Das Labeling System – ein freier Baukasten für kontrollierte Vokabulare

Michael Piotrowski, Florian Thiery, Kai-Christian Bruhn

DARIAH-DE
Digital Research Infrastructure for the Arts and Humanities

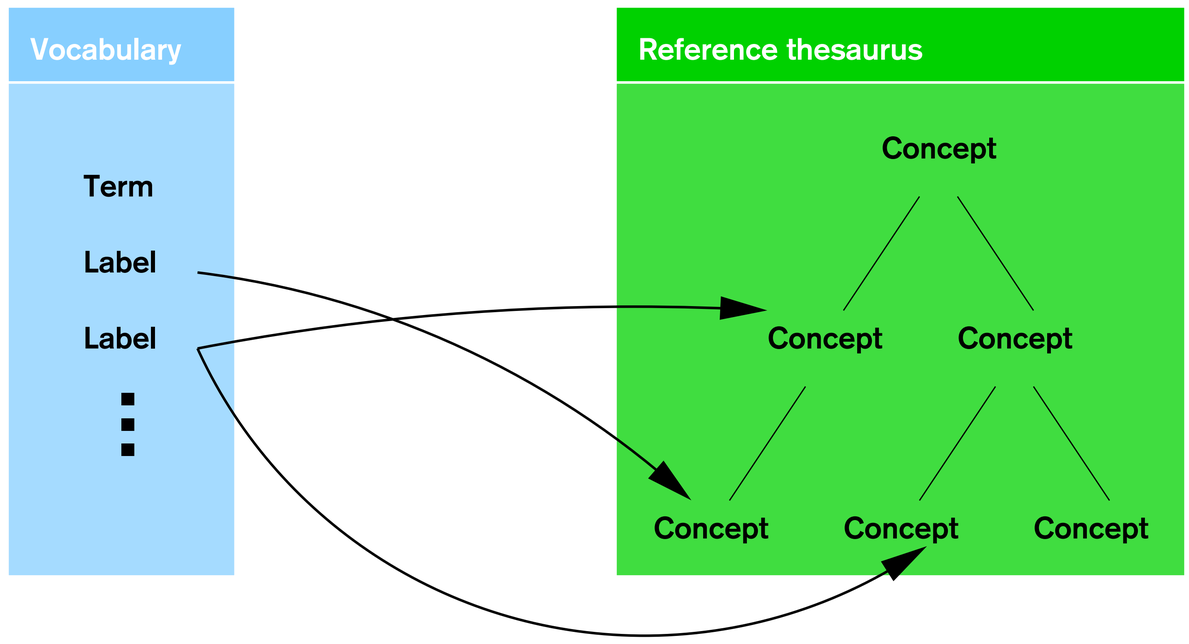
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Problem

- Maschinenlesbare Annotationen sind Voraussetzung für semantische Verarbeitung von Daten.
- Kontrollierte Vokabulare sind notwendig, um Annotationen maschinell verarbeitbar zu machen.
- Kontrollierte Vokabulare abstrahieren von natürlichsprachlichen Ambiguitäten und Konnotationen: entscheidend für die semantische Verarbeitung von Forschungsdaten.
- Für projektübergreifende Zusammenarbeit und semantischen Datenaustausch müssen Vokabulare nicht nur kontrolliert, sondern auch formell oder informell standardisiert sein.
- Standardisierte kontrollierte Vokabulare ermöglichen Austausch, Kombination und gemeinsame Analyse annotierter Daten aus verschiedenen Quellen sowie Implementierung generischer Werkzeuge.
- Erstellung und Wartung standardisierter kontrollierter Vokabulare sind jedoch teuer.
- Alle beteiligten Parteien müssen zu gemeinsamem Verständnis der Begriffe kommen und Balance zwischen möglichst breiter Anwendbarkeit und möglichst präziser Analyse andereseits finden.
- Ziele insbesondere in den Geisteswissenschaften schwierig zu erreichen: Potentielle Forschungsfragen extrem weit gefächert, Kategorisierung der Daten ist häufig ein essenzieller Teil des Forschungsprozesses selbst.

→ **Eklatanter Mangel an standardisierten kontrollierten Vokabularen in den Geisteswissenschaften.** Digital-Humanities-Projekte sind gezwungen, eigene, projektspezifische Vokabulare zu definieren. Projektspezifische Vokabulare können internen Bedarf kurzfristig befriedigen, sind aber **nicht interoperabel und verhindern zukünftigen Austausch und Nachnutzung annotierter Daten.**

Ansatz



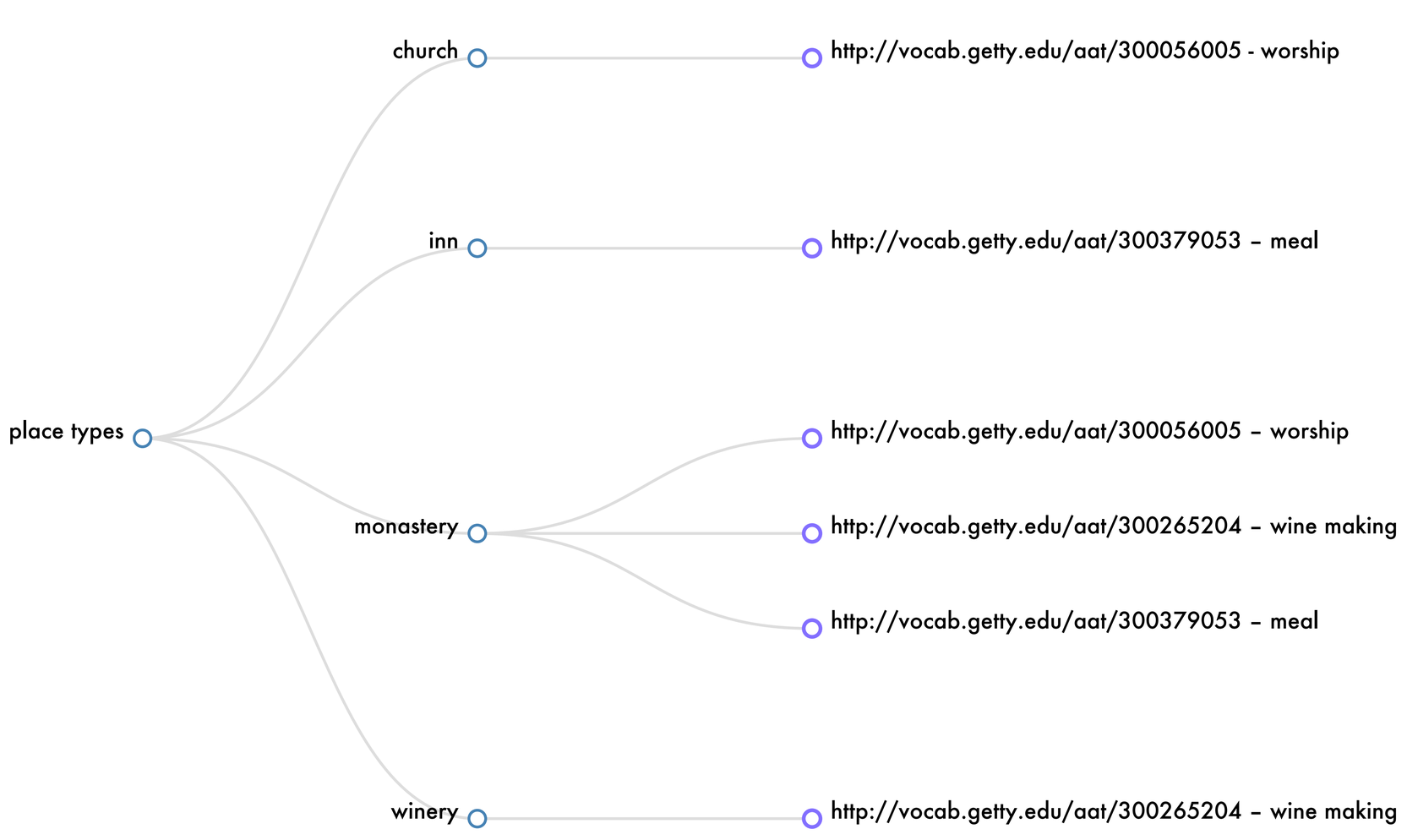
Da es in der geisteswissenschaftlichen Forschung praktisch unmöglich ist, kontrollierte Vokabulare zu definieren, die alle denkbaren Anwendungen abdecken *und* generell akzeptiert sind, schlagen wir ein anderes Vorgehen vor.

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 der im jeweiligen Forschungsgebiet üblichen Bezeichnungen benutzen, während gleichzeitig die Interoperabilität mit anderen Projekten über den Referenzthesaurus gewährleistet ist.

Implementierung: Das Labeling System

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 (`broadMatch`, `narrowMatch`, `closeMatch`, `exactMatch`, `relatedMatch`, `seeAlso`, `sameAs` und `isDefinedBy`).
Die Benutzeroberfläche ermöglicht die Visualisierung der definierten Vokabulare in einer Baumstruktur. Mit der **CSV-Upload-Funktion** kann ein CSV-Dokument hochgeladen werden, das mehrere `prefLabels`, `altLabels`, `notes` und `definitions` eines Labels enthält und zudem Relationen `broader/narrower/related` innerhalb von Vokabularen sowie die Links zu externen Ressourcen beschreibt.
Externer Zugriff auf Vokabulare über **SPARQL-Schnittstelle**.
Das Labeling System basiert auf ausgereiften **Open-Source**-Komponenten und ist selbst ebenfalls **frei verfügbar**.



Institut für Raumbezogene Informations- und Messtechnik
Hochschule Mainz
University of Applied Sciences

IEG
Leibniz-Institut für Europäische Geschichte

http://labeling.i3mainz.hs-mainz.de/

Michael Piotrowski <piotrowski@ieg-mainz.de>
Florian Thiery <florian.thiery@hs-mainz.de>
Kai-Christian Bruhn <bruhn@hs-mainz.de>