Verschiedene Sichtweisen – verschiedene Sprachen: Codesysteme für landwirtschaftliche Kulturen und wie sich Interoperabilitätsbarrieren überwinden lassen

Daniel Martini [1], Esther Mietzsch[1], Nils Reinosch[1], Jascha Jung[1] und Desiree Batzer-Kaufmann[1]

Abstract: In landwirtschaftlichen Datenbeständen werden für wichtige Konzepte wie Maschinen, Betriebsmittel, Kulturen oder Produkte unterschiedliche Bezeichner verwendet. Dabei werden in verschiedenen Systemen oder im Kontext verschiedener Anwendungsfälle teilweise auch für äquivalente oder zumindest nahverwandte Konzepte unterschiedliche Bezeichner genutzt. Dies führt zu Herausforderungen bei der Zusammenführung von Daten oder für die Interoperabilität im Datenaustausch. Wenn gemäß Spezifikationen des Semantic Web jedoch jedem Konzept eine URI als global eindeutiger Identifier zugewiesen wird, lassen sich Identitäten und Äquivalenzbeziehungen zwischen Konzepten über entsprechende Prädikate beschreiben. Werden die URIs anschließend auch in Datenbeständen genutzt, führt dies im Zusammenspiel mit den Beschreibungen der Beziehungen ohne weiteres Zutun zu einer Verknüpfung mit verwandten Konzepten. Gemeinsame Auswertungen von Daten aus verschiedenen Quellen lassen sich so ohne vorherige, aufwändige Transformation durchführen.

Keywords: Semantic Web, Interoperabilität, Identifikationssysteme, Wissensorganisation

1 Einleitung

In der Landwirtschaft existieren eine Vielzahl anwendungsfallspezifischer oder an bestimmte Softwaresysteme gebundener Datenbestände, die in unterschiedlichen Datenmodellen und -formaten abgebildet sind. Innerhalb dieser Datenrepräsentationen werden für wichtige Konzepte wie beispielsweise Maschinen, Betriebsmittel, Kulturen oder Produkte unterschiedliche Bezeichner verwendet. Zweck dieser Bezeichner ist eine eindeutige Identifikation, sodass diese beispielsweise in einem relationalen Datenbanksystem auch als Primärschlüssel herangezogen werden können. Im Folgenden werden Zusammenstellungen von solchen schlüsselartig verwendbaren Bezeichnern, also die Menge von in einem bestimmten Kontext nutzbaren Codes, „Codesysteme“ genannt.

[1] Kuratorium für Technik und Bauwesen in der Landwirtschaft e.V. (KTBL), Datenbanken und Wissenstechnologien, Bartningstraße 49, 64289 Darmstadt, d.martini@ktbl.de, https://orcid.org/0000-0002-6953-4524; e.mietzsch@ktbl.de, n.reinosch@ktbl.de, j.jung@ktbl.de, d.batzer-kaufmann@ktbl.de

Heutzutage genutzte Codesysteme haben meist zwei Schwachpunkte: Zum einen sind die Bezeichner jeweils nur innerhalb ihres Kontextes wirklich eindeutig. Das heißt, sie beziehen sich in der Regel nur auf die Bezeichnung von Entitäten innerhalb bestimmter Softwaresysteme oder gar auf die Bezeichnung nur bestimmter Klassen von Objekten. Werden zum Beispiel im Umfeld relationaler Datenbanken übliche, seriell erzeugte, ganzzahlige Nummern als Bezeichner verwendet, so sind diese bereits außerhalb ihres Tabellenkontextes nicht mehr eindeutig. Auch systemübergreifend standardisierte Codesysteme sind außerhalb ihres Kontextes oft nicht mehr eindeutig. Überlappungen von Wertebereichen treten z. B. zwischen InVeKoS-Kulturcodes und DDI-Nummern des Standards ISO11783 [ISO15] auf.

Zum Anderen existieren verschiedene Codesysteme, die für äquivalente oder zumindest nahverwandte Konzepte unterschiedliche Bezeichner beinhalten. Dies ist häufig darin begründet, dass diese im Kontext unterschiedlicher Anwendungsfälle oder innerhalb nur schwach vernetzter, verschiedener fachlicher Interessensgemeinschaften parallel entwickelt wurden. Dies ist beispielsweise der Fall im Bereich von Codesystemen landwirtschaftlich angebauter Kulturen. Auf diese wird im Abschnitt 2.1 detaillierter eingegangen.

Sowohl die mangelnde Eindeutigkeit als auch die Heterogenität an zur Verfügung stehenden Codesystemen führen zu Herausforderungen bei der Zusammenführung von Daten für übergreifende Auswertungen oder für die Interoperabilität im Datenaustausch beispielsweise zwischen Farmmanagement-Informationssystemen und Entscheidungsunterstützungssystemen. So müssen häufig vorab Datentransformationen durchgeführt und individuell Zuordnungstabellen für Bezeichner und Codesysteme erstellt werden.

Verschiedene Spezifikationen des World Wide Web Consortium (W3C) adressieren genau diese Schwachpunkte. Außerdem wird in den FAIR-Prinzipien [Wi16] zur Sicherstellung der Auffindbarkeit und Interoperabilität von Datenbeständen gefordert, dass (Meta)daten global eindeutige Bezeichner erhalten sollen (Prinzip F1), über die diese auch abrufbar sind (Prinzip A1), dass (Meta)daten mit Hilfe formaler Sprachen für die Wissensrepräsentation beschrieben werden sollen (Prinzip I1) und dass sie qualifizierte Verweise auf weitere (Meta)daten beinhalten sollen (Prinzip I3). Was dies bedeutet und wie eine praktische Umsetzung erfolgen kann, wird hier anhand eines im Rahmen verschiedener Projekte entwickelten Systems zur Handhabung von Daten, die verschiedene Codesysteme für landwirtschaftliche Kulturen beinhalten, dargestellt.

2 Material und Methoden

2.1 Einbezogene Codesysteme und Datenbestände

Die folgenden Codesysteme für landwirtschaftliche Kulturen wurden eingebunden:

- **EPPO-Kulturcodes** wurden speziell für Anwendungsfälle des Pflanzenschutzes entwickelt. Die bereitgestellten Bezeichner werden beispielsweise im Rahmen der Zulassungs- und Registrierungsverfahren und in der Pflanzenschutzberatung genutzt. Gepflegt wird das System heute von der European Plant Protection Organisation (EPPO). Kulturpflanzen werden über fünfstellige Buchstabencodes bezeichnet, z. B. SOLTU für die Kartoffel.
- **InVeKoS-Kulturcodes** zur Bezeichnung von Kulturen werden im Verfahren zur Beantragung der Agrarförderung im Rahmen des Integrierten Verwaltungs- und Kontroll-Systems (InVeKoS) genutzt. Dabei werden numerische Codes verwendet: z. B. 115 für Winterweichweizen. Außerdem werden übergeordnete Codes für die Anbaudiversifizierung spezifiziert, der genannte Winterweichweizen gehört z. B. zur Systematik der Winterweizen, Code 1.28.2.1. Die Codezuweisungen sind zwar in allen Bundesländern identisch, es werden jedoch jeweils unterschiedliche Untermengen verwendet.
- **Kulturcodes des Beratungssystems PSInfo**: PSInfo ist ein vom Dienstleistungszentrum Ländlicher Raum Rheinpfalz bereitgestelltes System der Agrarberatung[2]. Die dort verwendeten Codes entsprechen weitgehend den EPPO-Codes, es wurden jedoch Erweiterungen hinzugefügt, die nach der Nutzung von Kulturen z. B. als Blatt- oder Wurzelgemüse differenzieren.
- **Kulturbezeichner des Bundessortenamtes** werden in der Sortenzulassung auf nationaler Ebene in Deutschland und in Veröffentlichungen zu Sortendaten genutzt. Diese Kulturpflanzen-Bezeichner bestehen aus einer Folge von einem bis drei Buchstaben, z.B. HWS für Sommerhartweizen.

Bereits ein Vergleich einiger der genannten Beispiele in den verschiedenen Codesystemen fördert eine Herausforderung zutage: So unterscheidet sich das Verständnis des Begriffes „Kultur“: Während die Codes der EPPO weitgehend auf den Artbegriff abheben, unterscheiden die Bundessortenamt-Codes Sommer- und Winterformen. Sowohl PSInfo als auch das InVeKoS-Verfahren unterscheiden zusätzlich nach Nutzung (z. B. Speisekartoffeln (602), Stärkekartoffeln (601) und Pflanzkartoffeln (606)), letzteres System beinhaltet außerdem Codes z. B. für Brachen, Aufforstungen und Hecken.

Als umfassendere semantische Ressource wurde außerdem der AGROVOC-Thesaurus der FAO [FAO21] eingebunden.

2.2 Vorverarbeitung und Datenrepräsentation

Aufgrund der beschriebenen Aspekte ist für das Semantic Web spezifiziert, dass alle Konzepte – Ressourcen genannt – einen kontextfrei global eindeutigen Bezeichner zugewiesen bekommen sollen. Hierfür werden üblicherweise Uniform Resource Identifier (URIs, [BFM05]) genutzt. Die in den Codesystemen enthaltenen Bezeichner wurden daher durch

[2] https://www.pflanzenschutz-information.de/, Stand 31.10.2021

Hinzufügen eines Präfixes in URIs überführt. Aus dem InVeKoS-Code 115 wird z. B. `<https://srv.ktbl.de/daten/cc/invekos/115>`. Syntaktisch lassen sich URIs über die Deklaration von Namespace-Präfixen abkürzen, d. h. die Kurzform der o. g. URI lautet `cci:115`. Die dadurch entstehenden Semantic-Web-Ressourcen lassen sich mit Hilfe des Resource Description Framework (RDF, [CWL14]) durch logische Aussagen näher beschreiben. Dabei werden „Sätze“ (Tripel) bestehend aus Subjekt, Prädikat und Objekt gebildet. Die Aussage `cci:115 rdf:type skos:Concept` drückt z. B. aus, dass der Code cci:115 als SKOS-Konzept typisiert wurde, dabei ist `cci:115` das Subjekt, `rdf:type` das Prädikat und `skos:Concept` das Objekt der Tripelaussage. SKOS steht für Simple Knowledge Organisation System [MB09] und ist eine Empfehlung des W3C zur Modellierung von Begriffssystemen. Über solche Aussagen lassen sich mit Hilfe entsprechender Prädikate auch begriffliche Verwandtschaften ausdrücken. Dabei lässt sich nicht nur die einfache Gleichheit zweier Konzepte, wie sie oft in Zuordnungstabellen modelliert wird, abbilden, sondern es lassen sich auch fein-granulare Äquivalenzbeziehungen und Ober-/Unterbegriffsrelationen oder losere Beziehungen darstellen. Hierfür stellt SKOS die Prädikate `skos:exactMatch`, `skos:broadMatch`, `skos:narrowMatch` und `skos:relatedMatch` zur Verfügung, die in dem Zusammenhang genutzt wurden. Auch den einbezogenen Codesystemen als Ganzes wurde eine URI zugewiesen, sodass sich diese als eigene Ressourcen mit weiteren Metadaten beschreiben lassen. In den Codesystemen enthaltene Codes wurden mittels des Prädikats `skos:inScheme` diesen Ressourcen zugewiesen. Außerdem wurden die Codesysteme mittels des DublinCore- [DCM20] sowie des PROV-Vokabulars [LSM13] mit ihren Quellen- und Provenienzangaben versehen.

Die Rohdaten wurden dabei aus Webseiten und teils auch PDF-Dokumenten bezogen und mit Hilfe von Python-Skripten als RDF aufbereitet. Bislang erfolgt die Aufbereitung dabei individuell für jede Datenquelle. Der hierfür im Internet als Open-Source-Software frei verfügbare Werkzeugkasten beinhaltet jedoch eine Reihe komfortabel nutzbarer Bibliotheken und Mappingwerkzeuge wie beispielsweise rdflib[3] oder pyTARQL[4], sodass sich damit verbundene Aufwände in Grenzen halten. Bereits jetzt als RDF bereitgestellte Datensätze lassen sich direkt ohne jegliche Vorverarbeitung nutzen. So waren für die Einbindung des AGROVOC beispielsweise keine vorbereitenden Schritte erforderlich. Dieser erfüllt die o. g. Anforderungen und FAIR-Prinzipien als RDF-Datensatz bereits vollständig.

2.3 Bereitstellung

Zur gezielten Abfrage werden die Daten über einen SPARQL-Service-Endpoint bereitgestellt. Zum Einsatz kommt dabei Apache Jena Fuseki. Abfragen können mit Hilfe der

[3] https://github.com/RDFLib/rdflib, Stand 6.12.2021
[4] https://github.com/RDFLib/pyTARQL. Stand 6.12.2021

SPARQL Query Language [HS13] formuliert werden und erlauben das Ausgeben praktisch beliebiger Zusammenstellungen enthaltener Aussagen mittels SELECT bzw. CONSTRUCT-Abfragen. Im Ergebnis stand somit nun ein semantisches Netz von Beziehungen zur Verfügung, das gezielt Abfragen und Auswertungen für eine Darstellung von Zusammenhängen verschiedener Codesysteme erlaubt.

3 Ergebnisse, Diskussion und Fazit

Mit Hilfe des Datenbestandes lassen sich über entsprechende Abfragen an die SPARQL-Schnittstelle nun beispielsweise folgende Fragen beantworten:

- Welche Bezeichner für Kulturen sind im InVeKoS-Verfahren des Bundeslandes Niedersachsen valide?
- Aus welchem Quelldokument wurden diese gewonnen, wo lassen sich diese nachlesen?
- Welche Bezeichner stellen Codesysteme für Winterweizen zur Verfügung?
- Welche genauen Entsprechungen gibt es für den Bezeichner 115 im InVeKoS-Verfahren in anderen Codesystemen?
- Welche weniger genauen Entsprechungen stehen zur Verfügung?

Es steht mithin ein System zur Verfügung, das für Datentransformationen notwendige Informationen bereitstellen kann – mehr noch: Wenn die Datenbestände selbst nach Spezifikationen des Semantic Web aufgebaut sind und URIs für ihre Codesysteme nutzen, können diese ohne Vorverarbeitung zusammengeführt werden und die in dem hier skizzierten Ansatz bereits vorhandenen Beziehungen für Querverknüpfungen nutzen.

Oft wird die Heterogenität vorhandener Codesysteme zum Anlass genommen, eine stärkere Standardisierung einzufordern – es soll ein einziges Codesystem geschaffen werden, das alle Anforderungen abbildet. Anzunehmen, das Problem ließe sich hierdurch lösen, ist indes unrealistisch: Abstimmungsprozesse hierfür sind aufwändig und angesichts der großen Anzahl Beteiligter nur schwer zu organisieren. Außerdem existieren Unterschiede in Granularität und hierarchischen Kategorisierungen, die in der Regel durch die jeweiligen Anwendungsfälle gut begründet sind und sich nur schwer auflösen lassen.

Wenn hingegen die Bereitsteller von Codesystemen und Standards einige einfache Grundprinzipien des Semantic Web wie die Nutzung von URIs anstatt nur einfacher alphanumerischer Bezeichner beachten würden, könnten Dritte gemäß der hier dargestellten Mechanismen Relationen zwischen Codesystemen beschreiben und durch Maschinen auswertbar und abfragbar machen.

4 Ausblick

Neben den genannten Codesystemen existieren im internationalen Umfeld weitere, die ebenfalls die Bezeichnung landwirtschaftlicher Kulturen beinhalten, z. B. die USDA Commodity Codes oder die Codes des europäischen Sortenamtes, der CPVO. Außerdem pflegt die EPPO Codesysteme auch für Schaderreger. Der Ausbau des semantischen Netzes um diese Bezeichner ist geplant.

Teilweise erfordert die Einpflege von Relationen noch manuelle Nacharbeit. Aktuell werden jedoch auch Mechanismen der automatischen Ableitung (Alignment) erprobt. Je mehr dabei auf bereits vorhandene Begriffssysteme und Relationen zurückgegriffen werden kann, desto eher lassen sich gute Ergebnisse erzielen.

Literaturverzeichnis

[BFM05] Berners-Lee, T.; Fielding, R.; Masinter, L.: RFC3986: Uniform Resource Identifier (URI): Generic Syntax. Internet Engineering Task Force (IETF), Network Working Group, 2005. https://www.rfc-editor.org/rfc/rfc3986.txt, Stand 31.10.2021.

[CWL14] Cyganiak, R.; Wood, D.; Lanthaler, M.: RDF 1.1 Concepts and Abstract Syntax. World Wide Web Consortium (W3C), 2014. https://www.w3.org/TR/rdf11-concepts/, Stand 31.10.2021.

[DCM20] DCMI Usage Board: DCMI Metadata Terms. 2020. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, Stand 31.10.2021.

[FAO21] Food and Agriculture Organization of the United Nations (FAO): AGROVOC – Semantic data interoperability on food and agriculture. Rome, 2021. https://doi.org/10.4060/cb2838en, Stand 31.10.2021.

[HS13] Harris, S.; Seaborne, A.: SPARQL 1.1 Query Language. World Wide Web Consortium (W3C), 2013. http://www.w3.org/TR/sparql11-query/, Stand 31.10.2021.

[ISO15] International Organization for Standardization: ISO 11783-10:2015 – Tractors and machinery for agriculture and forestry — Serial control and communications data network — Part 10: Task controller and management information system data interchange. 2015.

[LSM13] Lebo, T.; Sahoo, S.; McGuiness, D.: PROV-O: The PROV Ontology. World Wide Web Consortium (W3C), 2013. https://www.w3.org/TR/prov-o/, Stand 31.10.2021.

[MB09] Miles, A.; Bechhofer, S.: SKOS Simple Knowledge Organization System Reference. World Wide Web Consortium (W3C), 2009. https://www.w3.org/TR/skos-reference/, Stand 31.10.2021.

[Wi16] Wilkinson, M. D. et.al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18, Stand 31.10.2021.