Datenmanagement – Gegenstand und Dienst der Computerlinguistik

Thorsten Trippel
Eberhard Karls Universität Tübingen
thorsten.trippel@uni-tuebingen.de

Datenmanagement wird durch die Forschungsföderungsorganisationen (etwa in Horizon 2020 der EU, die Allianz der deutschen Wissenschaftsorganisationen oder in DFG geförderten Projekten) mehr und mehr Teil der Forschungslandschaft. Für die Computerlinguistik ist das Forschungsdatenmanagement aber auch Teil des Forschungsgebietes: Datenmodellierung und Transformation für die nachhaltige Datenspeicherung gehören in den Bereich der Texttechnologie und Textlinguistik, ebenso die Modellierung der beschreibenden Daten zu Datensätzen.

Die Anreicherungen der Metadaten etwa durch die Erkennung der Sprache in einem Datensatz können als Gegenstand der automatischen Sprachverarbeitung gesehen werden, die Erstellung von beschreibenden Datenkategorien und deren Definition dagegen als angewandte Lexikographie. Gleichzeitig dienen die Forschungsdaten und Metadaten als Grundlage für Fragestellungen der semantischen Netze und damit dem Forschungsgebiet der Linked Data.

Die FAIR-Prinzipien als Grundphilosophie für das wissenschaftliche Datenmanagement setzen für sprachliche Inhalte voraus, dass Werkzeuge zur Suche und zur Weiterverarbeitung zur Verfügung stehen, durch die Forschungsdaten aufgefunden, zugänglich, interoperabel und nachnutzbar werden. Forschungsinfrastrukturen wie CLARIN-D (siehe Hinrichs & Trippel, 2017) haben daher neben einem Servicecharakter für die Linguistik einen starken Forschungsschwerpunkt in der Computerlinguistik. Auf dem Poster werden wir am Beispiel von CLARIN zentralen Werkzeuge und Dienste im Datenmanagement mit (computer-) linguistischen Methoden und Ansätze darstellen.

References: Hinrichs, E. & Trippel, T. (2017). CLARIN-D: eine Forschungsinfrastruktur für die sprachbasierte Forschung in den Geistes- und Sozialwissenschaften. *Bibliothek Forschung und Praxis*, 41(1), pp. 45-54. DOI :10.1515/bfp-2017-0015