

Semantic Web im Agrarbereich: Wiederverwendung von Ontologien und Aufbereitung von Daten mit agroRDF

Daniel Martini, Mario Schmitz

Kuratorium für Technik und Bauwesen in der Landwirtschaft e. V. (KTBL)
Bartningstraße 49
64289 Darmstadt
d.martini@ktbl.de
m.schmitz@ktbl.de

Abstract: In landwirtschaftlichen Datenaustauschszszenarien rückte zunehmend die Frage in den Fokus, wie Daten aus verschiedenen Quellen und in verschiedenen Formaten in Prozessen integriert und einheitlich zugänglich gemacht werden können. Flexibilität und Erweiterbarkeit sind dabei entscheidende Anforderungen an die Technologie, in der Daten abgebildet werden. Semantische Technologien bieten die Möglichkeit, die in anderen Repräsentationstechnologien vorgefundenen Restriktionen zu überwinden. Entsprechende Entwicklungsarbeiten wurden im Rahmen der Arbeit an agroXML durchgeführt.

1 Zielsetzung und Motivation

In den letzten Jahren rückte zunehmend die Frage in den Fokus, wie Daten im landwirtschaftlichen Kontext aus verteilten Quellen und in verschiedenen Formaten in Prozessen integriert und einheitlich zugänglich gemacht werden können. Flexibilität und Erweiterbarkeit sind dabei Kernanforderungen an die Technologie, in der Daten abgebildet werden. Es zeigte sich, dass XML diese Anforderungen nicht erfüllen kann. Das dynamische Hinzufügen von Inhalten ist nur schwer handhabbar, da das grundlegende Modell von XML eher für die Auszeichnung von Dokumenten mit hohem Anteil von textlichem Inhalt ausgelegt ist. Strukturen müssen vorab bekannt sein, um verarbeitet werden zu können, variable Querbezüge zwischen einzelnen Entitäten sind wie auch in Datenbanken nur ineffizient darstellbar. Semantische Technologien bieten zum Einen eine Darstellungsform für beliebige Datenstrukturen, sodass sich Datenbestände, die nach verschiedensten Modellierungsparadigmen (relational, hierarchisch, objekt-orientiert usw.) aufgebaut sind, leicht überführen lassen, zum Anderen können Datenmodelle leicht ad-hoc je nach Bedarf erweitert werden. Softwarewerkzeuge können zur Laufzeit auf diese Erweiterungen reagieren, da Schema und Daten einheitlich dargestellt werden, und somit gezielte Abfragen nach neuen Eigenschaften, Klassen oder Typen möglich sind. Diese grundlegenden Eigenschaften semantischer Technologien waren Motivation, zu überprüfen, inwiefern bei der Repräsentation landwirtschaftlicher Daten Vorteile insbesondere in Bezug auf Integration und Interoperabilität erzielt werden können.

2 Semantische Modelle und agroRDF

Semantische Modelle versuchen außer der Struktur von Daten auch deren Bedeutung festzuhalten. Da Computer sich selbst keinen Begriff von der Bedeutung von Konzepten machen können, wird hierbei darauf zurückgegriffen, die Bezüge zwischen Konzepten explizit und maschinenlesbar festzuhalten. Eine solche explizite Konzeptualisierung einer Anwendungsdomäne in einer formalen Sprache heißt Ontologie [Gr93]. Dem Rechner werden hierdurch “vorformulierte” Assoziationen zu bestimmten Termini, Datenentitäten o.ä. geliefert. In aller Regel bedient man sich dabei der Datenstruktur des Graphen, in dem zentrale Elemente Knoten und Kanten sind. Für die syntaktische Darstellung von Graphen gibt es verschiedene Sprachen, die meist auf dem Konzept der Adjazenzliste aufsetzen, bei der jeweils durch eine Kante verbundene Knoten gemeinsam notiert werden (vgl. hierzu z. B. das Tripelmodell des Resource Description Frameworks (RDF) [MM04]). agroRDF ist ein solches, explizit formuliertes, semantisches Modell für den Agrarbereich.

Die grundlegenden Konzepte und Inhalte von agroRDF wurden bereits 2011 beschrieben [MSK11]. Aus technischer Sicht ist agroRDF ein Satz von modular aufgebauten, überschaubaren und einfachen Ontologien, festgehalten in der formalen Sprache RDF Schema [BG04] unter Verwendung der sogenannten Turtle-Syntax [BB13] für RDF [KC04]. agroRDF baut auf agroXML auf. Verknüpfungen von den bereits vorhandenen XML Schemas in die RDF Schemas werden über den SAWSDL-Standard der W3C hergestellt [FL07]. Dabei werden Schemaelemente explizit durch weitere Attribute mit im semantischen Modell definierten Klassen oder Eigenschaften verbunden. Aus fachlicher Sicht deckt agroRDF dieselben Konzepte der landwirtschaftlichen Domäne wie agroXML ab. Die Klassen und Eigenschaftsdefinitionen und deren Zusammenhänge umfassen beispielsweise: Grundlegende Stammdaten des Betriebes (Anschrift, Kontaktdaten wichtiger Ansprechpartner...), Ackerflächen, Betriebsmittel (Pflanzenschutzmittel, Düngemittel...), Arbeitsprozesse (Aussaat, Ernte, Düngung...), statische Maschinendaten (Leermasse, Abmessungen...). Neben der Außenwirtschaft wurden im Rahmen des InfrAgrar-Projektes Konzepte der Innenwirtschaft ergänzt. Das Projekt beschränkte sich auf die Schweinehaltung und die Wertschöpfungskette von der Ferkelerzeugung bis zum Schlachthof. Hinzugefügt wurden Klassen und Eigenschaften zur Beschreibung von Orten (Stall, Bucht...), Tieren (Geschlecht, Abstammung...) und den jeweils zugewiesenen Ereignissen (Geburt, Umbuchten, Schlachtung...).

In der anfänglichen Entwicklung wurden alle Definitionen von Klassen und Eigenschaften auf Basis der agroXML Schemas eigenständig ausformuliert. Später wurden zunehmend externe Ontologien wiederverwendet. So wird die QUDT-Ontologie [HK13] zur Beschreibung physikalischer Größen genutzt, die vCard-Ontologie [IM13] liefert Konstrukte zur Einbindung von Daten zu Personen und Organisationen. Das folgende kurze Beispiel zeigt einige relevante Aspekte dieser gezielten Wiederverwendung auf:

```
1   @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
2   @prefix vcard: <http://www.w3.org/2006/vcard/ns#> .
3   @prefix qudt-unit: <http://qudt.org/vocab/unit#> .
4   @prefix machine: <http://www.agroxml.de/rdf/vocabulary/machine#> .
5   :tractor0815
```

```

6           a machine:Machine ;
7           machine:length "4.18"^^qudt-unit:Meter .
8   :Operator
9           rdfs:subClassOf vcard:Individual .
10          :operator3141
11          a :Operator ;
12          vcard:fn "Otto Operator" .

```

Über die in Zeile 1-4 deklarierten namespaces sind in aller Regel formale Beschreibungen der jeweils zugehörigen Vokabularien zugänglich. So findet sich beispielsweise unter der in Zeile 3 gegebenen URL die Spezifikation der im QUDT-Vokabular definierten Einheiten. Dort sind wiederum weitere in RDF beschriebene fachliche Informationen zu finden, z. B. offizielle Einheitenkürzel und Umrechnungsfaktoren, sodass generische Softwarefunktionen zur Bearbeitung beliebiger Einheiten implementiert werden können. Genutzt wird eine der Einheiten in Zeile 7. Zeile 12 zeigt die Wiederverwendung einer Eigenschaft aus dem vCard-Vokabular. In Zeile 8 und 9 wird hierbei zunächst beschrieben, dass der `:Operator` eine Art `vcard:Individual` ist. Hierdurch werden sämtliche Eigenschaften, die für `vcard:Individuals` gelten (u. a. auch Telefonnummern und Adressen) zugänglich. Eigene Modellierungen dieser Daten sind nicht notwendig. Beschreibungen von Daten und Daten selbst können wie in den Blöcken von Zeile 8-9 und 10-12 in RDF beliebig miteinander gemischt werden. Insgesamt fällt die kompakte, gut lesbare Syntax auf, die die Entwicklungsarbeit gegenüber anderen Repräsentationen erleichtert.

Umfangreichere Datensätze auf Basis von realen Daten wurden in den genannten Projekten erstellt und verwendet. Bestehende Datenbestände aus relationaler Datenhaltung können mit Werkzeugen wie D2RQ (<http://d2rq.org>) in Infrastrukturen des Semantic Web eingebunden werden. In einem Mapping wird hierbei spezifiziert, welche Tabellen in welche Klassen und welche Spalten in welche Eigenschaften abgebildet werden sollen. Es können weitere Bedingungen angegeben werden und mit Hilfe von Translation Tables beliebige Eigenschaften an Datensätze angehängt werden, sodass hiermit ein mächtiges Werkzeug zur Verfügung steht, um in relationalen Datenbanken abgelegte Daten in ein semantisches, graphenorientiertes Modell zu überführen, zugänglich zu machen und die Datensätze mit weiteren Bezügen zusätzlich anzureichern und aufzubereiten.

3 Fazit

Ein Vergleich des Quelltextes der agroXML Schemas mit den agroRDF Schemas zeigt, dass eine identische Ausdrucksstärke mit etwa einem sechstel an Volumen erreicht werden kann. Die Notation von RDF Schemas in Turtle ist deutlich übersichtlicher als XML Schema und unterstützt Entwickler daher bei der Einarbeitung und Implementation. Konsequente Wiederverwendung von Konzepten aus bekannten, etablierten Vokabularien hilft, Interoperabilität auch über eine beliebige Anzahl an Webdiensten hinweg zu erzielen und ermöglicht damit erst die im Umfeld anderer Technologiestacks propagierte lose Kopplung. Die Einbindung von externen Vokabularien in eigene Datenmodelle erfolgt allerdings nach wie vor nicht von selbst sondern muss aktiv betrieben werden. Für die dafür im Vorfeld notwendige Recherchearbeit existieren derzeit noch keine all-

gemein anerkannten Werkzeuge. Registries oder spezielle Suchmaschinen könnten hier in Zukunft noch zu deutlicher Erleichterung der Entwicklungsarbeit beitragen. Es kann jedoch festgehalten werden, dass Werkzeuge zur Bearbeitung und Bereitstellung von Daten als RDF in Anwendungen heutzutage ausgereift sind. Somit wird es möglich, die in der Landwirtschaft oft bestehenden komplexen Realweltbeziehungen von Entitäten jenseits vorgefertigter Schlüssel-Fremdschlüsselbeziehungen oder fester hierarchischer Strukturen abzubilden. Trotz ihrer Mächtigkeit ist die Technologie insgesamt einfach zu handhaben, sodass durch ihre Nutzung ein beträchtliches Potenzial der Vereinfachung landwirtschaftlicher Informationsmanagement-Infrastrukturen besteht.

4 Danksagung

Die Förderung des iGreen-Projektes, in dem Grundlagen zu agroRDF entwickelt wurden, erfolgte aus Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01IA08005X. Die Förderung des InfrAgrar-Projektes, in dem Teile der dargestellten Arbeiten durchgeführt wurden, erfolgt aus Mitteln des Bundesministeriums für Ernährung, Landwirtschaft und Verbraucherschutz (BMELV) im Rahmen des Programms zur Innovationsförderung.

Literaturverzeichnis

- [BG04] Brickley, D.; Guha, R. V.: RDF Vocabulary Description Language 1.0: RDF Schema. World Wide Web Consortium, 2004. <http://www.w3.org/TR/rdf-schema/>.
- [BB13] Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G.: Turtle – Terse RDF Triple Language. World Wide Web Consortium, 2013. <http://www.w3.org/TR/turtle/>.
- [FL07] Farrell, J.; Lausen, H.: Semantic Annotations for WSDL and XML Schema. World Wide Web Consortium, 2007. <http://www.w3.org/TR/sawSDL/>.
- [Gr93] Gruber, T. R.: A Translation Approach to Portable Ontology Specifications. In: Knowledge Acquisition, 5(2), 1993. S. 199-220.
- [HK13] Hodgson, R., Keller, P. J., Hodges, J., Spivak, J.: QUDT – Quantities, Units, Dimensions and Data Types Ontologies. 2013. <http://qudt.org>.
- [IM13] Iannella, R., McKinney, J.: vCard Ontology for describing People and Organisations. World Wide Web Consortium, 2013. <http://www.w3.org/TR/vcard-rdf/>.
- [KC04] Klyne, G.; Carroll, J. J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. World Wide Web Consortium, 2004. <http://www.w3.org/TR/rdf-concepts/>.
- [MM04] Manola, F., Miller, E.: RDF Primer. World Wide Web Consortium, 2004. <http://www.w3.org/TR/rdf-primer/>.
- [MSK11] Martini, D., Schmitz, M., Kunisch, M.: Datenintegration zwischen Standards in der Landwirtschaft auf Basis semantischer Technologien. In (Clasen, M.; Schätzel, O.; Theuvsen, B., Hrsg.): Qualität und Effizienz durch informationsgestützte Landwirtschaft – Referate der 31. GIL-Jahrestagung 24.-25. Februar 2011, Oppenheim. GI-Edition Lecture Notes in Informatics, 181, 2011. S. 133-136.
- [PH12] Perry, M., Herring, J.: GeoSPARQL – A Geographic Query Language for RDF Data. Open Geospatial Consortium, 2012.