Das Labeling System – ein freier Baukasten für kontrollierte Vokabulare

Michael Piotrowski
Florian Thiery
Kai-Christian Bruhn

10. November 2014

Maschinenlesbare Annotationen sind die Voraussetzung für die semantische Verarbeitung von Daten. Diese Aussage gilt unabhängig davon, ob es sich bei den Daten um natürlichsprachigen Text oder um strukturierte Datensätze in einer Datenbank handelt, und unabhängig davon, ob es um einfaches Sortieren und Filtern geht oder um komplexes automatisches Schließen. Kontrollierte Vokabulare (ob in einer einfachen Terminolgieliste oder als Taxonomien, Thesauri oder Ontologien strukturiert) sind dabei unbedingt notwendig, um Annotationen maschinell verarbeitbar zu machen; ohne terminologische Kontrolle sind Annotationen für die maschinelle Verarbeitung kaum nützlicher als an den Rand eines Buches gekrizelte Notizen. Kontrollierte Vokabulare abstrahieren von natürlichsprachlichen Ambiguitäten und Konnotationen; sie sind daher entscheidend für die semantische Verarbeitung von Forschungsdaten. Um projektübergreifende Zusammenarbeit und den semantischen Austausch von Daten zu ermöglichen, müssen Vokabulare nicht nur kontrolliert, sondern auch formell oder informell standardisiert sein. Standardisierte kontrollierte Vokabulare ermöglichen den Austausch, die Kombination und die gemeinsame Analyse annotierter Daten aus verschiedenen Quellen sowie die Implementierung generischer Werkzeuge für die semantische Verarbeitung.

Erstellung und Wartung standardisierter kontrollierter Vokabulare sind jedoch zeitaufwändig und damit teuer. Zu den größten Herausforderungen zählen, dass alle beteiligten Parteien zu einem gemeinsamen Verständnis der Begriffe kommen, und dass die richtige Balance zwischen möglichst breiter Anwendbarkeit einerseits und möglichst präziser Analyse andereseits gefunden werden. Diese Ziele sind insbesondere in den Geisteswissenschaften schwierig zu erreichen: nicht nur sind die Forschungsfragen, die potentiell an einen gegebenen Datensatz gerichtet werden können, extrem weit gefächert, sondern die Kategorisierung der Daten ist häufig ein essenzieller Teil des Forschungsprozesses selbst. Es gibt daher einen eklatanten Mangel an standardisierten kontrollierten Vokabularen in den Geisteswissenschaften, der Digital-Humanities-Projekte letztlich dazu zwingt, eigene, projektspezifische Vokabulare zu definieren. Projektspezifische Vokabulare lösen können den internen Bedarf zwar kurzfristig befriedigen, sind aber nicht interoperabel und verhindern den zukünftigen Austausch und die Nachnutzung der annotierten Daten.

Unser Poster stellt einen neuen konzeptuellen Ansatz zur Lösung dieser Probleme vor und beschreibt die Implementierung dieses Ansatzes in einem Softwarewerkzeug, dem *Labeling System*.

Da es in der geisteswissenschaftlichen Forschung praktisch unmöglich ist, kontrollierte Vokabulare zu definieren, die alle denkbaren Anwendungen abdecken und generell akzeptiert sind, schlagen wir ein anderes Vorgehen vor. Bei unserem Ansatz definieren Projekte ihre eigenen Vokabulare, aber anstelle natürlichsprachlicher

Definitionen werden die Terme mit einem oder mehreren Konzepten in einem Referenzthesaurus verknüpft. Der projektspezifische Term dient also quasi als »Label« für eine Menge gemeinsamer Konzepte. Dieser Ansatz ermöglicht es Projekten Vokabulare entsprechend ihrer Bedürfnisse und unter Verwendung im jeweiligen Forschungsgebiet üblichen Bezeichnungen benutzen, während gleichzeitig die Interoperabilität mit anderen Projekten über den Referenzthesaurus gewährleistet ist.

Das Labeling System ist eine Webanwendung, die es Benutzern ermöglicht, SKOS-Vokabulare zu erstellen und auf einfache Weise deren Terme mit einem oder mehreren Konzepten in einem oder mehreren Referenzthesauri zu verknüpfen. Die Benutzeroberfläche ermöglicht die Visualisierung der definierten Vokabulare in einer hierarchischen Baumstruktur und ermöglicht den Zugriff auf Vokabulare über eine SPARQL-Schnittstelle. Das Labeling System basiert auf ausgereiften Open-Source-Komponenten und ist selbst ebenfalls frei verfügbar.