Tabellarische Annotationen in der Sprachwissenschaft

Use Case

In einer linguistischen Untersuchung werden verschiedene im 19. Jahrhundert publizierte Texte auf bestimmte sprachliche Merkmale und den Kontext, in dem bestimmte Bezeichnungen verwendet werden, untersucht. Für die Organisation und Auswertung der Daten wurde ein gängiges Tabellenkalkulationsprogramm (Calc, Excel,...) genutzt. Darin wurden die relevanten Textstellen gesammelt, annotiert, klassifiziert und kodiert. Dabei wurden die einzelnen Daten jeweils in eigene Spalten eingetragen, etwa die Referenz zur Quelle, die Textstelle selbst oder die Kodierung. Alle Abkürzungen, Kodes und Klassen werden in einer separaten Dokumentation aufgelistet und beschrieben.

Kontext

- Geistes- und Sozialwissenschaften
- Geisteswissenschaften
- Sprachwissenschaften
- Tabellenformat
- Annotation

Empfohlenes Vorgehen

Da in dieser Tabelle keine besonderen Formatierungen, Formeln zur Berechnung von Feldinhalten oder Makros genutzt wurden, die für eine spätere Verwendung der Daten essentiell sind, wurde empfohlen die Tabelle in das CSV-Format (Comma-Separated Values) zu exportieren. Unter Formatierungen fällt hier beispielsweise die Nutzung von Farben (Einfärbung von Zellen, Textfarbe), um Informationen zu vermitteln. Idealerweise sollten Farben höchstens zur besseren Visualisierung für den Menschen, aber nicht als alleiniger Informationsträger verwendet werden. Stattdessen bietet sich an, den Informationsinhalt nochmals gesondert in einer zusätzlichen Spalte zu hinterlegen. Diese Spalte - und damit die Information - wird dann beim Export in das CSV-Format erhalten bleiben.

Grund

Für die Langzeitverfügbarkeit von Daten sollten grundsätzlich offene, nicht-proprietäre Dateiformate gewählt werden. Bei Microsoft Office wird beispielsweise für Tabellenkalkulationen das Format XLSX verwendet. Das Format ist grundsätzlich nach dem Office Open XML Standard

gestaltet, jedoch ist es ohne entsprechende Programme nur mit Aufwand möglich, die Daten aus diesem sehr komplexen Standard zu extrahieren. Das CSV-Format hingegen kann sowohl mithilfe verschiedener Tabellenkalkulationsprogramme als auch mit jedem einfachen Texteditor gelesen und bearbeitet werden.

Konsequenzen und Kosten

Für die nähere Zukunft werden auch Dateien im XLSX-Format weiterhin zugänglich und damit lesbar sein. Im Sinne der Langzeitverfügbarkeit erschweren komplexe Dateiformate (wozu auch das XLSX-Format gehört) eine dauerhafte Verfügbarkeit und Kuratierung. In diesen Fällen kann deshalb möglicherweise nur eine technische Erhaltung der Daten sichergestellt werden.

Für eine Speicherung im CSV-Format entstehen keine weiteren Kosten und nur minimaler Zeitaufwand, da CSV-Dateien beispielsweise direkt aus Microsoft Excel heraus gespeichert werden können.

Nach der Speicherung und vor dem Import in ein Langzeitarchivierungssystem sollten die Daten unbedingt auf Vollständigkeit und vor allem auf die Korrektheit der darin enthaltenen Daten überprüft werden, um möglicherweise auftretende Formatierungsfehler beheben zu können.

Weitere Hinweise

- CSV (Wikipedia)
- Office Open XML (Wikipedia)