Daten auswerten

Peter Willadt

2018-12-10

Wozu überhaupt?

- Betriebliche Software unterstützt nicht jede gewünschte Auswertung.
- Benutzer ist mit Tabellenkalkulation vertrauter als mit Datenbank.
- Aufbereitung für Präsentationen ist besser als mit betrieblicher Software.
- Aufbereiten von Daten aus einem Anwendungsprogramm für ein anderes.

Daten importieren

Batch-Export

- Originaldaten werden vor versehentlichen Beschädigungen geschützt.
- Mit etwas Glück ist das Exportformat dokumentiert.

Permanente Datenverbindung

- Daten können mit einem Mausklick aktualisiert werden.
- Tabellenkalkulation kann Zusammenfassungen abfragen, wenn der Umfang der Daten sehr groß ist.

Dateiformate

CSV weit verbreitet, auch bei älterer Software

XML bei moderner Software, im Internet-Umfeld

Text mit fester Spaltenbreite typisch für Druckausgaben

Es ist eher untypisch, dass betriebliche Software fehlerfreie Office-Dateien erzeugt.
Falls Sie einen komplizierten Import immer wieder machen müssen, kann Ihnen ein Programm in einer Skriptsprache wie *Perl* die Aufbereitung erleichtern.

Beispieldaten

CSV

Kunde;*Datum*;*Betrag*
Maier;13.11.2007;99,00 €

vielleicht aber auch so:

Kunde,*Datum*,*Betrag*
"Maier",11/13/07,99.00

XML

```
<Buchung>
    <Kunde>Maier</Kunde>
    <Datum>
        <Tag>13</Tag>
        <Monat>11</Monat>
        <Jahr>2007</Jahr>
    </Datum>
    <Betrag Waehrung="Euro">
        99.00</Betrag>
</Buchung>
```

XML

- XML (*extensible markup language*) ermöglicht strikte, maschinenüberprüfbare Datenformate.
- XML spielt auch eine wichtige Rolle beim Datenaustausch im WWW.
- XML kann automatisiert in das gewünschte Format gewandelt werden.
- XML-Bearbeitung erfordert viel Einarbeitung.

CSV

- CSV (*comma-separated values*) ist nicht wirklich standardisiert.
- EDIFACT ist ein Standard für kaufmännische Software, der auf CSV aufsetzt, aber genaue Datenbeschreibungen vorgibt.
- CSV-Export und Import ist fast immer verfügbar.
- CSV-Import in die Tabellenkalkulation lässt sich nicht automatisieren.

CSV-Definitionen

Satztrenner meist das Zeilenende (dessen tatsächliche Kodierung ist betriebssystemabhängig)

Feldtrenner in USA das Komma, in Deutschland das Semikolon

Texttrenner meist doppelte Anführungszeichen
Texttrenner werden benötigt, wenn in Texten der Feldtrenner vorkommt.

Dezimalpunkt In USA der Punkt, in Europa das Komma

CSV-Spezialitäten

Datumsformat Von tt.mm.jj bis jjjj-mm-tt wird alles verwendet.

Tausenderpunkte In USA das Komma, in Deutschland der Punkt, vielleicht auch ein Leerzeichen

Währungsangaben Vielleicht vorhanden, vielleicht ausgeschrieben, vielleicht als Symbol

Führende Nullen z.B. bei Postleitzahlen

Leerfelder am Zeilenende sind ein Problem, wenn die Daten mit anderer Software weiterverarbeitet werden.

CSV-Import in Excel

- Wenn eine Datei mit .CSV endet, gibt Excel Feldtrenner, Texttrenner usw. landestypisch vor.
- Falls die Vorgabe nicht funktioniert:
  - Datei umbenennen (.PRN, .TXT) oder
  - mit *Daten/Text in Spalten* nachbearbeiten.
- Zu Anfang Feldtrenner usw. setzen.
- Im nächsten Schritt Datenformate für kritische Felder (z.B. Datum) setzen (im Notfall »Text«)

Zeichensätze

Falls in den Daten auch Umlaute, das Euro-Zeichen oder ähnliches vorkommen, kontrollieren Sie dies nach dem Import. Wenn Sie statt der erwarteten Zeichen Datensalat erhalten, verändern Sie im ersten Schritt des Imports den Zeichensatz. Die häufigsten Zeichensätze für in Deutschland erzeugte Daten sind:

Unicode UTF8 für moderne Software, die z.B. in Java geschrieben ist und Import aus Linux

Windows-1252 für typische Windows-Software

DOS-850 für ältere Software

Texttrenner

- Wenn ein Textfeld den Feldtrenner enthält, sollte es in Texttrenner eingeschlossen werden, damit beim Import keine zusätzlichen Felder entstehen.
- Typischer Texttrenner ist das doppelte Anführungszeichen: `127;"Schraube; MS85";0,95`
- Wenn in einem Text der Texttrenner vorkommt, wird er
  - verdoppelt (`"Schraube; MS85 mit 3""-Gewinde"`) oder
  - maskiert (`"Schraube; MS85 mit 3\"-Gewinde"`),

  das lässt sich beim Import hoffentlich einstellen.

Überschriften

Falls die importierten Daten keine Überschriften haben, fügen Sie Überschriften zu, das erleichtert die Weiterverarbeitung. Achten Sie darauf, dass jede Spalte eine eindeutige Überschrift bekommt.
Es ist besser, wenn Überschriften aus einem Wort bestehen, verwenden Sie ggf. den Unterstrich oder notfalls CamelCase bei mehrteiligen Überschriften.

```
Lohn_brutto;LohnNetto
```

Daten prüfen

Was prüfen?

- Importfehler (Zahlen- und Datumsformat, Sonderzeichen, sporadische Zusatzfelder)
- Logische Fehler (Inkonsistenzen, unplausible Werte)

Und dann?

- Bei Importfehlern nochmals importieren.
- Offensichtliche Datenfehler korrigieren, auch in der exportierenden Software.
- Falls fragwürdige Daten ausgelassen werden: dokumentieren

Ausgeschlossene Daten dokumentieren

Wenn Sie Daten weglassen müssen, sollten Sie das dokumentieren, insbesondere

- wenn Sie Daten zu wissenschaftlichen Zwecken auswerten oder
- wenn die Daten buchhalterisch relevant sind.

Wie prüfen?

- Dezimalzahlen, Datumswerte, Texte mit Umlauten kontrollieren
- Extremwerte ansehen: Sortieren, `min()`, `max()`
- Filtern, Duplikate beseitigen

Sortieren

Sortieren verändert die Reihenfolge der Daten. Wenn Sie nur einen Teil der Spalten markieren, sind Ihre Daten kaputt.

- Geben Sie dem Datenblock einen Namen (mitsamt Überschriften).
- Markieren Sie den gesamten Datenblock, bevor Sie sortieren. (Neuere Excel-Versionen warnen Sie, falls Sie das nicht tun.)
- Wenn die Ursprungs-Reihenfolge wichtig ist, fügen Sie eine weitere Spalte mit fortlaufenden Nummern zu, bevor Sie sortieren.

Standardfunktionen

- Anzahl(), Anzahl2()
- Summe(), Mittelwert()
- Min(), Max()
- Einfache Berechnungen (z.B. Anzahl * Einzelpreis)

Weiteres

- Zählenwenn(), Summewenn()
- SVerweis()
- Gliedern
- Pivot-Tabelle

Zählen

- Filtern
- Anzahl(*Bereich*) zählt Zellen mit Zahlen, Anzahl2(*Bereich*) zählt Zellen, die nicht leer sind.
- Zählenwenn(*Bereich*;*Bedingung*) zählt Zellen, auf die die Bedingung zutrifft.
  - Bei Test auf Gleichheit können Sie den Vergleichswert einfach eintragen (Texte mit Anführungszeichen, Zahlen ohne):
    `=ZählenWenn($a2:$a20;100)` zählt alle Zellen im Bereich A2 bis A20, in denen die Zahl 100 steht.
  - Bei anderen Tests muss der komplette Test in Anführungszeichen, um den Formelparser zu überlisten:
    `=ZählenWenn($a2:$a20;">100")` zählt Zellen im Bereich mit Inhalten über 100.

Addieren, Durchschnittswerte bilden

- Summe(*Bereich*) addiert alle Werte im Bereich.
- SummeWenn(*Prüfbereich*; *Bedingung*; *Wertebereich*) addiert Werte im Wertebereich, falls die Bedingung auf den Prüfbereich zutrifft.
- Mittelwert(*Bereich*) bildet das arithmetische Mittel.
- Modalwert(*Bereich*) liefert den häufigsten Wert im Bereich.