gv_schaetzung
ZIP SIARD file
  addFile: content/schema0/table0/table0.xml/ >crc32 >stuff
  addFile: content/schema0/table0/table0.xsd/ >crc32 >stuff
  addFolder: content/schema0/table0/
  addFile: content/schema0/table1/table1.xml/ >crc32 >stuff
  addFile: content/schema0/table1/table1.xsd/ >crc32 >stuff
  addFolder: content/schema0/table1/
  addFile: content/schema0/table2/table2.xml/ >crc32 >stuff
  addFile: content/schema0/table2/table2.xsd/ >crc32 >stuff
  addFolder: content/schema0/table2/
  addFile: content/schema0/table3/table3.xml/ >crc32 >stuff
  addFile: content/schema0/table3/table3.xsd/ >crc32 >stuff
  addFolder: content/schema0/table3/
  addFile: content/schema0/table4/table4.xml/ >crc32 >stuff
  addFile: content/schema0/table4/table4.xsd/ >crc32 >stuff
  addFolder: content/schema0/table4/
  addFile: content/schema0/table5/table5.xml/ >crc32 >stuff
  addFile: content/schema0/table5/table5.xsd/ >crc32 >stuff
  addFolder: content/schema0/table5/
  addFolder: content/schema0/
  addFolder: content/
  addFolder: header/
  addFile: header/metadata.xsd/ >crc32 >stuff
  addFile: header/metadata.xml/ >crc32 >stuff
SIARD file created: C:/Programme/csv2siard/new.siard
Conversion completed
```

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 Zeichensatzkonvertierung sind implizit, z.BI us-ascii zu ISO-8859-1, siehe die Fussnote zu **CHARSET** weiter oben.

8 Installierte Dateien

12 Folgende Dateistruktur wird beim Installieren von **csv2siard** angelegt:

