csv2siard v.1.7   
*Anwendungshandbuch*

**Inhalt**

[1 Programmbeschreibung 2](#_Toc317668428)

[2 cvs2siard installieren 3](#_Toc317668429)

[3 csv2siard konfigurieren 3](#_Toc317668430)

[4 Beispiel GV-Daten in eine SIARD konvertieren 5](#_Toc317668431)

[5 Beliebige CSV-Dateien in eine SIARD-Datei konvertieren 6](#_Toc317668432)

[6 Präferenzen 7](#_Toc317668433)

[7 Konsoleausgabe 8](#_Toc317668434)

[8 Konvertierung von CSV zu Datenbankfeldern 9](#_Toc317668435)

[9 Unterstützte Datumformate 10](#_Toc317668436)

[10 Installierte Dateien 11](#_Toc317668437)

# 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 spezifische 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*** und als Zweites eine Testsammlung zur Veranschaulichung von unterschiedlichen Datenfeldern mit dem Datenmodell ***datatype-model.xml*** und den Dateien in ***csvtest[[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** anlegen |

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

# csv2siard konfigurieren

|  |  |
| --- | --- |
| **3a** | 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. |
| **3b** | Ein Kommandozeilenfenster öffen: |

|  |  |
| --- | --- |
| **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 NO\_DB\_MODEL**  **csvpath :: path where to find the csv files**  **siardfile :: SIARD file to be created**  **prefs :: configuration file (default preferences.prefs)**  **version :: 1.7** |
| **3e** | 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**    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.[[8]](#footnote-8) 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') [[9]](#footnote-9) # 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 '"') [[10]](#footnote-10) # Optional field quotation  FILE\_MASK (default '\*.dat') # Wild card is replaced with table name  # or is converted to tablename  CHECK\_COLUMN (default true) [[11]](#footnote-11) # Check column count,   # not applicable with MS-Excel CSV  CHECK\_NAMES (default true) [[12]](#footnote-12) # 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  INVALID\_ENTITIES (default false)[[13]](#footnote-13) # Ignore invalid XML character  VERBOSITY (default false) # Display additional messages  ***# Optional content settings***) [[14]](#footnote-14)  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) [[15]](#footnote-15) # Database description  OWNER (default '(...)') # Data owner prior to archiving  SIARD\_SCHEMA (default 'schema0') # default schema  SIARD\_USER (default 'admin') [[16]](#footnote-16) # default user  TIMESPAN (default '(...)') # Data creation time span |

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

# Konsoleausgabe

|  |  |
| --- | --- |
| **7** | Die Konsoleausgabe 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 Aufbaue 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 Zeichensatzkonvertierung 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 | integer | INTEGER |
| -232767 | ctype\_digit | SMALLINT | integer | INTEGER |
| -2147483647 | ctype\_digit | INTEGER | integer | INTEGER |
| 2147483647 | ctype\_digit | BIGINT | integer | INTEGER |
| 345.6789 | is\_numeric | FLOAT | double | FLOAT |
| 1.23457E+15 | is\_numeric | REAL | double | FLOAT |
| 1.23457E+22 | is\_numeric | DOUBLE | double | FLOAT |
| 1234567891 | is\_numeric | NUMERIC | decimal | NUMERIC |
| 12345678.25 | is\_numeric | DECIMAL | decimal | NUMERIC |
| A | xml\_encode | CHAR | string | CHARACTER VARYING |
| ABV | xml\_encode | VARCHAR | string | CHARACTER VARYING |
| Victor jagt zwölf Boxkämpfer quer über den Sylter Deich | xml\_encode | LONGVARCHAR | string | CHARACTER VARYING |
| 2003-12-31 | convert2XMLdate | DATE | date | DATE |
| 01:02:03 | convert2XMLdate | TIME | time | TIME |
| 2003-12-31T01:02:03 | convert2XMLdate | TIMESTAMP | dateTime | TIMESTAMP |
| 00011011[[17]](#footnote-17) | bit->hex | BIT | hexBinary | BIT |
| PK□□[[18]](#footnote-18) | bin->hex | BINARY | hexBinary | BIT VARYING |
| VGhpcyBpcyBhbiBlbmNvZGVkIHN0cmluZw==[[19]](#footnote-19) | base64->hex | VARBINARY | hexBinary | BIT VARYING |
| R0lGODlhDAAKAJEAAP///3N1B1FRUQAAACwAAAAADAAKAAACGpSPB8ttDcELNE5Ac5ACVww+ESOOnLkkqlEAADs= | base64->hex | LONGVARBINARY | hexBinary | BIT VARYING |
| R0lGODlhDAAKAJEAAP///3N1B1FRUQAAACwAAAAADAAKAAACGpSPB8ttDcELNE5Ac5ACVww+ESOOnLkkqlEAADs= | base64->hex | BLOB | hexBinary | BLOB |
| The quick brown fox jumps over the lazy dog | xml\_encode | CLOB | string | CHARACTER VARYING |
| http://ch.php.net/manual/en/function.base64-decode.php |  | REF | string | CHARACTER VARYING(255) |
| TRUE | to\_bool | BOOLEANINT | boolean | BOOLEAN |
| FALSE | to\_bool | BOOLEANCHAR | 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() [[20]](#footnote-20) |  |
| 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" |

# Installierte Dateien

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

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** wird 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 **csvtest** haben die Dateiendung **.csv**, die Preference Datei **preferences.prefs** muss in diesem Fall geändert oder **csvtest/csvtest.prefs** verwendet werden. [↑](#footnote-ref-6)
7. Bambalam PHP EXE Compiler/Embedder: <http://www.bambalam.se/bamcompile/>. [↑](#footnote-ref-7)
8. 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-8)
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. [↑](#footnote-ref-9)
10. 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-10)
11. 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. [↑](#footnote-ref-11)
12. 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. [↑](#footnote-ref-12)
13. Gewisse Zeichen sind auch als XML-Entities codiert nicht in einer XML Datei erlaubt, siehe <http://www.w3.org/TR/2000/REC-xml-20001006#charsets>. Mit dieser Preference Einstellung wird diese Einschränkung aufgehoben. Es wird aber so möglicherweise eine fehlerhafte SIARD Datei erstellt. Diese Option dient im wesentlichen der Fehlersuche. [↑](#footnote-ref-13)
14. Werden Sonderzeichen oder Umlaute in den *optional content settings* verwendet, muss die Preference Datei UTF-8 codiert gespeichert werden. [↑](#footnote-ref-14)
15. 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. [↑](#footnote-ref-15)
16. 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. [↑](#footnote-ref-16)
17. 8-Bit codiert 0x1B bzw. ESC [↑](#footnote-ref-17)
18. Uncodierte Signatur einer ZIP Datei (vier Byte 0x504B0304) [↑](#footnote-ref-18)
19. Base64 codiert „This is an encoded string“ [↑](#footnote-ref-19)
20. 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-20)