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

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

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.

received 12 December 2013
accepted 30 January 2014

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

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

Information plays an increasing role in agriculture as a basis for farm management decisions. Fundamental data for preparation of investments and planning of on-farm production processes and work procedures have been collected since a longer time already by the Association for Technology and Structures in Agriculture (KTBL) and made available after processing and testing for reliability. If, for instance, a farmer considers starting production of quality wheat and wants to carry out an economic profitability analysis, necessary data such as fixed and variable costs for new machinery, agricultural operating supplies or labour are available.

In the past, the most important medium were publications such as Faustzahlen Landwirtschaft (Estimates for agriculture) [1] or the Datensammlung Betriebsplanung (data collection for farm business planning) [2]. Printed material also dominated for provision of other important data for preparation and planning within farm production e.g. crop variety field trial results and variety lists. However, for a number of years now the internet plays an increasing role in publishing and dissemination of data in agriculture as well. So far, in general, data offered online have been closely integrated with their application and only accessible to the user via already prepared interaction routes established by the respective developers in the graphical user interfaces. A stringent separation of data services and application logic is mostly missing within the increasingly popular apps for smartphones as well. Further use and processing of data within other systems e. g. in farm management information systems or advisory tools or the aggregation and integration of data bases from different organisations, is therefore difficult to achieve, requiring highly demanding work for maintenance or import. Users of data are thus forced to use the respective accompanying applications. Especially with smartphones this results in an accumulation of numerous applications in systems, often covering similar functionalities or with overlapping contents that are anyways incompatible.

Fundamental technical methods

Technologies around the semantic web by now permit more elegant and flexible solutions for the user compared with the above paradigms whereby application and data are bound closely together - or also compared with traditional web services based on SOAP (Simple Object Access Protocol [3]) and XML (eXtensible Markup Language [4]). Within the framework of the linked open data (LOD) initiative of the World Wide Web Consortium (W3C), the idea is to try to separate data from evaluation logic and user interfaces and to provide access through simple Internet protocols (mainly HTTP) [5; 6]. In particular, data are also provided in machine-readable, standardised formats so that they can be accessed via simple URL requests and automatically read by applications. For representation of data in LOD services the Resource Description Framework (RDF) from the W3C is recommended and is generally also used by most suppliers. Data sets come with accompanying vocabularies, which can be created e. g. in RDF schema [7]. Core aspect is the model-

ling of data as a directed graph. Such a graph comprises nodes and edges. Nodes can be linked with any number of further nodes via the respective edges. Similar data structures are used for solving problems in a number of application areas such as routing for navigation systems or in communication networks (mobile phone networks, Internet). In the context of the semantic web they help build-up a network of statements. Thereby, two nodes on an adjacent edge together with their relating edge create a statement, which takes on the form subject-predicate-object or thing-property-value, a so-called triple. Using a very simple example, the most important characteristics of this form of representation can be explained:

	Subject	Predicate	Object
Tripel 1:	FarmerXY	owns	Machine0815
Tripel 2:	Machine0815	is a	tractor
Tripel 3:	Machine0815	purchase price	83,000 Euro

Triple 1 relates a farmer to a machine. The two nodes of the graph are farmer XY and machine 0815. In this case, the edge would be the relationship "owns". The object in this statement is used as a subject in the following statement (triple 2). Hereby a new edge is opened from node machine 0815, which via a "isA" relationship specifies the node more precisely. The existing object in this triple – the tractor – can also be described in more detail in further triples, e. g., via a subclass relation as a type of farm machinery. Triple 3 illustrates the usage of so-called literals, which are generally used to add attributes and their values to objects in graphs (e. g. numerical or character strings). No further edges can proceed from literals although they can be connected with a data type (in this example "Euro"). Triples such as those presented above are represented in a syntax such as turtle [8]. Additionally, all nodes and edges with the exception of the literals are represented as Uniform Resource Identifiers (URIs [9]). Hereby, data sets can be connected with further data on other servers so that it becomes now possible to create a "global data space" [5]. In the section "linked open data service" there is an example illustrating turtle syntax and usage of URIs. An important property of a graph representation is that data sets of all types can be transformed into this generic data structure. Entries in data base tables (rows) can e. g. be represented as objects, the columns of the tables can be represented as attributes. With suitable tools, transformations can be carried out "virtually", i. e., they take place on-the-fly during run-time and require no change of existing infrastructures for data storage. Additionally, relationships between data entities are explicitly modelled and specified. This means that relationships represented in relational data bases only as key-foreign key-relations are, in this type of modelling, given an identifier that also allows a targeted request regarding this relationship (in the above given triples e. g. which nodes are connected through "owns"?). Predetermination of such relationships between entities or fixed tree structures are not necessary, i.e. later on, further types of objects can be flexibly added, that can use already existing relationships as well as bring in new attributes. Because data and schemes or vocabularies are represented in the same format, data sets can be described with simple means in such a way, that any application built on top of them can directly use new added data without any new compilation or manual adjustments. In relational databases such an approach would be possible through queries into the so-called "information schema" and dynamically creating the user interfaces based on the results of these queries. Implementation of this is comparatively complicated, impractical and unportable between databases from different manufacturers.

In summary, we can say that using the described representation technologies, data can be structured and delineated in a way, that is flexible and closely oriented on entity relations of the real world.

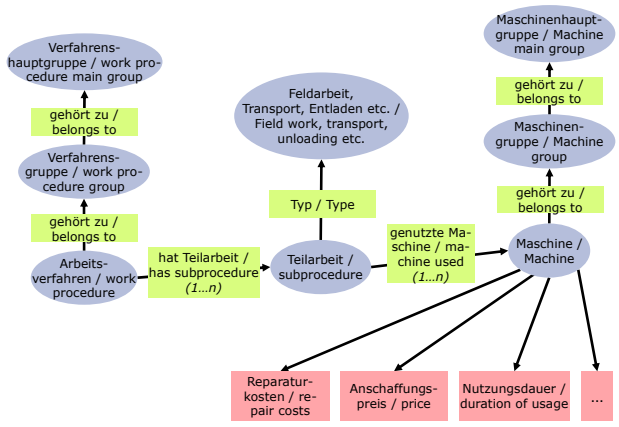
Even complex queries can in most cases be formulated simply by selecting sets of nodes in the graph which fulfil certain conditions. In the RDF environment the query language SPARQL [10] is generally used. Through the characteristics mentioned an application-independent usage of data is simplified. Additionally, data sets are especially accessible for searches, navigation and the answering of complex questions, that require incorporating a larger number of relations.

On a technical level, by using the HTTP protocol, the above mentioned generic data formats and RDF based vocabularies for describing the data access is unified to such an extent that for client systems requesting data, only a minimum of predetermination by the developer is necessary in comparison with other technology (see above).

This enables interconnection between and facilitated extension of data collections.

Within the iGreen project and a long-term cooperation with the FAO, the authors were able to expedite fundamental developments in this area. Among other things, the agroRDF vocabulary was created on the basis of agroXML [11]. Building on such preliminary work the KTBL has now, on the basis of the existing KTBL data base, implemented a linked open data service with information regarding work procedures and machinery costs that allows remote requests on these data by computers. On top of this service SearchHaus – a spin-off from Karlsruhe Institute for Technology (KIT) – applied a semantic search as one possible application enabling targeted finding of data sets. The question, e. g., could be the investment needed for a wheat seed drill. The KTBL data base can then be searched using simple search terms, e. g. "purchase price machinery wheat seed". The questions can be freely formulated and no special knowledge of structure or querying of the data base are required. For possible results, different interpretations are shown, which are determined with the aid of relationships existing in the data. The results can then finally be narrowed down using dynamically calculated facets.

Fig. 1



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

Linked open data service

The linked open data service is implemented using the open source tool d2rq [12] on top of the existing Oracle 11 data base. The d2rq mapping language is used to establish a mapping that specifies how the existing relational data base tables are to be transferred into a RDF graph model. A RDF schema based on agroRDF is applied for describing the data contents. Special concepts of the KTBL data base were built as appropriate extensions. As a result, on the one hand, an HTTP service that can deliver data in HTML format for presentation for the user in the browser as well as in machine-readable turtle format [8] for RDF is supplied. On the other hand, a SPARQL end-point for formally specified queries is available.

At the moment, the service is limited to data on machinery and field work procedures. **Figure 1** shows the most important entities and their relationship to each other. The figure is a heavily simplified presentation, which could be represented in only 10 triples. It shows no connections to external services and should be regarded only as an overview. The RDF schema actually used for describing the data sets involved 190 triples. The data itself includes currently 104 342 triples. Added to this are 180 triples for the description of the KTBL planning and calculation applications which can also be found in the search, as well as the 283 triples of the agroRDF machinery vocabulary. Real data sets are, therefore, mostly substantially more extensive than the three triples needed in the preceding section for illustration of selected examples. Alongside the existing data in the local data base, connections to the multi-lingual AGROVOC thesaurus of the FAO [13] are also built into the data sets so that for some types of machinery, the appropriate concepts as well as the translations for descriptions in numerous languages can be requested. Hereby all that has to be done – as described above – is that the appropriate URIs within the linked open data server of the FAO have to be applied in the subject, predicate or object position of a triple. Such a process of connection can in most cases be conducted in a semi-automatic manner. Also in the case demonstrated here, different simple tools not described in detail were used in combination with subsequent manual readjustments at a few points.

Because the sequential order of statements in the data set is irrelevant and because triples, through the above-mentioned characteristics of the RDF graph data model, can be added later, a system supporting this process – if necessary in an iterative manner – is very simple to realise for the developer.

Thus, client applications need only follow the existing connections when attaching to further external services. Because of the standardisation of the protocol level (HTTP) and the use of a standardised syntax and the self-description of the data sets using vocabularies, an explorative approach to accessing data with the help of a single generic application is fundamentally possible.

The linked open data service can be entered using any given URI included in the data set. In the same way as external services can be tied into own data sets can, vice versa, a tie-in of own services into any other data sets take place.

In the given prototype, the machine-readable results of an HTTP request of the URI http://www.agroxml.de/data/data/machine/110406[1] e. g. looks in a shortened form like this:

```
@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…]
```

It shows the data output for the machine identified by this URI – a KTBL standard machine class, in this case a certain type of combine harvester – in the already mentioned turtle syntax.

Further formats, e.g., on the basis of XML or JavaScript object notation [14] would be basically possible. They are however not yet implemented. In general, turtle was more efficient to parse than XML and at the same time more powerful than JSON as far as the representation of links is concerned. In a way, therefore, it unites the advantages of both worlds. The service can also deliver a simple HTML view suitable for presentation in standard web browsers.

[1] In that this concerns a prototype still under development, all the following URIs are still provisional and currently still lie in a password-protected area. Those interested in accessing the current stage of development should contact the authors.

More attractive interfaces, however, may be realised by requesting the data in the form given above and subsequently processing it. The machinery costs data in the service – also in the context of different work procedures – can thus also be used in farm management information systems (FMIS) for planning of work procedures. The presentation then takes place seamlessly within the respective native FMIS graphical interfaces. Apart from the possibility of entering the linked open data service over another external service or from an earlier accessed and known URI, the general entry point via the root directory http://www.agroxml.de/data/ can be used from which several levels of links lead to all contained data sets. For formal queries a so-called SPARQL end point is available at http://www.agroxml.de/data/sparql. Here, queries on the graphs in the background can be entered using the SQL-similar language SPARQL [10].

The data semantically processed in this way serve as the basis for the search engine.

Linked open data search engine

With the semantic search from SearchHaus, users can simply use key words to search in the KTBL data, just as they are used to doing on web search engines. The search engine also accepts key word requests, interprets the requests with the help of the data base and delivers to the user the appropriate results from the KTBL data. The search engine internally aggregates and groups the complex, structured data so that a rapid answer to the requests is possible, even with large amounts of data involved. In this compact data representation the search result is then determined through exploration algorithms and ranking procedures [15].

Figure 2 shows as an example the search results for the key words "combine harvester purchase price" ("Mähdrescher Anschaffungskosten"). In this case, three possible interpretations were found:

- KTBL standard machinery containing the terms combine harvester and those having purchase price as attribute
- Field operations in which operations are carried out with combine harvesters as machine and those with a purchase price as attribute
- KTBL standard machinery, which belongs to the machinery group "combine harvesters" and that have a purchase price (this is not the same as the first interpretation!)

In standard configuration, only a few results are shown for the first interpretation. For the other interpretations, the information can be suitably expanded using the button with the downward arrow. Not presented in the illustration are resulting hits in further data sources such as KTBL applications and printed publications. These can be opened up using the buttons on the left hand area of the screen. Note at that point that in the case of the semantic search, not only a full text search is carried out in the text fields of the data base but also the relationships and associations between entities and their attributes ("purchase price") are evaluated. Hereby also hyponyms and hyperonyms can be recognized, e. g., "purchase price" as a type of fixed cost, insofar as this is described in the accompanying data vocabulary. In a second stage, the user can restrict the search result interactively through automatically calculated facets to meet information requirements more exactly. The button "show all results" ("alle Ergebnisse anzeigen") leads to the screenview presented in **Figure 3**. The available facets are presented in the left hand area. In the example, the search results are already restricted to all KTBL standard machines with depreciation between 5,100 and 25,500 EUR per year. The attribute "height" ("Höhe") is opened. Possible value ranges are shown for selection. Further attributes can be added in the tabular view on the right. As an example, only the "purchase price" ("Anschaffungskosten") is shown.

Adding new attributes to the facets requires no programming input, they are automatically incorporated. The value ranges are then appropriately calculated.

Therefore, it is sufficient to simply attach attributes in the data set. There is no necessity to make additions explicitly known to the interface. Such a functionality would be very difficult to realize with traditional relational data base technology, exactly as with the requests or searches that, as described above, also include the attribute descriptions (column names in relational data storage) or relationships between entities. Here the semantic technology shows its advantages coming from syntactic unification of data and the accompanying description of data (metadata).

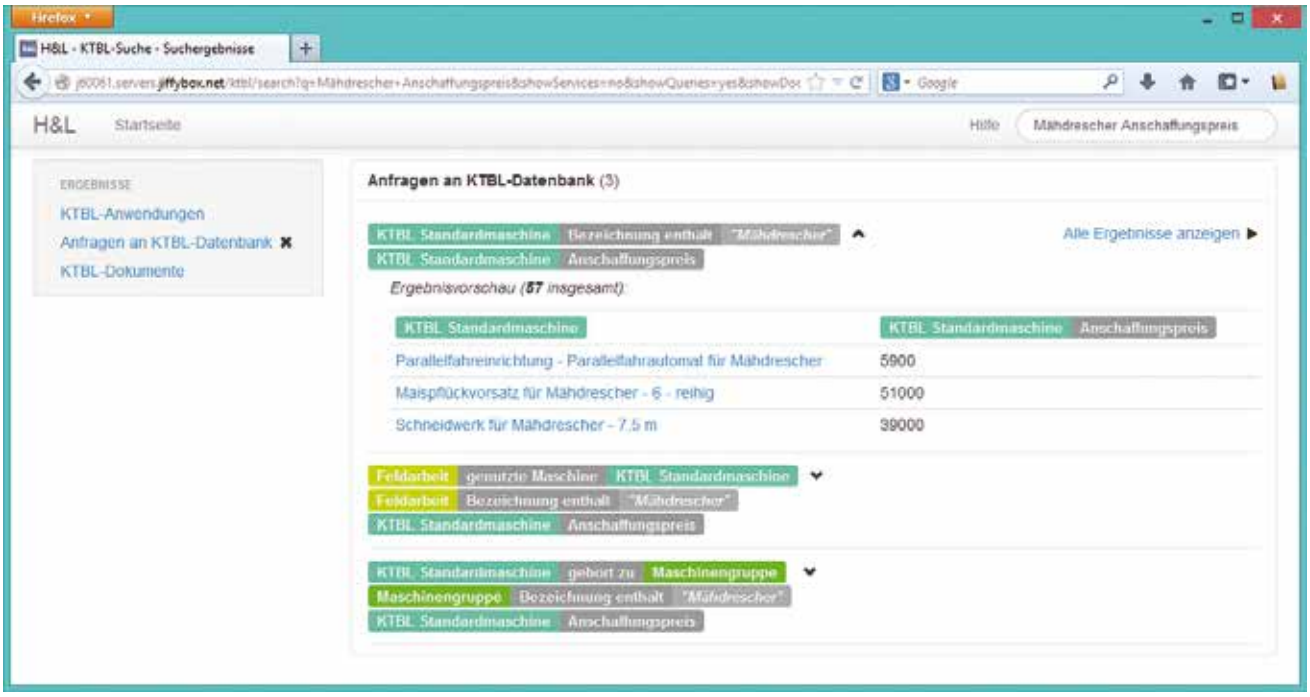
Conclusions

The linked open data service, as well as the semantic search, are very easy to integrate in existing infrastructures. At the relational data base level, no adjustments were necessary on the existing access mechanisms, data base schemas or tables. In the mid term, implementation can be optimised by preparation of additional views, although there was no urgent need for this in prototype operation. The authors currently plan to add further relationships in automated post-processing of the data sets generated from the data base, so that the navigation in the service and possibilities for searching for certain relations are improved still further. Already, however, it has been possible with limited effort within a period of around three months to achieve a complete implementation regarding functional specifications.

A by-product of the linked open data service is that standard Internet search engines can now access KTBL data, because the fine-grained distribution of data across different URIs and the links they contain allows crawling and indexing. With the implemented semantic search, domain-specific relationships can be evaluated more precisely. Internet search engine providers therefore also increasingly engage in research and development of solutions for the interpretation of semantics for improving search engine results.

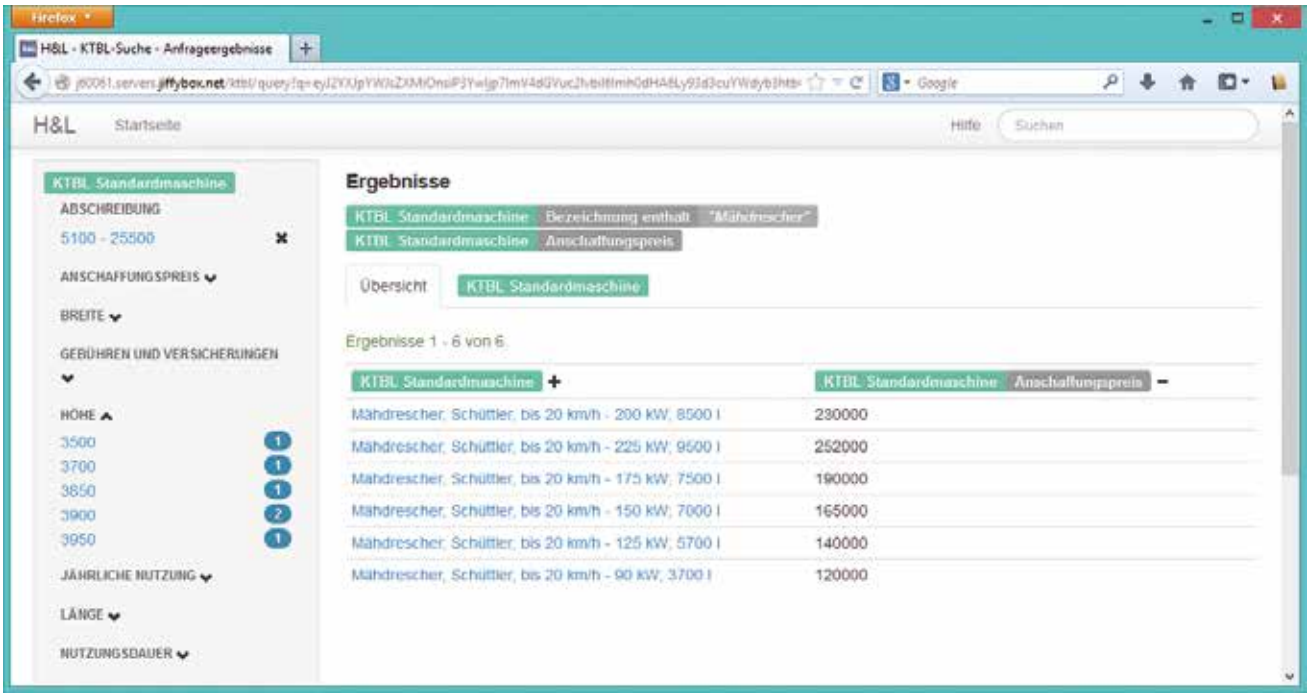
Currently the prototype is being tested internally. There are first indications that users will profit from the possibilities of

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

Fig. 3



Refining Search Results using automatically calculated facets according to properties and their value spaces as given in the data set

being able to achieve a rapid overview of existing data. A preparation for external use is planned for 2014. The aim is that advisers, farmers, and other users, will be able to find data and publications more quickly and accurately in the KTBL data for answers to their questions.

Cyganiac and Jentzsch have prepared a diagram of worldwide available linked open data services that also includes data with a more or less closer relationship to agriculture, e. g. AGROVOC (multilingual agricultural specialist thesaurus), Geonames (providing regionality/locality names and terms), Eurostat (sta-

tistical data), Drug Bank (incl. veterinary drugs active ingredients). Also the services of the network dbpedia.org, centrally presented in the above mentioned diagram, contains a series of agricultural concepts and their semantic descriptions. Additionally, the European Union is currently working at providing data relevant for consumer protection as linked open data, which also includes data concerning plant protection substances. In the above context, so far only data from AGROVOC are used from the network illustrated in the diagram. Application potentials have, however, also the data from dbpedia.org. With future further expansion of the technology also in the agricultural domain additional information from other providers fitting with search questions can be included on the basis of the same technology. Obvious examples are data concerning crops, varieties, applied machinery and inputs (plant protection ingredients, fertiliser, etc.). Work being carried out at the moment is targeted on processing presently publicly available data in this direction and to include this information or provide necessary links.

The machine-readable interfaces can be used by developers in production of farm management information systems and decision support systems in order to implement new functionalities. For instance it would be possible to precisely request and import standard cost sets for machinery. Interest has also been indicated in standard production procedures currently not yet contained in the data set. They could, however, be provided in a further development stage. From the KTBL point of view, also the own, new applications could be enriched by data from external organisations.

In details, limitations have been identified. For instance in the d2rq mapping language it is not possible to attach language information to textual labels as common in RDF if the content is generated by a join of several data base tables. Also the presentation, interpretation and sorting of physical quantities can still be improved. This deficit can also be seen in the examples shown above in which markup of the respective quantities with units (e. g. height in metres, purchase price in euros) is still completely missing. In agroRDF, the QUDT ontology [17] is used for representation of units, quantities and dimensions. The plan is to incorporate this vocabulary here as well, to enable conversions and a more suitable representation in the interfaces.

Future work on the linked open data service will also include integration of further planning data material, connections with further external data sets, optimisation of the graph-oriented representation of relational data, further extension of the vocabulary in the breadth and also on the multilingual level and preparation of accompanying documentation.

In total, the work represents a future-oriented approach to the establishment of innovative and more flexible solutions in using the KTBL data base information and especially for answering specialist questions that, in the long term will be more possible through integrated usage of data recorded, processed and stored within different organisations. Ideally, this integration will take place at machine level without any prior agreement on technical details required between participating organisations.

References

[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

Authors

Daniel Martini is team manager of the agroXML working group and **Dr. Martin Kunisch** is chief executive (com.) at the Association for Technology and Structures in Agriculture e.V. (KTBL), Bartningstraße 49, 64289 Darmstadt, E-Mail: d.martini@ktbl.de

Daniel Herzig and **Günter Ladwig** are managing directors of SearchHaus - Daniel Herzig & Günter Ladwig Softwarelösungen, GbR, Alter Schlachthof 39 / F3, 76131 Karlsruhe

Acknowledgements

Part of the results reported here were produced within the iGreen Project supported by the Federal Ministry for Education and Research under the Development funding code number 01IA08005.