Anbietungs- und Archivierungsformate für Datenbanken und Datentabellen

Richtlinie des Landesarchivs Baden-Württemberg

Stand: 5.3.2023

1. Die Datenbanken und Datentabellen enthalten lesbaren Text oder Zahlen im Zeichenformat Unicode (UTF-8 oder UTF-16).
2. Die Daten und die Feldbeschreibungen werden in einem definierten CSV-Format wiedergegeben. Nach Rücksprache kann anstelle des CSV-Formats auch ein anderes geeignetes Format (z.B. generisches XML mit Schema, SIARD, JSON) vereinbart werden.
3. Die Bedeutung der einzelnen Felder, der Datentabellen und eventuelle Zusammenhänge sind in geeigneter Weise dokumentiert.
4. CSV-Definition

4.1 Jede Zeile enthält einen Datensatz. Die Zeilen werden durch die Zeichen `CRLF` oder `LF` voneinander getrennt.

4.2 Die letzte Zeile kann mit `CRLF` oder `LF` enden.

4.3 In der ersten Zeile sind die Namen der in der Tabelle repräsentierten Datenfelder als Überschriften wiedergegeben. Die Zahl der Überschriftenfelder muss der Zahl der Felder pro Zeile entsprechen.

4.4 In allen Zeilen sind die Felder durch Semikolon voneinander getrennt. Leerzeichen werden als Bestandteil eines Felds betrachtet und nicht ignoriert.

4.5 Jedes Feld kann mit Anführungszeichen (`"`) an Anfang und Ende begrenzt werden.

4.6 Felder, die Zeilentrenner (`CRLF`), Anführungszeichen oder Semikola enthalten, sind mit Anführungszeichen an Anfang und Ende zu begrenzen.

4.7 Wenn Anführungszeichen zur Begrenzung eines Felds benutzt werden, muss jedes Anführungszeichen innerhalb eines solchen Feldes durch ein zweites Anführungszeichen aufgehoben werden, um die Bedeutung als Feldbegrenzer zu anullieren.

Erläuterungen

Zu Punkt 2: Alternativen

Bei komplexeren Datenbanken kann es sinnvoll sein, die Daten in einem anderen geeigneten Format abzugeben. In diesen Fällen bedarf es weiterer Vereinbarungen zwischen der abgebenden Behörde und dem Landesarchiv.

Zu Punkt 4: CSV-Formatdefinition

Das Format CSV ist ein platzsparendes Mittel zur Ablage textbasierter Daten in Tabellenform. Hinsichtlich der Feldtrenner (field separators, field delimiters) und der Textbegrenzer (text delimiters) existieren viele Varianten. Es kann zu Problemen kommen, wenn die zur Tabellenauszeichnung vorgesehenen Zeichen auch in den Feldern der Tabelle auftreten. Solche Felder müssen mit Textbegrenzern maskiert werden.

Bei der IETF (Internet Engineering Task Force) liegt seit Oktober 2005 die MIME-Type-Definition RFC 4180 vor. Sie ist offiziell registriert, ist öffentlich zugänglich und definiert das Format ausführlich: https://www.rfc-editor.org/rfc/rfc4180.txt.

Die Richtlinie des Landesarchivs entspricht daher weitgehend der Definition der RFC 4180, mit zwei Ausnahmen:

- Als Feldtrenner wird in der RFC 4180 nur das Komma genannt. Da die meisten Anwendungen im deutschsprachigen Raum stattdessen ein Semikolon benutzen, schreibt das Landesarchiv ebenfalls die Benutzung eines Semikolon vor.
- Außerdem ist die Einfügung von Feldnamen in der ersten Tabellenzeile verpflichtend.