Daniel Martini, Daniel Herzig, Günter Ladwig und Martin Kunisch

Semantische Suche: Planungsdaten des KTBL finden und maschinell weiterverarbeiten

Der Aufwand für die Recherche nach relevanten Daten für die Planung und Vorbereitung von Arbeiten und Investitionen in der landwirtschaftlichen Produktion sowie die anschließende Aufbereitung und Einpflege der Daten für die Verwendung in Kalkulationswerkzeugen und Farm-Managementsystemen stellen wesentliche Herausforderungen für datenbasierte Entscheidungen dar. In diesem Beitrag wird eine Lösung vorgestellt, die das gezielte Auffinden von Planungsdaten im KTBL-Datenangebot durch den Einsatz einer semantischen Suchmaschine vereinfacht und eine leichte Wiederverwendung und Weiterverarbeitung ermöglicht, da die KTBL-Daten nach den Prinzipien von Linked Open Data bereitgestellt werden.

eingereicht 12. Dezember 2013
akzeptiert 30. Januar 2014

Schlüsselwörter
Semantic Web, Linked Open Data, Webdienste, Suchmaschine

Keywords
Semantic web, linked open data, web services, search engine

Abstract
Martini, Daniel; Herzig, Daniel; Ladwig, Günter and Kunisch, Martin

Semantic search: finding KTBL's planning data and reusing them in IT systems

Landtechnik 69(1), 2014, pp. 12–18, 3 figures, 17 references

The effort to investigate relevant data for planning purposes and preparation of labour and investments in agricultural production as well as reworking and entering them for reuse in calculation tools and farm management information systems are major challenges for decisions based on data. The following paper presents a solution which on the one hand simplifies targeted finding of planning data within KTBL's data sets using a semantic search engine and on the other hand enables simple reuse and processing of these data by providing them using Linked Open Data principles.

Information spielt in der Landwirtschaft eine zunehmende Rolle als Basis für Entscheidungen im betrieblichen Management. Grundlegende Daten zur Vorbereitung von Investitionen und zur Planung von Produktionsverfahren und Arbeitsprozessen auf dem landwirtschaftlichen Betrieb werden vom Kuratorium für Technik und Bauwesen in der Landwirtschaft e. V. (KTBL) bereits seit langer Zeit gesammelt, aufbereitet und qualitätsgesichert bereitgestellt. Wenn beispielsweise ein Landwirt überlegt, in den Anbau von Qualitätsweizen einzusteigen und eine Wirtschaftlichkeitsbetrachtung durchführen möchte, kann er dafür notwendige Daten, wie fixe und variable Kosten für gegebenenfalls notwendige neue Maschinen oder Lohnunternehmeraufträge oder den Arbeitszeitbedarf, erhalten. Wichtigstes Medium hierfür waren Publikationen in gedruckter Form wie beispielsweise die Faustzahlen Landwirtschaft [1] oder die Datensammlung Betriebsplanung [2]. Auch in anderen Bereichen, aus denen für die Vorbereitung und Planung landwirtschaftlicher Produktion wichtige Daten bereitgestellt werden, z. B. Sortenversuchsergebnisse und Sortenlisten, war die Papierform lange dominierend. Bereits seit einigen Jahren spielt jedoch das Internet auch im Agrarbereich eine zunehmende Rolle bei der Veröffentlichung und Verbreitung von Daten. Bislang sind online angebotene Daten in aller Regel eng integriert mit ihren Anwendungen und dem Nutzer lediglich über vorgefertigte, von den jeweiligen Entwicklern in den Be-

dienoberflächen festgelegten Interaktionswegen zugänglich. Auch bei den zunehmend beliebten Apps für Smartphones fehlt meist eine konsequente Trennung von Datendiensten und Anwendungslogik. Eine Weiternutzung und Verarbeitung der Daten in anderen Systemen, wie z. B. in Farm-Management-Informationssystemen oder Beratungswerkzeugen, oder ein Zusammenführen von Datenbeständen verschiedener Organisationen ist mithin nur erschwert möglich und erfordert aufwendigste Arbeiten für das Einpflegen oder den Import. Die Nutzer der Daten sind gezwungen, die jeweils vorgesehenen Anwendungen zu verwenden. Besonders bei Smartphones sammeln sich dadurch eine Vielzahl von Anwendungen auf Systemen an, die oft ähnliche Funktionalitäten abdecken oder sich inhaltlich überlappen, aber dennoch inkompatibel sind.

Grundlegende technische Methoden

Technologien im Umfeld des Semantic Web erlauben inzwischen für die Datennutzer elegantere und flexiblere Lösungen gegenüber den oben beschriebenen Ansätzen, bei denen Anwendungen und Daten eng verdrahtet sind, oder auch gegenüber traditionellen Webservices auf Basis von SOAP (Simple Object Access Protocol [3]) und XML (eXtensible Markup Language [4]). So wird im Rahmen der Linked-Open-Data (LOD)-Initiative des World Wide Web Consortium (W3C) versucht, Daten von Auswertungslogik und Benutzerschnittstelle zu trennen und mit einfachen Internet-Protokollen (in erster Linie HTTP) zugänglich zu machen [5; 6]. Insbesondere werden die Daten hier auch maschinenlesbar in standardisierten Formaten bereitgestellt, sodass diese über einfache URL-Aufrufe abgerufen und automatisiert in Anwendungen eingelesen werden können.

Für die Repräsentation der Daten in LOD-Diensten wird das Resource Description Framework (RDF) vom W3C empfohlen und im Allgemeinen auch von den meisten Anbietern genutzt. Zu den Daten gehören jeweils Vokabulare, die z. B. in RDF-Schema [7] erstellt werden können. Kernaspekt ist dabei die Modellierung von Daten als gerichteter Graph. Ein solcher Graph besteht aus Knoten und Kanten. Knoten können über die jeweiligen Kanten mit einer beliebigen Anzahl weiterer Knoten verknüpft sein. Genutzt werden solche Strukturen zur Lösung von Problemen in einer Reihe von Anwendungsbereichen wie z. B. Routing für Navigationssysteme oder in Kommunikationsnetzwerken (Mobilfunknetz, Internet). Im Kontext des Semantic Web wird mit Ihrer Hilfe ein Netz von Aussagen aufgebaut. Dabei bilden jeweils die beiden an eine Kante angrenzenden Knoten zusammen mit ihrer zugehörigen Kante eine Aussage der Form Subjekt-Prädikat-Objekt oder Ding-Eigenschaft-Wert, ein sogenanntes Tripel. Anhand eines sehr einfachen Beispiels lassen sich die wichtigsten Charakteristika dieser Darstellungsform erläutern:

	Subjekt	Prädikat	Objekt
Tripel 1:	LandwirtXY	Besitzt	Maschine0815
Tripel 2:	Maschine0815	istEin	Traktor
Tripel 3:	Maschine0815	Anschaffungspreis	83000 Euro

Tripel 1 verknüpft einen Landwirt mit einer Maschine. Die beiden Knoten des Graphen sind dabei LandwirtXY sowie Maschine0815. Die Kante wäre in dem Fall die Beziehung „besitzen“. Das in dieser Aussage verwendete Objekt wird in der nächsten Aussage (Tripel 2) als Subjekt verwendet. Hierdurch wird vom Knoten Maschine0815 ausgehend eine neue Kante eröffnet, die über eine „istEin“-Beziehung diesen Knoten näher spezifiziert. Das in diesem Tripel vorhandene Objekt Traktor könnte ebenfalls in weiteren Tripeln näher beschrieben werden, so könnte z. B. über eine Unterklassenrelation beschrieben sein, dass der Traktor eine Art von Landmaschine ist. Tripel 3 illustriert die Verwendung von sogenannten Literalen, die im Allgemeinen dazu verwendet werden, im Graphen Objekten Eigenschaften mit Werten (z. B. numerisch oder Zeichenketten) zuzuweisen. Von Literalen können keine weiteren Kanten ausgehen, sie können aber mit einem Datentyp (im Beispiel „Euro“) verknüpft sein. Tripel wie die oben dargestellten werden in einer Syntax wie z. B. Turtle [8] repräsentiert. Außerdem werden alle Knoten und Kanten mit Ausnahme der Literale als Uniform Resource Identifier (URI [9]) dargestellt. Hierdurch können Datensätze mit weiteren Daten auf anderen Servern verknüpft werden, sodass es überhaupt erst möglich wird, einen „weltweiten Datenraum“ zu schaffen [5]. Im Abschnitt „Linked Open Data Dienst“ findet sich ein Beispiel hierfür, das Turtle-Syntax und Nutzung von URIs illustriert. Eine wichtige Eigenschaft der Graphenrepräsentation ist, dass in diese allgemeine und generische Datenstruktur Datensätze jeglicher Art einfach überführt werden können. Einträge in Datenbanktabellen (Zeilen) können beispielsweise als Objekte repräsentiert werden, die Spalten der Tabelle können als Eigenschaften dargestellt werden. Mit geeigneten Werkzeugen kann eine solche Überführung „virtuell“ durchgeführt werden, d. h. sie erfolgt während der Laufzeit und erfordert keinen Austausch von vorhandenen Infrastrukturen zur Datenspeicherung. Beziehungen zwischen Daten werden außerdem explizit modelliert und spezifiziert. Das heißt, Zusammenhängen, die in relationalen Datenbanken lediglich als Schlüssel-Fremdschlüssel-Relationen abgebildet würden, wird in dieser Art der Modellierung ein Bezeichner gegeben, der auch eine gezielte Abfrage nach dieser Beziehung erlaubt (in den oben gegebenen Tripeln z. B.: Welche Knoten sind durch „besitzen“ verknüpft?). Außerdem ist eine Vorabfestlegung solcher Beziehungen zwischen Entitäten oder vorgefertigten Baumstrukturen nicht notwendig, d. h. es können z. B. später flexibel weitere Arten von Objekten hinzugefügt werden, die sowohl bereits vorhandene Beziehungen nutzen als auch neue Eigenschaften mitbringen können. Da Daten und Schemas oder Vokabularien in dem gleichen Format repräsentiert werden, können Datensätze mit einfachen Mitteln so beschrieben werden, dass darauf aufbauende Anwendungen neue Hinzufügungen von Daten ohne jegliche Neukompilierung oder manuelle Anpassung direkt nutzen können. In relationalen Datenbanken wäre ein solcher Ansatz durch Abfragen in das sogenannte „information schema“ und darauf aufbauend durch den dynamischen Aufbau der Bedienoberfläche möglich, was

aber im Vergleich relativ aufwendig zu realisieren, schwer zu handhaben und zudem herstellerabhängig ist.

Zusammenfassend lässt sich feststellen, dass Daten durch die beschriebenen Repräsentationstechnologien flexibel und sehr nah an in der Realwelt existierenden Bezügen orientiert dargestellt werden können. Auch komplexe Abfragen können meist einfach formuliert werden, indem im Graphen gezielt Mengen von Knoten mit bestimmten Eigenschaften selektiert werden. Im RDF-Umfeld wird hierfür in aller Regel die Abfragesprache SPARQL [10] genutzt. Durch die genannten Eigenschaften wird eine anwendungsunabhängige Nutzung vereinfacht. Datensätze werden außerdem besonders zugänglich für Suche, Navigation und die Beantwortung komplexer Fragestellungen, die die Einbeziehung einer Reihe von Bezügen erfordern.

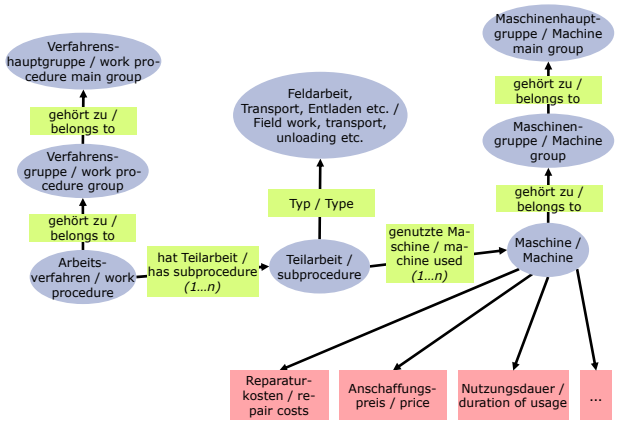
Auf technischer Ebene wird durch die Nutzung des HTTP-Protokolls, den oben genannten generischen Datenformaten und einfachen RDF-basierten Vokabularen zur Beschreibung der Daten der Zugriff soweit vereinheitlicht, dass für Clientsysteme, die Daten abrufen wollen, im Vergleich zu anderen Technologien (s. o.) nur ein Mindestmaß an Vorabfestlegung durch die Entwickler notwendig ist. Das erlaubt Verknüpfung und Erweiterung von Datenbeständen.

Im Rahmen des iGreen-Projektes und einer langjährigen Kooperation mit der FAO konnten die Autoren grundlegende Entwicklungen in diesem Bereich vorantreiben. So wurde unter anderem das agroRDF-Vokabular auf Basis von agroXML erstellt [11]. Aufbauend auf diesen Vorarbeiten hat das KTBL nun auf Basis der vorhandenen KTBL-Datenbank einen Linked-Open-Data-Service mit Informationen zu Arbeitsverfahren und Maschinenkosten aufgesetzt, der den maschinellen Abruf dieser Daten erlaubt. Auf diesen Dienst wurde von SearchHaus, einem Spin-Off des Karlsruher Instituts für Technologie (KIT), als eine mögliche Anwendung eine semantische Suche aufgesetzt, die ein gezieltes Auffinden von Datensätzen ermöglicht. Wird beispielsweise die Frage untersucht, wie hoch die Investitionskosten für Maschinen sind, die zur Aussaat von Weizen verwendet werden, so kann mit einfachen Stichwortanfragen, z. B. „anschaffungspreis maschinen weizen saat“, der KTBL-Datenbestand durchsucht werden. Die Anfragen können frei formuliert werden und es sind keine speziellen Kenntnisse über Aufbau oder Abfrage der Datenbank notwendig. Für mögliche Treffer werden verschiedene Interpretationen angezeigt, die anhand der in den Daten vorhandenen Beziehungen ermittelt werden. Die Ergebnisse können anschließend über dynamisch berechnete Facetten eingeschränkt werden.

Linked Open Data Dienst

Der Linked-Open-Data-Service wurde mithilfe des Open-Source-Werkzeugs d2rq [12] auf die bestehende Oracle-11-Datenbank aufgesetzt. Hierfür wurde mit der d2rq mapping language ein Mapping erstellt, das spezifiziert, wie die vorhandenen relationalen Datenbanktabellen in ein RDF-Graphen-Modell zu überführen sind. Zur Beschreibung der Dateninhalte wurde

Abb. 1



Stark vereinfachter, schematischer Überblick über die im derzeitigen Prototypen vorhandenen Daten und deren Beziehungen ohne Verknüpfungen zu externen Diensten

Fig. 1: Heavily simplified, schematic overview on data and their relations represented within the current prototype without links to external services

ein RDF-Schema erstellt, das auf agroRDF basiert. Spezifische Konzepte der KTBL-Datenbank wurden als entsprechende Erweiterungen umgesetzt. Als Ergebnis stehen zum einen ein http-Service, der Daten sowohl im HTML-Format zur Präsentation für den Nutzer im Browser als auch im maschinenlesbaren Turtle-Format [8] für RDF ausliefern kann, zum anderen ein SPARQL-Endpoint für formal spezifizierte Abfragen zur Verfügung.

Fachlich beschränkt sich der Dienst derzeit auf Daten zu Maschinen und Feldarbeitsverfahren. **Abbildung 1** zeigt die wichtigsten Entitäten und deren Beziehungen zueinander. Die stark vereinfachte Darstellung, die sich in nur 10 Tripeln repräsentieren lassen würde, stellt keine Verknüpfungen zu externen Diensten dar und soll lediglich einen Überblick geben. Das tatsächlich verwendete RDF-Schema zur Beschreibung des Datensatzes umfasst 190 Tripel. Die Daten selbst umfassen derzeit 104342 Tripel. Hinzu kommen außerdem 180 Tripel für die Beschreibung der KTBL-Anwendungen, die in der Suche ebenfalls gefunden werden können, sowie die 283 Tripel des agroRDF-Maschinenvokabulars. Reale Datensätze sind also meist deutlich umfangreicher als das im vorhergehenden Abschnitt zur Illustration gewählte Beispiel mit nur drei Tripeln. Neben den in der lokalen Datenbank vorhandenen Daten wurden in den Datensatz auch Verknüpfungen zu dem multilingualen AGROVOC-Thesaurus der FAO [13] eingebaut, sodass dort für einige Maschinenarten die entsprechenden Konzepte sowie die Übersetzungen der Bezeichnung in eine Vielzahl von Sprachen abgerufen werden können. Hierfür müssen lediglich – wie oben beschrieben – die entsprechenden URI des Linked-Open-Data-Servers der FAO in die Subjekt-, Prädikat oder Objekt-Position eines Tripels eingesetzt werden. Ein solcher Prozess der Verknüpfung erfolgt meist semi-automatisch, auch im vorliegenden Fall wurden verschiedene einfache Werkzeuge genutzt, die

hier nicht im Detail dargestellt werden, und anschließend an einigen Stellen manuell nachgepflegt. Da die Reihenfolge der Aussagen im Datensatz unerheblich ist und Tripel wegen der oben beschriebenen Charakteristika des Graphenmodells in RDF problemlos später hinzugefügt werden können, lassen sich für diesen Prozess aber für Entwickler sehr einfach zu handhabende – ggfs. auch iterativ durchführbare – Abläufe realisieren. Clientanwendungen müssen so anschließend lediglich den vorhandenen Verknüpfungen folgen um weitere, externe Dienste einzubinden. Wegen der Vereinheitlichung der Protokollebene (http) und der Nutzung standardisierter Syntax und der Selbstbeschreibung der Datensätze über Vokabularien ist ein explorativer Ansatz mithilfe einer einzigen, generischen Anwendung grundsätzlich möglich.

Ein Einstieg in den Linked-Open-Data-Dienst kann über einen beliebigen, im Datensatz enthaltenen URI erfolgen. Genau wie externe Dienste in den eigenen Datensatz eingebunden werden können, kann so auch umgekehrt eine Einbindung des eigenen Dienstes in beliebigen, anderen Datensätzen erfolgen. Im gegebenen Prototypen sieht das maschinenlesbare Ergebnis eines http-Aufrufs des URI http://www.agroxml.de/data/data/machine/110406[1)] beispielsweise verkürzt so aus:

```
@prefix ktbldb: <http://www.agroxml.de/lod/
vocabulary/db#> .
@prefix agrordf-mach: <http://www.agroxml.de/lod/
vocabulary/machine#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-
schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://www.agroxml.de/data/resource/
machine/110406>
      a ktbldb:StandardMachine ;
      rdfs:label „Mähdrescher, Schüttler, bis
20 km/h - 200 kW; 8500 l“ ;
      ktbldb:abschreibung „23000“^^xsd:decimal ;
      ktbldb:anschaffungspreis
               „230000“^^xsd:decimal ;
      ktbldb:belongsTo <http://www.agroxml.de/data/
resource/machinegroup/219> ;
      ktbldb:versicherung „60“^^xsd:decimal ;
      agrordf-mach:height „3900“^^xsd:decimal ;
      […weitere Daten folgen…]
```

Es handelt sich hierbei um eine Ausgabe der Daten zu einer über diesen URI identifizierten Maschine – eine KTBL-Standardmaschine, in diesem Fall eine bestimmte Art von Mähdrescher – in der oben bereits erwähnten Turtle-Syntax. Weitere Formate, beispielsweise auf Basis von XML oder der Java Script Object Notation [14], wären grundsätzlich möglich, sind derzeit aber noch nicht implementiert. Turtle war generell einfacher einzulesen als XML und gleichzeitig mächtiger als JSON, was die Darstellung von Verknüpfungen anbelangt, es vereint also in gewisser Weise die Vorteile beider Welten. Der Dienst kann jedoch auch eine einfache HTML-Darstellung, die für die Anzeige in Standard-Webbrowsern geeignet ist, ausliefern. Ansprechendere Oberflächen lassen sich aber realisieren, indem die Daten in der obigen Form abgeholt und anschließend aufbereitet werden. Die im Dienst enthaltenen Daten zu Maschinenkosten – auch im Kontext verschiedener Arbeitsverfahren – können so auch in Farm-Management-Informationssystemen (FMIS, z. B. Schlagkarteien) für die Planung von Arbeitsprozessen eingesetzt werden. Die Darstellung erfolgt dann nahtlos in der jeweiligen nativen grafischen Oberfläche des FMIS.

Abgesehen von der Möglichkeit, den Linked-Open-Data-Service über eine aus einem anderen, externen Dienst oder aus früheren Aufrufen bekannte URI zu „betreten“, kann außerdem der allgemeine Einstiegspunkt über das Wurzelverzeichnis http://www.agroxml.de/data/ genutzt werden, von dem aus Verknüpfungen über mehrere Stufen zu allen enthaltenen Datensätzen führen. Für gezielte Abfragen steht ein sogenannter SPARQL-Endpunkt unter http://www.agroxml.de/data/sparql zur Verfügung. Anfragen können hier über die SQL-ähnliche Sprache SPARQL [10] an den dahinter liegenden Graphen abgesetzt werden.

Die in dieser Weise semantisch aufbereiteten Daten bilden die Grundlage für die Suche.

Linked Open Data Suchmaschine

Mit der semantischen Suche von SearchHaus können die Benutzer einfach in den KTBL-Daten mit Schlüsselwörtern suchen, so wie sie es von Websuchmaschinen gewohnt sind. Die Suchmaschine nimmt Schlüsselwortanfragen entgegen, interpretiert die Anfragen mithilfe des Datenbestands und liefert dem Benutzer die passenden Ergebnisse aus den KTBL-Daten. Intern aggregiert und gruppiert die Suchmaschine die komplexen, strukturierten Daten, sodass eine schnelle Beantwortung der Anfragen auch über große Datenbestände möglich ist. Auf dieser kompakten Datenrepräsentation werden dann mittels Explorationsalgorithmen und Rankingverfahren die Suchergebnisse ermittelt [15].

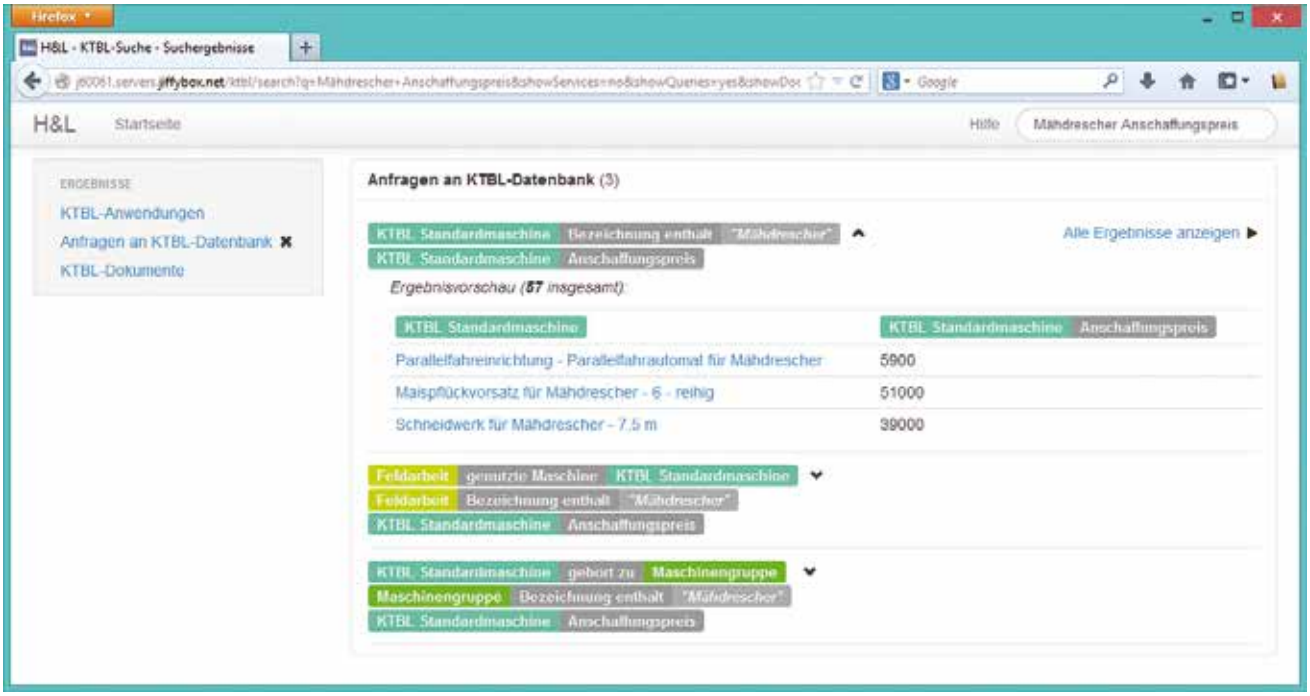
Abbildung 2 zeigt beispielhaft das Suchergebnis für die Schlüsselwörter „Mähdrescher Anschaffungspreis“. In dem Fall wurden drei mögliche Interpretationen gefunden:

- KTBL-Standardmaschinen, deren Bezeichnung den Begriff Mähdrescher enthält, und die ein Attribut Anschaffungspreis haben
- Feldarbeiten, für die Mähdrescher als Maschine zur Ausführung genutzt werden, die wiederum einen Anschaffungspreis haben
- KTBL-Standardmaschinen, die zur Maschinengruppe der Mähdrescher gehören und einen Anschaffungspreis haben (nicht deckungsgleich mit der ersten Interpretation!)

Standardmäßig werden nur für die erste Interpretation einige wenige Ergebnisse angezeigt. Für die anderen Interpretationen kann über die Schaltfläche mit dem nach unten weisenden Pfeil die Anzeige entsprechend erweitert werden. Nicht dargestellt sind in der Abbildung außerdem Treffer in weiteren Datenquellen wie KTBL-Anwendungen und Druckwerke. Die-

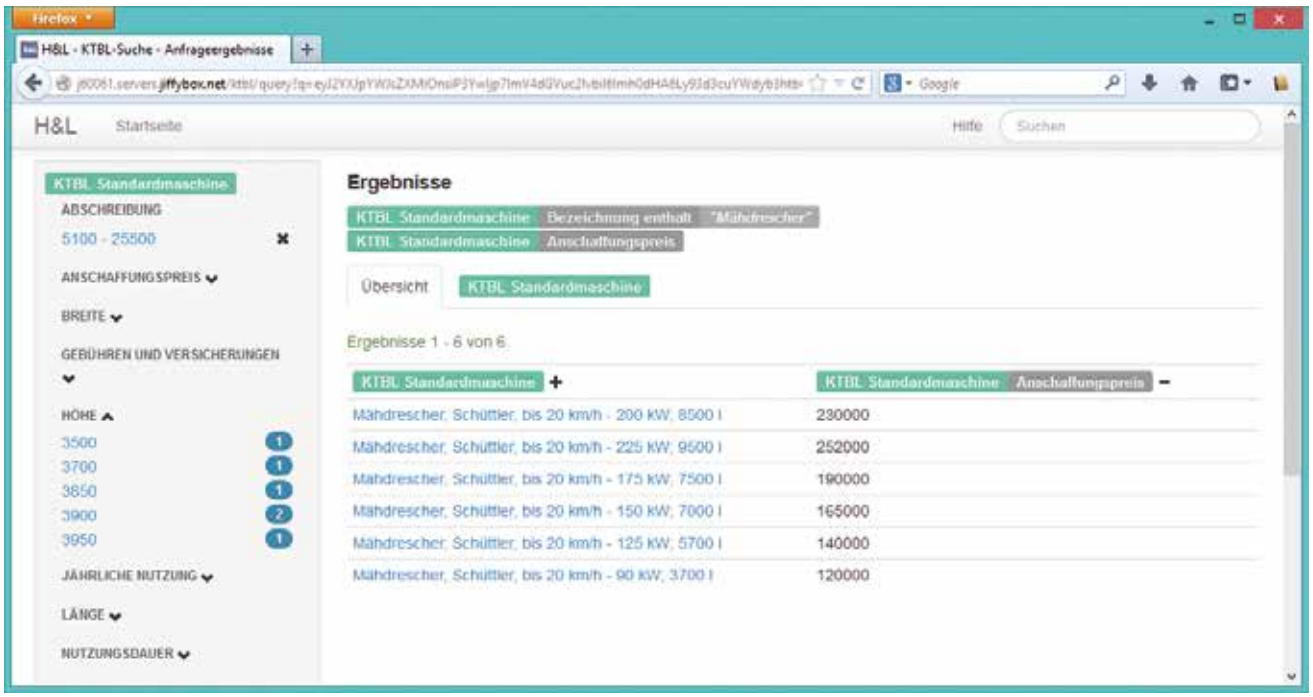
[1)] Da es sich um einen Prototypen handelt, an dem noch gearbeitet wird, sind alle im folgenden aufgeführten URIs vorläufig und liegen aktuell noch in einem passwortgeschützten Bereich. Bei Interesse an einem Zugang zum derzeitigen Stand können die Autoren kontaktiert werden.

Abb. 2



Einstieg in die Exploration der Daten, im Beispiel mit der Begriffskombination „Mähdrescher Anschaffungspreis“. Erläuterungen im Text
Fig. 2: Entry point for data exploration, in this example using the term combination „Mähdrescher Anschaffungspreis“ (combine harvester purchase price). Further explanation within the text

Abb. 3



Verfeinerung der Suchergebnisse mittels automatisch anhand der im Datensatz vorhandenen Eigenschaften und deren Wertbereiche berechneter Facetten
Fig. 3: Refining Search Results using automatically calculated facets according to properties and their value spaces as given in the data set

se können über die Schaltfläche im linken Bildschirmbereich eingeblendet werden. Ersichtlich wird an dieser Stelle, dass bei der semantischen Suche nicht nur eine Volltextsuche in den Textfeldern der Datenbank durchgeführt wird, sondern auch die Beziehungen und Zusammenhänge zwischen Entitäten und deren Eigenschaften („Anschaffungspreis“) ausgewertet werden. Dabei können auch Ober-/Unterbegriffsbeziehungen, z. B. „Anschaffungspreis“ als eine Art von fixen Kosten, genutzt werden, sofern diese im zu den Daten gehörenden Vokabular beschrieben sind.

In einem zweiten Schritt kann der Benutzer das Suchergebnis interaktiv mittels automatisch berechneter Facetten einschränken und somit genau an sein Informationsbedürfnis anpassen. Die Schaltfläche „Alle Ergebnisse anzeigen“ führt zu der Ansicht, die in **Abbildung 3** dargestellt ist. Im linken Bereich werden die zur Verfügung stehenden Facetten dargestellt. Im Beispiel wurden die Suchergebnisse bereits eingeschränkt auf alle KTBL-Standardmaschinen, deren Abschreibung zwischen 5.100 und 25.500 EUR pro Jahr liegt. Die Eigenschaft „Höhe“ ist aufgeklappt, die möglichen Wertebereiche zur Auswahl werden angezeigt. Sobald eine Auswahl erfolgt, passt sich die Ansicht im rechten Bereich an. Hier können außerdem weitere Eigenschaften tabellarisch miteingeblendet werden, beispielhaft ist hier lediglich die Eigenschaft „Anschaffungspreis“ angezeigt.

Die Aufnahme neuer Eigenschaften in die Facetten erfordert keinen Programmieraufwand, sondern wird automatisch erkannt. Die Wertebereiche werden dann entsprechend berechnet. Es reicht also aus, Attribute lediglich im Datensatz anzuhängen, die Zufügung muss der Oberfläche nicht separat bekannt gemacht werden. Eine solche Funktionalität wäre auf Basis herkömmlicher relationaler Datenbanktechnologie im Vergleich sehr aufwendig zu realisieren, genauso wie Abfragen oder Suchen, die wie oben beschrieben auch die Attributbezeichnungen (Spaltennamen in relationaler Datenhaltung) oder Beziehungen zwischen Entitäten einbeziehen. Hier spielt die semantische Technologie die Vorteile aus, die sich durch die syntaktische Vereinheitlichung der Daten und der zugehörigen Beschreibung der Daten (Metadaten) ergibt.

Schlussfolgerungen

Sowohl der Linked-Open-Data-Service als auch die semantische Suche ließen sich sehr einfach in die bestehende Infrastruktur integrieren. An der relationalen Datenbank im Hintergrund waren keine Anpassungen an bestehenden Zugangsmechanismen, Datenbankschemas oder Tabellen notwendig. Mittelfristig kann die Implementation durch die Erstellung zusätzlicher Sichten (VIEWs gemäß SQL-Standard) optimiert werden, ein zwingender Bedarf ergab sich hierfür jedoch im prototypischen Betrieb nicht. In der Planung ist derzeit, durch eine automatisierte Nachbearbeitung des aus der Datenbank generierten Datensatzes weitere Beziehungen zuzufügen, sodass die Navigation im Dienst und die Möglichkeiten, nach bestimmten Zusammenhängen zu suchen, noch verbessert werden. Bereits jetzt konnte jedoch mit überschaubarem Aufwand innerhalb eines Zeitraumes von etwa drei Monaten eine vollständig spezifikationsgemäß funktionierende Implementation erzielt werden.

Ein Nebeneffekt des Linked-Open-Data-Dienstes ist es, dass auch Standard-Internetsuchmaschinen KTBL-Daten nun finden können, da sie wegen der feingranulierten Aufteilung in über verschiedene URIs abrufbare Daten und die darin enthaltenen Verknüpfungen mithilfe der von den Betreibern verwendeten Crawlern indiziert werden können. Mit der spezifisch implementierten semantischen Suche können derzeit fachliche Beziehungen noch gezielter ausgewertet werden, auch Internetsuchmaschinen-Betreiber befassen sich zunehmend mit dem Thema der Interpretation von Semantik um die Treffergenauigkeit zu verbessern.

Derzeit befindet sich der Prototyp im internen Probebetrieb. Erste Indikatoren lassen darauf schließen, dass die Nutzer von den Möglichkeiten profitieren werden, sich schnell einen Überblick über vorhandene Daten verschaffen zu können. Eine Bereitstellung für Externe ist für das Jahr 2014 geplant. Ziel ist, dass Berater, Landwirte und andere Nutzer innerhalb des KTBL-Datenangebotes schneller und zielgerichteter Anwendungen, Daten und Veröffentlichungen finden, die für ihre Fragestellungen Antworten liefern können.

Cyganiac und Jentzsch haben ein Diagramm von weltweit verfügbaren Linked-Open-Data-Diensten sowie deren Verknüpfungen untereinander erstellt [16]. Bereits jetzt stehen eine Vielzahl von Daten innerhalb des Linked-Open-Data-Netzwerkes zur Verfügung, darunter auch Daten mit mehr oder weniger engem Bezug zur Landwirtschaft, z. B. AGROVOC (multilingualer landwirtschaftlicher Fachthesaurus), Geonames (Herstellung von Regionalbezug), Eurostat (Statistische Daten), Drug Bank (u. a. Tierarzneimittelwirkstoffe). Auch der in o. g. Diagramm zentral dargestellte Dienst des Netzwerkes, dbpedia.org, enthält eine Reihe von landwirtschaftlichen Konzepten und deren semantische Beschreibung. Die Europäische Union arbeitet derzeit außerdem an der Bereitstellung von Daten aus dem Bereich des Verbraucherschutzes – hierzu zählen unter anderem auch Daten zu Pflanzenschutzmitteln – als Linked Open Data. Im obigen Kontext werden aus dem im Diagramm dargestellten Netzwerk bislang lediglich Daten aus dem AGROVOC verwendet. Potenzial für eine Nutzung haben jedoch auch die Daten aus dbpedia.org. Bei weiterer Verbreitung der Technik auch im landwirtschaftlichen Bereich können in Zukunft weitere Informationen anderer Anbieter, die zu der Anfrage passen, angezeigt werden. Naheliegend ist es beispielsweise Daten zu Kulturen, Sorten, eingesetzten Maschinen und Betriebsmitteln (Pflanzenschutzmittel, Düngemittel usw.) auf Basis derselben Technologien bereitzustellen. Derzeit durchgeführte Arbeiten zielen darauf ab, schon jetzt öffentlich verfügbare Daten dahingehend aufzubereiten und mit einzubinden bzw. zu verlinken.

Die maschinenlesbare Schnittstelle kann von Entwicklern bei Herstellern von Farm-Management-Informationssystemen und Entscheidungsunterstützungssystemen genutzt werden, um neue Funktionalitäten zu implementieren. Beispielsweise

wäre es möglich, Standardkostensätze für Maschinen gezielt abzufragen und zu importieren. Interessensbekundungen liegen auch für Standardproduktionsverfahren vor, die derzeit noch nicht im Datensatz enthalten sind, die aber eine weitere Ausbaustufe oberhalb der Arbeitsverfahren darstellen können. Auch seitens des KTBL können so neue Anwendungen durch Daten externer Einrichtungen angereichert werden.

Im Detail wurden auch Limitationen gefunden. So ist es beispielsweise in der d2rq mapping language nicht möglich, in RDF übliche Angaben zur Sprache an Bezeichnungen anzuhängen, die aus Inhalten mehrerer Datenbanktabellen zusammengesetzt wurden. Auch die Abbildung, Interpretation und Sortierung von physikalischen Größen kann noch verbessert werden. Dieses Defizit wird auch in den oben gegebenen Beispielen ersichtlich, in denen Auszeichnungen mit Einheiten (z. B. Höhe in Meter, Anschaffungspreis in Euro) noch komplett fehlen. In agroRDF wird die QUDT-Ontologie [17] für die Repräsentation von Einheiten und Größen verwendet. Diese soll auch hier in Zukunft genutzt werden, um Umrechnungen und eine ansprechendere Darstellung in Bedienoberflächen zu ermöglichen.

Künftige Arbeiten am Linked-Open-Data-Service umfassen voraussichtlich außerdem die Einbindung weiterer Planungsdatensätze, die Verknüpfung mit weiteren, externen Datensätzen, die Optimierung der graphen-orientierten Repräsentation der relationalen Daten, eine weitere Ausarbeitung des Vokabulars in der Breite und auch auf mehrsprachiger Ebene und die Bereitstellung von begleitender Dokumentation.

Insgesamt stellen die Arbeiten einen zukunftsweisenden Ansatz zur Etablierung von innovativen und flexibleren Lösungen zur Nutzung des KTBL-Datenangebots und insbesondere zur Beantwortung von Fachfragen dar, die langfristig die integrierte Nutzung von Daten erlauben, die in verschiedenen Organisationen erhoben, bereitgestellt und aktuell gehalten werden. Im Idealfall erfolgt dies auf Maschinenebene ohne vorherige Abstimmung technischer Details zwischen den beteiligten Einrichtungen.

Literatur

[1] Kuratorium für Technik und Bauwesen in der Landwirtschaft (Hg.) (2009): Faustzahlen für die Landwirtschaft. 14. Auflage, Darmstadt, KTBL e.V.

[2] Kuratorium für Technik und Bauwesen in der Landwirtschaft (Hg.) (2012): Betriebsplanung Landwirtschaft 2012/13 – Daten für die Betriebsplanung in der Landwirtschaft. 23. Auflage, Darmstadt, KTBL e.V.

[3] Mitra, N.; Lafon, Y. (2007): SOAP Version 1.2 Part 0: Primer (Second Edition). http://www.w3.org/TR/soap12-part0/, Zugriff am 23.1.2014

[4] Bray, T.; Paoli, J.; Sperberg-McQueen, C. M.; Maler, E.; Yergeau, F. (2008): Extensible Markup Language (XML) 1.0. Fifth Edition, http://www.w3.org/TR/xml/, Zugriff am 23.01.2014

[5] Heath, T.; Bizer, C. (2011): Linked Data – Evolving the Web into a Global Data Space. Morgan & Claypool Publishers

[6] Berners-Lee, T. (2009): Linked Data. http://www.w3.org/DesignIssues/LinkedData.html, Zugriff am 11.12.2013

[7] Brickley, D.; Guha, R.V. (2004): RDF Vocabulary Description Language 1.0. http://www.w3.org/TR/rdf-schema/, Zugriff am 11.12.2013

[8] Prud'hommeaux, E.; Carothers, G. (2013): Turtle – Terse RDF Triple Language. http://www.w3.org/TR/turtle/, Zugriff am 11.12.2013

[9] Berners-Lee, T.; Fielding, R.; Masinter, L. (2005): RFC 3986 – Uniform Resource Identifier (URI): Generic Syntax. http://www.rfc-editor.org/info/rfc3986, Zugriff am 23.1.2014

[10] Harris, S.; Seaborne, A. (2013): SPARQL 1.1 Query Language. http://www.w3.org/TR/sparql11-query/, Zugriff am 11.12.2013

[11] Martini, D.; Schmitz, M.; Kunisch, M. (2011): Datenintegration zwischen Standards in der Landwirtschaft auf Basis semantischer Technologien. GIL-Jahrestagung: Qualität und Effizienz durch informationsgestützte Landwirtschaft, Gesellschaft für Informatik in der Landwirtschaft e.V., 24.–25. Februar 2011, Oppenheim, S. 133–136

[12] Bizer, C.; Cyganiak, R. (2012): D2RQ – Accessing Relational Databases as Virtual RDF Graphs. http://d2rq.org, Zugriff am 11.12.2013

[13] Food and Agricultural Organization of the United Nations (2012): AGROVOC. http://aims.fao.org/standards/agrovoc/about, Zugriff am 11.12.2013

[14] Crockford, D. (2006): RFC 4627 – The application/json Media Type for JavaScript Object Notation (JSON). http://www.rfc-editor.org/info/rfc4627, Zugriff am 23.1.2014

[15] Ladwig G.; Tran T. (2010): Combining Keyword Translation with Structured Query Answering for Efficient Keyword Search. Proceedings of the 7th Extended Semantic Web Conference ESWC '10, Springer

[16] Cyganiak, R.; Jentzsch, A. (2011): The Linking Open Data cloud diagram. http://lod-cloud.net, Zugriff am 11.12.2013

[17] Hodgson, R.; Keller, P. J.; Hodges, J.; Spivak, J. (2013): QUDT – Quantities, Units, Dimensions and Data Types Ontologies. http://qudt.org, Zugriff am 11.12.2013

Autoren

Daniel Martini ist Teamleiter des Arbeitsschwerpunkts agroXML und **Dr. Martin Kunisch** ist Hauptgeschäftsführer (kommissarisch) am Kuratorium für Technik und Bauwesen in der Landwirtschaft e. V. (KTBL), Bartningstraße 49, 64289 Darmstadt, E-Mail: d.martini@ktbl.de

Daniel Herzig und **Günter Ladwig** sind Geschäftsführer bei Search-Haus - Daniel Herzig & Günter Ladwig Softwarelösungen, GbR, Alter Schlachthof 39 / F3, 76131 Karlsruhe

Hinweise

Teile der vorliegenden Ergebnisse wurden im Rahmen des vom Bundesministerium für Bildung und Forschung unter dem Förderkennzeichen 01IA08005 geförderten iGreen-Projektes erarbeitet.