Horticulture Semantic (HortiSem)

Eine Service-Infrastruktur für automatisierte Annotation, Named Entity Linking, Suche und Abfrage von Informationsressourcen für den Gartenbau

Jascha Daniló Jung[1] und Daniel Martini[1]

Abstract: Im Projekt HortiSem wird ein semantisches Netzwerk entwickelt, das speziell für den Bereich Landwirtschaft gedacht ist. Das Projekt hat inzwischen wichtige Fortschritte gemacht. Zum einen wurde die API „destreak“ entwickelt, die es ermöglicht, vorgefertigte Abfragen mithilfe von http-Requests auszuführen. Dadurch können einfache Abfragen leicht in Suchmasken integriert werden, ohne dass eine SPARQL-Query formuliert werden muss. Zudem wurde ein Crawler implementiert, der neue Warndienstmeldungen von hortigate.de automatisch herunterlädt, annotiert, ins RDF-Format umwandelt und abschließend auf den Triplestore hochlädt. Dies zeigt, wie der Knowledge Graph von HortiSem kontinuierlich wachsen kann und neue Informationen automatisch hinzugefügt werden können. Darüber hinaus ist es möglich, einen eigenen Triplestore aufzusetzen und ihn mit den Daten von HortiSem zu verknüpfen. Dadurch können auch Daten integriert werden, die nicht öffentlich zugänglich sein sollen.

Keywords: Datenmanagement, Smart und Big Data, künstliche Intelligenz, Machine Learning, Natural Language Processing, Linked Open Data, semantisches Netzwerk

1 Einleitung

Im Horticulture Semantic (HortiSem) Projekt wird ein semantisches Netzwerk speziell für den Bereich der Landwirtschaft entwickelt [Ju22]. Dieses Netzwerk verbindet Informationen und stellt sie in Beziehung zueinander. Der Fokus liegt auf Begriffen aus den Kategorien „Kultur“, „Schädlinge“, „Nützlinge“ und „Pflanzenschutzmittel“. Der Schwerpunkt dieses Beitrags liegt auf Zugriffs- und Nutzungsmöglichkeiten, insbesondere der parallel entwickelten API „destreak“, die es ermöglicht, vorgegebene Abfragen an den Knowledge Graphen zu stellen, indem komplizierte SPARQL-Queries im Backend musterhaft implementiert und als vorgefertigte Vorlagen über einfachere HTTP-Requests gestellt werden können. Diese lassen sich auch leicht in bereits vorhandene Suchmasken implementieren. Ein Crawler wurde ebenfalls entwickelt, der neue Daten von hortigate[2] automatisch erkennen und verarbeiten kann. Im Folgenden soll diskutiert werden, welche Vorteile die Verwendung der „destreak“ API im Vergleich zu

1 Kuratorium für Technik und Bauwesen in der Landwirtschaft, Datenbanken und Wissenstechnologien, Bartningstraße 49, 64289 Darmstadt, j.jung@ktbl.de
2 https://www.hortigate.de/

konventionellen SPARQL-Queries im Zusammenhang mit der Nutzung des Knowledge Graphen von HortiSem im Bereich der Landwirtschaft bietet.

2 Bedeutung semantischer Netzwerke für die Landwirtschaft

Semantische Netzwerke sind ein Bereich der Künstlichen Intelligenz, der sich mit der Verarbeitung und Interpretation natürlicher Sprache befasst. Sie werden häufig verwendet, um Bedeutungen von Wörtern und Sätzen in einem bestimmten Kontext zu verstehen und zu analysieren. Die Forschung an und mit semantischen Netzwerken hat in den letzten Jahren deutliche Fortschritte gemacht. Eine der Herausforderungen in diesem Bereich ist es, das Verständnis von semantischen Netzwerken für die Verarbeitung natürlicher Sprache zu verbessern, insbesondere im Hinblick auf die sogenannte „kontextuelle Bedeutung“ von Wörtern und Sätzen. Diese kontextuelle Bedeutung ist wichtig, um die Bedeutung von Wörtern und Sätzen in einem bestimmten Kontext zu verstehen und zu analysieren.

In der Landwirtschaft könnten semantische Netzwerke beispielsweise zur Analyse von Wetterdaten benutzt werden, oder um Anweisungen für landwirtschaftliche Maschinen zu verstehen und umzusetzen, wodurch die Arbeit effizienter automatisiert werden könnte.

Insgesamt bieten semantische Netzwerke vielversprechende Möglichkeiten für die Landwirtschaft und könnten dazu beitragen, die Effizienz und Präzision landwirtschaftlicher Prozesse zu verbessern. Die andauernde Forschung in diesem Bereich wird dazu beitragen, die Anwendungsmöglichkeiten von semantischen Netzwerken weiter zu erweitern und zu verbessern.

3 HortiSem Workflow: Daten, Methoden und Infrastruktur

3.1 Übersicht

Der Workflow von HortiSem wird in Abbildung 1 gezeigt (s. unten).

Kurz zusammengefasst werden Daten aus dem Web gecrawlt und mit einem NER-Model annotiert. Die annotierten Daten werden in ein RDF-Format gemappt und in einem Triplestore gespeichert. Dort stehen sie durch einen öffentlichen SPARQL-Endpoint zur Verfügung. Zusätzlich ermöglicht die API „destreak“ einen vereinfachten Abruf der Daten per http-Requests. Die Daten im Triplestore verweisen wiederum auf die Datenquelle.

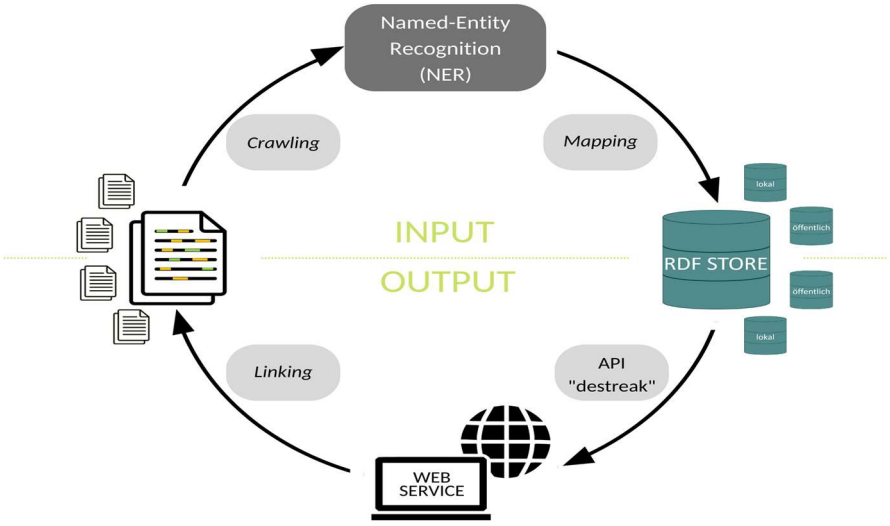

Abb. 1: Der HortiSem Workflow

Die einzelnen Schritte werden in den folgenden Unterkapiteln kurz beschrieben.

3.2 Knowledge Graphen: die Grundlage des semantischen Netzwerks

Ein semantisches Netzwerk ist ein Netzwerk aus Informationen, das Beziehungen und Bedeutungen zwischen Daten veranschaulicht. Im Falle des HortiSem-Projekts bilden bekannte Datenquellen über Kulturen, Schädlinge, Nützlinge und weitere Begriffe aus der Landwirtschaft die Basis (etwa aus dem AGROVOC Thesaurus[3] und von PS Info[4]). Diese Datenquellen werden in das RDF-Format gemappt und bilden das „Grundgerüst“ des Knowledge Graphen. Um den Knowledge Graphen zu erweitern, werden Beziehungen von neuen Daten zu bereits vorhandenen Daten beschrieben und das Netzwerk aus Informationen so erweitert.

3.3 Erweiterung des semantischen Netzwerks: neue Daten

Im Projekt HortiSem wurde ein automatisiertes Verfahren entwickelt, das neue Daten im RDF-Format in den Triplestore hochladen kann. Dabei werden die neusten PDF-Dateien von hortigate.de automatisch heruntergeladen, in Text-Format umgewandelt und bereinigt. Anschließend werden sie mit einem NER-Model annotiert, um Entitäten zu identifizieren. Gefundene Entitäten werden dann mit vorhandenen Daten im Knowledge

3 https://agrovoc.fao.org/browse/agrovoc/en/
4 https://www.pflanzenschutz-information.de/

Graphen abgeglichen und bei einem Treffer eine Verlinkung erstellt. Die PDF-Dateien werden dann in eine RDF-Repräsentation umgewandelt, die Metadaten zum Text sowie gefundene und verifizierte Entitäten mit entsprechenden Verlinkungen zum Knowledge Graphen enthält. Die RDF-Dateien werden schließlich in den Triplestore hochgeladen, während die PDF-Dateien gelöscht werden, da sie oft kostenpflichtige Informationen enthalten und daher nicht im Knowledge Graphen gespeichert werden. Stattdessen wird im Knowledge Graphen per URL auf die Quelle der Publikation verwiesen.

3.4 Vereinfachte Abfragen: die „destreak“ API

Triplestores wie Fuseki[5] ermöglichen die Abfrage von Daten mithilfe von SPARQL-Queries. Im Rahmen des Projekts HortiSem soll es einen frei verfügbaren SPARQL-Endpoint geben. SPARQL-Abfragen können für Laien allerdings sehr kompliziert sein, da sie eine spezielle Syntax und Kenntnisse in der Semantik der verwendeten Daten erfordern. Daher wird die API „destreak“ entwickelt.

Bei „destreak“ handelt es sich um eine API, die http-Requests verwendet. Sie kann eine einfachere Alternative zu SPARQL-Abfragen sein, da sie es Benutzern ermöglicht, Anfragen in einer intuitiver zu verwendenden Syntax zu stellen. Diese API ist nicht nur für Laien einfacher zu verwenden, sondern auch einfacher in bereits vorhandene Systeme zu integrieren. Im Gegensatz zu SPARQL-Abfragen, die spezielle Kenntnisse und Fähigkeiten erfordern, können http-Requests von vielen verschiedenen Systemen und Programmiersprachen verwendet werden. Dadurch wird der Zugang zu den Daten vereinfacht und ihre Verwendung benutzerfreundlicher.

Die Implementation von „destreak“ erfolgt mit FastAPI[6].

3.5 Vollständige Automatisierung

Der Prozess zum Einpflegen neuer Daten kann vollständig automatisiert werden. Der Crawler für hortigate.de beispielsweise ruft die neuste im Knowledge Graphen vorhandene Publikationsnummer ab und prüft dann die Webseite auf neuere Publikationen. Dieser Prozess kann regelmäßig automatisch durchgeführt werden. Neue Warndienstmeldungen können dann automatisch heruntergeladen, ihre wichtigen Inhalte und Metadaten in das RDF-Format umgewandelt und dem Triplestore hinzugefügt werden. Damit ist auch die Abfrage tagesaktueller Informationen möglich.

5 https://jena.apache.org/documentation/fuseki2/
6 https://fastapi.tiangolo.com/

3.6 Die Infrastruktur von HortiSem

Der Code, der dem gezeigten HortiSem Workflow zugrunde liegt, ist derzeit nicht öffentlich zugänglich. Nur der SPARQL-Endpoint ist testweise verfügbar. Zusätzlich ist geplant, eine Version von „destreak“ öffentlich zu machen. Es wird daran gearbeitet, die Setup-Informationen für lauffähige Docker-Container zur Verfügung zu stellen. Es ist auch möglich, einen eigenen Triplestore neben dem für das Projekt am KTBL aufgesetzten Triplestore zu hosten (siehe Abb. 2). Auf diesem lokalen Server könnten Unternehmen auch sensible Daten speichern. Auch können eigene Crawler und RDF-Mappings erstellt werden, um den lokalen Server damit zu erweitern. Die Daten von HortiSem können dann als Referenzpunkt und Teil des eigenen Knowledge Graphen verwendet werden. Das Ziel des Projekts ist es, eine große Zahl relevanter Daten in den Knowledge Graphen aufzunehmen und miteinander zu verknüpfen. Hierzu sollen neben Referenzen zu bereits vorhandenen Datenbeständen auch neue Knowledge Graphen hinzugefügt werden.

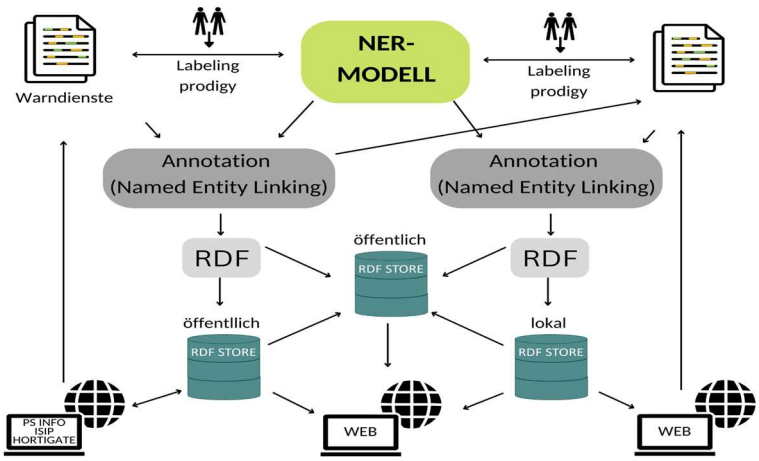

Abb. 2: Workflow und HortiSem-Infrastruktur

3.7 GUI

Für die vereinfachte Nutzung wurde für den Crawler, das automatische Mapping und den Upload der Daten auf den Triplestore ein einfaches GUI (Graphical User Interface) entwickelt. Dieses ermöglicht es, die neuesten Daten von hortigate.de automatisch herunterzuladen, zu annotieren und anschließend entweder auf einen lokalen Server oder direkt auf den KTBL-Server hochzuladen. Neben der Verwendung des Crawlers können auch eigene PDF-Dateien für die Annotation und das Mapping oder RDF-Dateien für den Upload auf den Triplestore von der Festplatte ausgewählt werden. Das GUI wird hauptsächlich für Testzwecke verwendet, um die automatisierten Schritte kontrolliert

starten und die Ergebnisse überprüfen zu können. Es dient aber auch zu Demonstrationszwecken, da ein solches Interface deutlich einfacher zu verstehen ist als die Verwendung von Shell- oder Command-line-Interfaces.

4 Fazit

Im Projekt HortiSem wurden Mappings für Daten aus verschiedenen Quellen erstellt und in den Knowledge Graphen eingefügt. Die Automatisierung des Prozesses der Umwandlung von PDF-Dateien in RDF-Dateien mit Bezug zum Knowledge Graphen wurde mit hortigate-Warndienstmeldungen demonstriert. Durch die Verwendung der „destreak“ API können Abfragen vereinfacht in Suchmasken integriert werden, sodass für häufige Abfragen keine komplizierten SPARQL-Abfragen erforderlich sind. Neue Daten können automatisch durch einen Crawler in den Knowledge Graphen hinzugefügt werden, wobei es möglich ist, eigene Daten auf einem eigenen Server zu hosten und mit den Daten von HortiSem zu verknüpfen. Die aktuelle Entwicklung konzentriert sich auf die Erweiterung des Knowledge Graphen und das Containerisieren von Triplestore und destreak, um das Hosten in „abgeschlossenen“ Umgebungen zu ermöglichen. Schließlich wird noch mit der automatischen Erweiterung des Netzwerks experimentiert, um HortiSem aktuell zu halten.

Literaturverzeichnis

[Ju22] Jung, J. D.; He, X.; Martini, D.; Golla, B.: Horticulture Semantic (HortiSem) – Natural Language Processing bei Entwicklung und Interaktion mit einem semantischen Netzwerk für die Landwirtschaft. In: Gandorfer, M., Hoffmann, C., El Benni, N., Cockburn, M., Anken, T. & Floto, H. (Hrsg.), 42. GIL-Jahrestagung, Künstliche Intelligenz in der Agrar- und Ernährungswirtschaft. Bonn: Gesellschaft für Informatik e.V.. (S. 141-146). 2022.