Hinweise zur Verwendung von CSV-Dateien

Hinweise zum Arbeiten mit dem CSV-Validator

Das CSV-Format

Das Akronym CSV steht für »**C**omma-**s**eparated **v**alues« (durch Komma separierte Werte). Eine CSV-Datei ist eine Textdatei, die mit jedem Texteditor erstellt und geöffnet werden kann und zur Übermittlung von Daten zwischen verschiedenen Programmen eignet. Für den *Aufbau* (1.), die *Formatierung von Datenfeldern* (2.) und die *Zeichencodierung* (3.) gelten folgende Hinweise:

1. Aufbau einer CSV-Datei

- Mit dem CSV-Datenformat lassen sich in erster Linie einfach strukturierte Daten speichern, weiterverarbeiten und austauschen. Es können aber auch kompliziertere und geschachtelte Datenstrukturen – diese mit zusätzlichen Regeln versehen oder in verketteten CSV-Dateien – gespeichert werden.
- Berechnungen sind nicht vorgesehen, aber programmabhängig möglich.
- Die Dateinamenserweiterung (Dateisuffix) lautet »*.csv«.
- Entgegen landläufiger Meinung ist das CSV-Format **kein** allgemeiner Standard, jedoch wird es im RFC 4180 (Requests For Comments, Sammlung von Dokumenten der Internet Engineering Task Force) grundlegend beschrieben.
- Die zu verwendende Zeichenkodierung ist **nicht** festgelegt; der 7-Bit-ASCII-Code mit max. 128 möglichen Zeichen gilt weithin als der kleinste gemeinsame Nenner (üblicherweise wird mit UTF-8, s. u., gearbeitet und in 7-Bit-ASCII geschrieben).

Innerhalb der CSV-Datei haben einige Zeichen eine Sonderfunktion. Diese Zeichen, verwendet für die *Trennung von Datenfeldern* und *Datensätzen* oder zur *Verwendung von Sonderzeichen*, dienen zur Strukturierung der Daten:

- *Trennung von Datenfeldern*: Zur Trennung von Datenfeldern (Spalten) innerhalb eines Datensatzes (Zeile) wird das Komma benutzt. Abhängig von Software und Benutzereinstellungen sind auch Semikolon, Doppelpunkt, Tabulatorzeichen, Leerzeichen oder andere Zeichen üblich.
  Zu beachten: Der Tabulator ist immer dann ein Trennzeichen, wenn die CSV-Datei im Unicode-Format gespeichert wurde. Das Trennzeichen kann auch mit 'sep=<Trennzeichen>' in der ersten Zeile der CSV-Datei explizit angegeben werden, z. B. 'sep=,' für Komma oder 'sep=;' für das Semikolon.
- *Trennung von Datensätzen*: In der Regel dient dazu der Zeilenumbruch des dateierzeugenden Betriebssystems. Unter Windows sind es zwei Zeichen: Carriage Return (CR) und Linefeed (LF).
- *Verwendung von Sonderzeichen*: Um Sonderzeichen innerhalb der Daten nutzen zu können (z. B. Komma in Dezimalzahlwerten oder Freitext), wird ein Feld- oder Textbegrenzungszeichen (in MS Excel als »Textqualifizierer« bezeichnet) benutzt. Dies ist normalerweise das hochgestellte doppelte Anführungszeichen (**"**)*. Ist der Feldbegrenzer selbst in den Daten enthalten (z. B. um Textbestandteile in Anführungszeichen zu setzen), wird dieser im Datenfeld verdoppelt (z. B. **""**eine Hervorhebung**""**).

  * Bei Eingabe des Anführungszeichens zu beachten: **"** (Unicode 0022) ist nicht gleich **“** (Unicode 201C).

2. Formatierung der Datenfelder

Die Formatierung der Daten in CSV-Dateien ist **nicht** festgelegt (s. o.). Das bedeutet, dass die verwendeten Formate zwischen den beteiligten Benutzern abgesprochen werden müssen. Besonders davon betroffen sind:

- *Der Kopfdatensatz*: Der erste Datensatz ist in der Regel ein »Kopfdatensatz«, der die Spaltennamen definiert. Jeder Datensatz (Zeile) muss die gleiche Anzahl Spalten enthalten.
- *Datums- und Zeitangaben*: Die Datums- und Zeitangaben und die Reihenfolge der Einzelangaben (d. h. Jahr, Monat, Tag, Stunde, Minute, Sekunde) kann nicht immer eindeutig erkannt werden.

- *Zahlenwerte mit und ohne führende Null(en)*: Zahlenfelder können mit fester Mindestbreite verwendet werden, dann werden Zahlenwerte mit führenden (vorlaufenden) Nullen ergänzt, bis die Mindestbreite erreicht ist. Beispiele:
  123 = ohne Mindestbreite, ohne führende Nullen
  00123 = Mindestbreite beträgt 5 Zeichen, fehlende Zeichen werden mit Null aufgefüllt
- *Währungsangaben sowie Dezimal- und Tausendertrennzeichen*
- *Negative Zahlenwerte*: Je nach Anwender (und Gewohnheit) werden negative Werte unterschiedlich dargestellt. Beispiele:
  -1 | - 1 | 1- | 1 - | –1 | – 1 | 1– | 1 – | (-)1
- *Leerfelder*: Der Feldinhalt **""** (= Leerfeld) wird manchmal als leerer Inhalt und manchmal als einzelnes Anführungszeichen (das **"** wird »ausmaskiert«) interpretiert. Beispiele:
  ""Hervorhebung**""** „Hervorhebung“
  "" das Feld ist leer, es ist kein Inhalt vorhanden
  " " das Feld ist nicht leer, es enthält einen Leerraum, z. B. einen Wortzwischenraum
- *Texte*: Im Gegensatz zu XML (Extensible Markup Language) sieht CSV **keinen Vermerk des benutzten Zeichensatzes** innerhalb der Datei vor. Die verwendete Zeichencodierung (z. B. UTF-8, s. a. u.) sollte daher zwischen allen Beteiligten im Vorfeld festgelegt werden. Dies ist vor allem wichtig bei der Verwendung von Umlauten (ä, ö, ü/Ä, Ö, Ü – ae, oe, ue/Ae, Oe, Ue).

3. Zeichencodierung nach dem Standard UTF-8

Die Abkürzung UTF-8 bedeutet »8-Bit Universal Character Set Transformation Format« (zu Deutsch: »Universelles 8-Bit-Zeichensatz-Umwandlungs-Format«). Mit dieser Zeichencodierung können sämtliche Sprachzeichen und Textelemente der Sprachen dieser Welt für die IT verwendet werden. Weitere Angaben dazu weiter unten im Abschnitt zum CSV-Validator.

4. Werkzeuge zum Bearbeiten von CSV-Dateien

Jeder Texteditor ist in der Lage, CSV-Dateien zu generieren. Verschiedene Tools und Editoren, kommerzielle und Open Source, sind verfügbar, um die Arbeit mit CSV-Tabellen zu erleichtern und diese übersichtlich darzustellen.

5. Der CSV-Validator des LAD – Hilfsmittel und Kontrollmöglichkeit

Der CSV-Validator ist unter folgender Adresse zu erreichen:

https://denkmalpflege-bw-csv.de

Mit dem CSV-Validator ist es möglich, CSV-Listen von Grabungsfirmen gegen die Vorlagen des LAD zu validieren. Die aktuellen CSV-Listen (**aktuell Listen zur dritten Fassung der Richtlinien**), diese enthalten die jeweils erlaubten Feldinhalte, befinden sich unter:

https://www.denkmalpflege-bw.de/geschichte-auftrag-struktur/archaeologische-denkmalpflege/firmenarchaeologie

Mittels eines einfach zu bedienenden Web-Interfaces, aufgerufen über einen aktuellen Browser sowie eingeschaltetem JavaScript, ist es möglich, CSV-Dateien auf korrekte Datenstrukturen zu prüfen und mögliche Fehler schnell zu erkennen. Sind dabei Fehler enthalten, werden sie rot markiert. Ist die CSV-Datei hingegen korrekt und interpretierbar, wird sie als Tabelle ausgegeben. Dabei werden zu keiner Zeit Daten gespeichert!

Bei der Benützung des CSV-Validators ist zu beachten:

- Über den Button »Hilfe« kann eine ausführliche Anleitung für den CSV-Validator aufgerufen werden.
- Als Feldtrenner ist das Semikolon zu verwenden, Dezimaltrenner ist der Punkt.
- Der Validator versucht das zu verwendende Schema aufgrund des Dateinamens zu setzen. Wird kein Schema erkannt (z. B. das für die Fotoliste), ist das zu verwendende Schema manuell über die Auswahl-Box auszuwählen.

- In der Kopfzeile müssen alle Felder (Spalten) der Vorlagen enthalten sein. In der Vorlage nicht definierte Spalten werden beim Export entfernt oder die Tabelle ggf. nicht erkannt.

6. Anmerkungen zu Zeichencodierungen

Zeichencodierung nach UTF-8
Verbindlich vorgesehen ist die Verwendung der UTF-8 Zeichenkodierung.

- UTF-8 steht für »8-Bit Universal Character Set Transformation Format« (»Universelles 8-Bit-Zeichensatz-Umwandlungs-Format«).
- UTF-8 ist die am weitesten verbreitete internationale Zeichencodierung, mit der sämtliche Sprachzeichen und Textelemente nahezu aller Sprachen der Welt für die EDV-Verarbeitung verwendet werden können.
- Wichtig: Wenn Unicode (s. u.) mit UTF-8-Kodierung verwendet werden soll, muss ein Unicode-fähiger Editor zum Einsatz kommen.

Zu beachten ist, dass lediglich eine konsequente Anwendung von UTF-8 und keiner Verwendung von MS Excel Probleme in der Darstellung von Umlauten, ß und weiteren Sonderzeichen verhindert.

7. Anmerkungen zu Microsoft Excel

Für viele Anwender, die Tabellendaten bearbeiten, gilt MS Excel als das Werkzeug der Wahl, denn Excel hat in mehreren Jahrzehnten Maßstäbe gesetzt und ist praktisch weltweit in allen Unternehmen im Einsatz. **Grundsätzlich** eignet sich Microsoft Excel jedoch **nicht** als »CSV-Editor«.

Zur Arbeit mit Excel und CSV-Listen dennoch im Folgenden einige Anmerkungen. Zunächst muss man unterschieden, wie Daten vor der eigentlichen Bearbeitung in ein Excel-Tabellenblatt übernommen werden (die direkte Eingabe von Hand bleibt hier außen vor). Möglich sind:

- 1. Methode: Die CSV-Datei wird mit einem **Doppelklick** geöffnet oder mit dem Dialog »**Datei öffnen**« geöffnet.
- 2. Methode: Aus anderen Anwendungen werden Daten via »**Kopieren und Einfügen**« eingefügt.
- 3. Methode: Die CSV-Datei wird als Text-Datei mit dem **Textkonvertierungs-Assistent** importiert.

1. Methode: Doppelklick (gilt auch für »Datei öffnen«)

Bei der herkömmlichen Installation von Excel unter Microsoft Windows werden CSV-Dateien mit Excel so verknüpft, dass Sie z. B. beim Öffnen mit Doppelklick standardmäßig mit Excel bearbeitet werden. Das ist einerseits sehr benutzerfreundlich, denn die CSV-Datei wird nicht in ihrer reinen Textform angezeigt, sondern bereits tabellarisch-strukturiert, andererseits aber auch sehr problematisch, denn Folgendes passiert:

- CSV-Dateien werden generell in Tabellenkalkulationsprogrammen unterschiedlich interpretiert. So ist es entscheidend, in welchem »Umfeld« CSV-Daten eingegeben, importiert oder gespeichert werden, da Werteeingaben immer nach internen Programm- oder benutzerspezifischen Vorgaben anpasst und (oft unbemerkt) Änderungen vorgenommen werden.
- Excel prüft den Zeichensatz nicht und geht davon aus, dass der Inhalt einer CSV-Datei einem Windows-spezifischen Zeichensatz entspricht. Dies hat zur Folge, dass beim Öffnen einer CSV-Datei die Dekodierung des Inhalts falsch durchgeführt wird und dadurch Umlaute und Sonderzeichen nicht korrekt dargestellt werden.
- Alle Spalten erhalten die gleiche Breite (unabhängig von der Zeichenmenge je Spalte/Zeile).
- Trennzeichen ist das Semikolon, wenn die CSV-Datei nach ANSI-Norm gespeichert ist, bzw. das Tabulatorzeichen bei Unicode. Das Trennzeichen kann aber auch das Komma sein. Welches Trennzeichen letztendlich verwendet wird, hängt von den Regions- und Spracheneinstellungen des Betriebssystems ab.
- Beim Öffnen analysiert Excel die Spalten und deklariert alle Spalten, in denen nur Ziffern stehen, automatisch als numerische Spalte, obwohl die Spalten in der CSV-Datei als Text-Spalten deklariert sind.
- In numerischen Spalten löscht Excel alle führenden Nullen.

- Zahlen mit führender Null (z. B. Postleitzahlen) werden zudem häufig als Ganzzahl interpretiert und die führende Null automatisch entfernt.
- Lange Zahlen werden als Exponentialzahl dargestellt.
- Zahlen können als Datumswerte interpretiert werden, obwohl es sich nicht um ein Datum handelt.

2. Methode: Kopieren und Einfügen

Beim Kopieren und Einfügen aus einer Nicht-Excel-Datei analysiert Excel die Spalten und deklariert alle Spalten, in denen nur Ziffern stehen, auch hier wieder automatisch als numerische Spalte. Eine Korrektur muss durch den Benutzer erfolgen.

3. Methode: Importieren mit dem Textkonvertierungs-Assistent

Tabellenkalkulationsprogramme prüfen beim Öffnen einer CSV-Datei z. B. den eingesetzten Zeichensatz und stellen den Inhalt der Datei korrekt dar. MS Excel verhält sich hier anders. Daher ist es der bessere Weg, die CSV-Datei nicht direkt zu öffnen, sondern durch den Textkonvertierungs-Assistenten zu importieren. Im Vergleich zum direkten Öffnen der CSV-Datei hat man hier die Möglichkeit, die Interpretation von Excel zu beeinflussen:

- Die Spaltenbreite wird an den Inhalt angepasst.
- Das Trennzeichen kann im Importdialog gewählt werden.
- Nach dem Öffnen der Datei können Spalten manuell als Textspalten deklariert werden, Spalten mit Zahlen werden also nicht mehr als numerische Spalten behandelt.

Weiterer Praxis-Kniff: Um zu verhindern, dass Excel eine automatische Interpretation der CSV-Datei vornimmt, sollte die CSV-Datei vorher durch Umbenennen zur Text-Datei (Suffix: *.txt) gemacht werden. Anschließend wird der Textkonvertierungs-Assistent aktiv, in dem eingestellt werden kann, wie die Datei interpretiert werden soll. Beachten: Die Textdatei nicht per Doppelklick öffnen, da sonst ein Texteditor gestartet wird.

Speichern der CSV-Datei in Excel

- Beim Speichern der CSV-Datei wird das Format durch Excel verändert (z. B. durch Programmartefakte wie Steuerzeichen), sodass die Datei nicht mehr dem Ursprungsformat entspricht. **Dies wird zu Fehlern bei der Weiterverarbeitung führen.**
- Die Excel-Arbeitsdatei sollte unter einem neuen Namen als CSV-Datei abgespeichert werden. Beim Speichern wird man gefragt – und das ein paar Mal –, ob man sich sicher ist, diese Datei im CSV-Format zu speichern. Antwortet man hier mit »Nein«, wird die Datei im Excel-Format (*.xls bzw. *.xlsx) gespeichert.

Fragen und Anmerkungen richten Sie bitte an

firmengrabungen@rps.bwl.de