Semantic Integration through Linked Data in the iGreen project

Gunnar Aastrand Grimnes, Malte Kiesel, Mohammed Abufouda, Armin Schröder

Knowledge Management Department
DFKI GmbH
Kaiserslautern
{gunnar.grimnes, malte.kiesel, mohammed.abufouda} @ dfki.de,
armin.schroeder1@gmx.de

Abstract: In this paper we outline the publication and usage of Linked Data in the iGreen project. Several existing structures and datasets have been enriched and represented as Linked Data and made available under the data.igreen-services.com platform.

1. Motivation

In the iGreen project[1] [Be01, TS01] we are attempting to realise a public-private knowledge sharing infrastructure, aiming at giving agriculturists and contractors access to richer, more accurate and more up-to-date decision support. To enable the communication between a diverse and changing set of stake-holders it is crucial to have explicit agreed upon standards for information interchange. In iGreen we have decide to use Linked Data technologies for defining these standards.

2. Linked Open Data

Linked Data[2] [HB01] is a community effort for publishing and linking structured datasets on the internet. Linked Data makes use of existing web standards HTTP and URIs for access and identification, the semantic web technologies Resource Description Framework (RDF) for knowledge representation, and the SPARQL query language for structured queries.

In the recent years, Linked Data has grown a lot and the Linked Data *Cloud* (the combined graph of interlinked datasets) has grown to over 50 billion facts (a fact is a *triple:* a *subject, predicate, object* statement, relating two URIs with some known property).

[1] http://igreen-projekt.de/

[2] http://linkeddata.org/

Linked Data has also enjoyed increased corporate and government interest, with large companies such as Best Buy, Facebook, Google, and the BBC, as well as the government data portals data.gov and data.gov.uk employing RDF and Linked Data technologies.

3. Linked Data in iGreen

The Linked Data vision and standard is an ideal match for the aims of the iGreen project. iGreen was early on dedicated to using semantic technologies for information integration and interchange. To this means we have developed a Linked Data portal for the iGreen project. This contains both structured developed internally in iGreen as well as public data, republished according to Linked Data principles. The portal can be found at http://data.igreen-services.com – it offers a traditional web view of the datasets, browsable using a normal web browser, as well as machine readable RDF representation and a SPARQL endpoint for structured queries. Currently, the site includes: The AgroRDF ontology, an RDFS version of the existing AgroXML standard for agricultural knowledge interchange; an RDF version of the the crop list from the Bundessortenamt (the German Federal Crop Registry); an RDF version of the plant protection registry from the Bundesamt für Verbraucherschutz und Lebensmittelsicherheit (BVL, the German Federal Office for consumer protection and food safety); and an RDF version of the ISOXML (ISO 11783) DDE Registry[1].

3.1. Enriching the datasets

In addition to just republishing the existing datasets, we have also created additional links between the crop list and the plant protection list, linking plant protection agents to the exact species for which it is approved. This allows querying across both datasets, enabling queries such as “which agents may I apply on my crop of Monalisa” — where the fact that *Monalisa* is a type of potato is stored in one dataset, and the information on which agents are appropriate for potatoes in the other.

The plant protection lists refer to usage areas taken from the EPPO Plant Protection Thesaurus[2]. Unfortunately, the restrictive terms of the EPPO prevent us from making use of the thesaurus, and the mappings between the crops lists and plant protection usage areas were created semi-automatically based on word-overlap in the labels.

We also mapped the *species* in the crop lists to concepts from DBpedia[3] [BLK01] where appropriate. DBpedia is a Linked Data version of Wikipedia, and DBpedia identifiers are commonly used for high-level data-integration in the Linked Open Data cloud. By linking species to their DBpedia concepts, extra information that is not available in the

[1] At the time of writing the state of the republishing rights for some of the datasets is still unclear, and they are currently password protected and for project internal use only. We hope to lift this restriction later in the year.
[2] http://eppt.eppo.org/
[3] http://dbpedia.org

dumps from the BVL can be pulled in. For example, we can import photos for illustrations, labels in other languages, or the full scientific classification tree.

4. Benefits of Linked Data

The main benefit of publishing Linked Data in iGreen is the globally unique identifiers we now have for talking about certain crops, plant protection agents, etc., this is an important first step to flexible interoperability. A second benefit is provided by the Linked Data being self-describing, if an HTTP identifier is resolved with a HTTP client library, a machine readable description is returned. This makes it possible to understand messages one has not seen before, and for programs to deal correctly with data and structures that did not yet exist when the program was written. For instance, if a new type of potato is approved for cultivation in Germany, and an iGreen OnlineBox (See [BT01] for more details on the iGreen OnlineBox and the rest of the iGreen architecture) receives a request for combating a certain pest on a crop of this new type of potato, a program can automatically look up the definition, discover that this is a type of potato, and query the plant-protection list for appropriate agents.

A third benefits comes from the flexible data representation of RDF and the powerful query capabilities offered by SPARQL. Although the datasets we have converted were already offered online, they had limited query capabilities through web-forms. The SPARQL endpoints support the SPARQL protocol, letting programs execute queries and retrieve results. This is like offering an API, but the searches you can do are not fixed, and you are free to construct your own API. This is important, as different people have different entry points to the data, for example, for the plant-protection data, some people are interested in what agents may be deployed against a certain pest, others on a certain crop, others again on which agents a certain company holds the approval rights for.

Another potential long-term benefit is that by making the data available as Linked Open Data, we encourage third parties to also make use of our identifiers and structures, thus extending the interoperability benefits far beyond the project.

5. Future Work

Future plans for the iGreen Linked Data portal include deploying collaborative tools for vocabulary specification. In the cases of the datasets already published, there are already processes and organisations in place for maintenance and control of the lists (i.e., ISOBUS for ISOXML, BVL). However, iGreen also needs structured data for domains such as fertilizers, machine types, tools, etc., where no definite and authoritative list exists. Instead, we hope to provide a collaborative platform for specifying, extending, and discussing the datasets we need in the project. We have already started a pilot study using

Semantic MediaWiki[1] for defining a machine ontology. Semantic MediaWiki extends the software used by Wikipedia with support for formal definition of categories, properties and instances. We have bootstrapped the machine ontology with the machine classes available from the KTBL data catalogue, and our domain experts will enhance and extend this.

With the availability of a wide range of Linked Data sets and supporting infrastructure we hope that iGreen can become a significant contributor to the agricultural space of the Linked Data Cloud.

Acknowledgements

The German ministry for education and research (Bundesministerium für Bildung und Forschung, BMBF) funded the iGreen project in the IKT-2020 framework under funding-code 01IA08005A.

References

[Be01] A. Bernardi: iGreen: Organisationsübergreifendes Wissensmanagement in öffentlich-privater Kooperation. In: Automatisierung und Roboter in der Landwirtschaft. KTBL-Tage-2010, Erfurt, Germany.

[BT01] A. Bernardi, C. Tuot: Raum-Zeit-bezogene Agrardaten für die Anforderungen von morgen: Semantische Datenspeicherung in dezentralen, offenen Architekturen. In: Proc. of GIL 2011, Oppenheim, Germany.

[BLK01] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann: DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Issue 7, Pages 154–165, 2009.

[HB01] Tom Heath and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

[TS01] C. Tuot; W. Schneider: Semantische Technologien für ein öffentlich-privates Wissensmanagement im Agrarbereich. GIL Jahrestagung. Gesellschaft für Informatik in der Land-, Forst- und Ernährungswirtschaft 2010.

[1] http://semantic-mediawiki.org/