KOST Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen

Ein Gemeinschaftsunternehmen von Schweizer Archiven

Das SIARD Format und die zugehörige Tool-Landschaft

Arbeitskreis Archivierung von Unterlagen aus digitalen Systemen

17. Tagung im Staatsarchiv Dresden
13./14. März 2013

Inhaltsübersicht

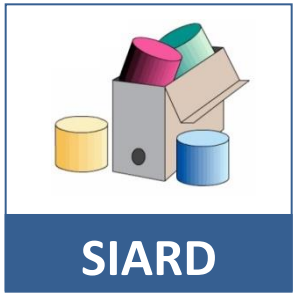
- Langzeitarchivierung von Datenbanken
  Eine Lösung: SIARD Format
- Die Tool-Landschaft um SIARD Format
- CSV <=> SIARD
- Das Tool *csv2siard*
- Kommandozeile / GUI *csv2siard*
- CSV Dateien konvertieren
- ODBC Quellen konvertieren

Langzeitarchivierung von Datenbanken

- Aus Registern werden bereits in den 70er Jahren Datenbank gestützte Fachanwendungen
- Proprietäre Lösungen und Export in CSV Dateien sollen den Datenaustausch zwischen Datenbanken ermöglichen, Archivierung steht lange Zeit nicht im Vordergrund
- In diesem Jahrhundert tritt neu der Datenexport in XML Dateien als Alternative hinzu, ohne dass sich ein Standard etablieren könnte
- SIARD Format verbindet XML basierte Datenspeicherung mit SQL:99 orientierter Metadatenbeschreibung, das Format wird in der aktuellen Form vom Schweizerischen Bundesarchiv erstmals 2008 in der Version 1.0 veröffentlicht
- Dieses Jahr soll SIARD-Format als eCH Standard 0165 freigegeben werden, dazu wurde eine erweiterte Formatbeschreibung erstellt

Die Tool-Landschaft um SIARD Format

- Ein Dateiformat verlangt auch Anwendungen, mit denen das Format erstellt, bearbeitet und genutzt werden kann
- Bei der Datenbankarchivierung wird eine relationale Datenbank serialisiert, das heisst in eine oder mehrere Dateien bestimmter Struktur umgewandelt



- ***SIARD-Suite*** des Schweizerischen Bundesarchives stellt ein Tool zur Umwandlung *Datenbank* ⇔ *SIARD Format* zur Verfügung
- Inbegriffen in SIARD-Suite ist mit ***SiardEdit*** ein Tool zum Betrachten der SIARD Daten und Editieren der SIARD Metadaten
- ***SIARD-Val*** der KOST ist ein eigentlicher Formatvalidator, der die Konformität zur SIARD Spezifikation eCH-0165 überprüft

SIARD-VAL

- ***csv2siard*** erlaubt, CSV Dateisammlungen direkt in eine SIARD Datei zu konvertieren

csv 2 siard

CSV <=> SIARD

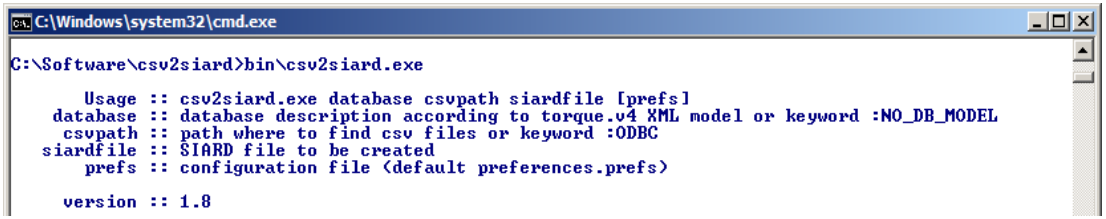
- Das CSV Format gilt gemeinhin als archivtauglich und lange erprobt, warum sollen CSV Dateien in SIARD konvertiert werden?
- CSV hat gewisse Schwächen, was die Formatspezifikation betrifft, jede CSV Datei ist in der Regel eine Individuallösung, was Trennzeichen, Zeichensatz, *Quotation* etc. betrifft
- Die Möglichkeiten zur Dokumentation auf Tabellen / Feldebene sind in SIARD viel umfänglicher und CSV Dateisammlungen können in einer SIARD Datei zusammengefasst werden
- Bei einer zeitnahen Konvertierung können Dateninkonsistenzen in den CSV Dateien erkannt und korrigiert werden
- Datenbanken und datenbankgestützte Fachanwendungen können in SIARD archiviert werden, auch wenn keine Möglichkeit für den direkten Zugriff auf die Datenbank mit *SIARD-Suite* besteht

Das Tool *csv2siard*

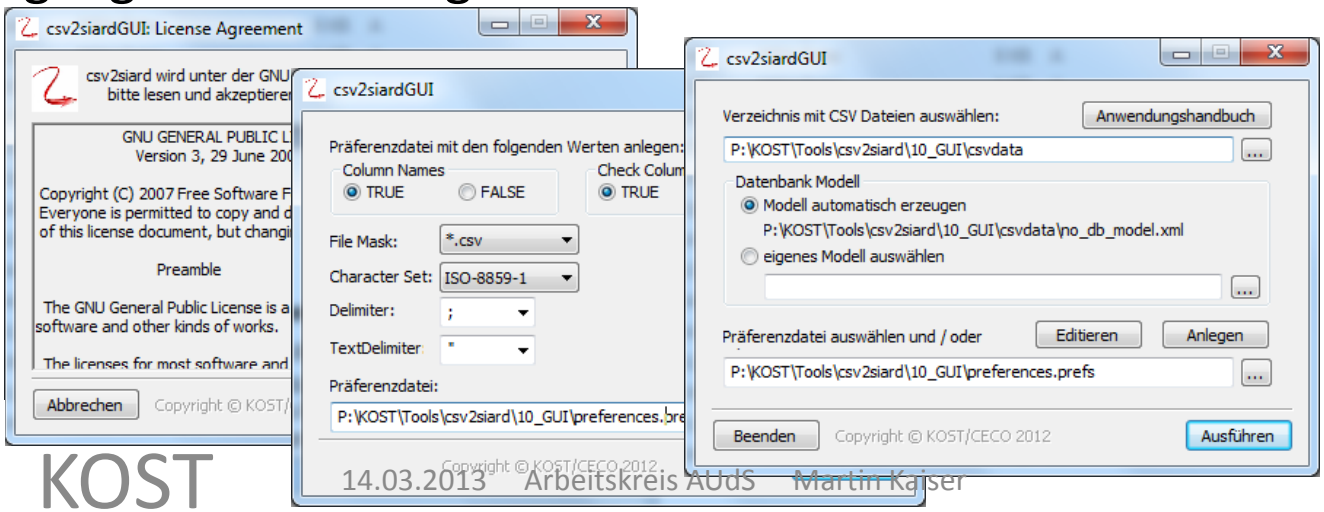
- CSV Dateien/Dateisammlungen können auch via Datenbank in eine SIARD Datei umgewandelt werden, mehrere Gründe haben zur Entwicklung von *csv2siard* geführt:
- Der archivseitige Wunsch, ohne Datenbank und Datenbankkenntnisse CSV in SIARD konvertieren zu können
- Den Konvertierungsvorgang bei Serien von CSV Dateisammlungen automatisieren zu können
- Der archivtheoretische Anspruch, dass ein archivtaugliches Format mit mehreren voneinander unabhängigen Tools erstellt und betrachtet werden kann
- Durch *csv2siard* ist der Nachweis erbracht, dass die SIARD Formatbeschreibung vollständig und inhärent ist

Kommandozeile / GUI csv2siard

- *csv2siard* steht einerseits als Kommandozeilentool zur Verfügung; damit können umfangreiche Konvertierungsprozesse in *Batch* Dateien automatisiert werden



- Andererseits gibt es auch eine GUI Version, die dem Nichtinformatiker den Umgang mit dem Programm erleichtert

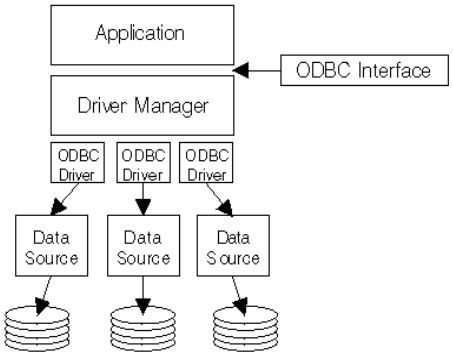


CSV Dateien konvertieren

- Die CSV <=> SIARD Konvertierung verläuft in zwei Schritten:
  - Im ersten Durchgang wird die Datenstruktur (Feldtyp, Feldlänge, etc.) der betrachteten CSV Dateien ***analysiert*** und in ein Datenschema geschrieben
  - Im zweiten Durchgang werden die CSV Dateien anhand dieses Schemas kontrolliert und in eine SIARD Datei ***konvertiert***
- Es ist auch möglich, das Datenschema (nach *Apache Torque 4.0*) manuell zu bearbeiten, zu spezifizieren und mit Kommentaren zu versehen und anschliessen die Konvertierung direkt mit diesem Schema vorzunehmen. Damit können Serienkonvertierungen systematisch durchgeführt werden

```
<?xml version="1.0" encoding="UTF-8"?>
<database name="gemdat5" xmlns="http://db.apache.org/torque/4.0/templates/database" xmlns:xsi="
http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="
http://db.apache.org/torque/4.0/templates/database database-torque-4-0.xsd">
  <table name="gv_minimal2">
    <column name="id" type="DECIMAL" primaryKey="true" description="Personen-ID"/>
    <column name="name" type="VARCHAR" size="1000" description="Personen-Suchbegriff"/>
    <column name="date" type="DATE" description="Geburtsdatum"/>
  </table>
</database>
```

ODBC Quellen konvertieren



- Die Umwandlung von ODBC Quellen verläuft analog zur Umwandlung von CSV Dateien
- Zusätzlich können auf ODBC Datenquellen auch mit SELECT über den Ursprungstabellen erstellte Abfragen in einer SIARD Datei gespeichert werden
- ODBC Quellen sind äusserst flexibel, sie können Datenbanken, aber auch Excel Tabellen oder CSV/Text Dateien sein
- Es bietet sich hiermit in *csv2siard* beinahe die gleiche Flexibilität bei der Datenaufbereitung wie bei der Gestaltung von Views in einer Datenbank
- Das SELECT *Statement* wird selbstredend auch in der SIARD Datei festgehalten

KOST Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen

Ein Gemeinschaftsunternehmen von Schweizer Archiven

Martin Kaiser

Koordinationsstelle für die Archivierung elektronischer Unterlagen
c/o Schweizerisches Bundesarchiv
Archivstrasse 24
CH-3003 Bern

T +41 79 464 08 60
E martin.kaiser@kost.admin.ch
W www.kost-ceco.ch