Chapter 8

Linked Data Usages in DataBio

**Raul Palma, Soumya Brahma, Christian Zinke-Wehlmann,
Amit Kirschenbaum, Karel Charvát, Karel Charvat Jr., and Tomas Reznik**

Abstract One of the main goals of DataBio was the provision of solutions for big data management enabling, among others, the harmonisation and integration of a large variety of data generated and collected through various applications, services and devices. The DataBio approach to deliver such capabilities was based on the use of Linked Data as a federated layer to provide an integrated view over (initially) disconnected and heterogeneous datasets. The large amount of data sources, ranging from mostly static to highly dynamic, led to the design and implementation of Linked Data Pipelines. The goal of these pipelines is to automate as much as possible the process to transform and publish different input datasets as Linked Data. In this chapter, we describe these pipelines and how they were applied to support different uses cases in the project, including the tools and methods used to implement them.

8.1 Introduction

Linked Data has been extensively used in the DataBio project as a federated layer to support large-scale harmonization and integration of a large variety of data collected from various heterogeneous sources and to provide an integrated view on them. Accordingly, as part of the project, we generated a large number of linked datasets. In fact, the triplestore populated during the course of DataBio with Linked Data has over 1 billion triples, being one of the largest semantic repositories related to agriculture. The dataset has been recognized by the EC Innovation Radar as ‘arable farming data integrator for smart farming.’ In addition, we have deployed different

R. Palma (✉) · S. Brahma

Poznan Supercomputing and Networking Center (PSNC), ul. Jana Pawła II 10, 61-139 Poznan, Poland

e-mail: rpalma@man.poznan.pl

C. Zinke-Wehlmann · A. Kirschenbaum

Institute for Applied Informatics (InfAI), University of Leipzig, Goerdelerring 9, 04109 Leipzig, Germany

K. Charvát · K. Charvat Jr. · T. Reznik

Lesprojekt Služby Ltd., Martinov 197, 27713 Záryby, Czech Republic

© The Author(s) 2021

C. Södergård et al. (eds.), *Big Data in Bioeconomy*,
https://doi.org/10.1007/978-3-030-71069-9_8

endpoints providing access to some dynamic data sources in their native format as Linked Data by providing a virtual semantic layer on top of them.

Given the huge number of data sources, and data formats that were addressed during the course of DataBio, such layer has been realized in DataBio through the implementation of instantiations of a ‘Generic Pipeline for the Publication and Integration of Linked Data,’ which have been applied in different uses cases related to the bioeconomy sectors. The main goal of these pipeline instances is to define and deploy (semi-) automatic processes to carry out the necessary steps to transform and publish different input datasets as Linked Data. Accordingly, they connect different data processing components to carry out the transformation of data into RDF [1] format or the translation of queries to/from SPARQL [2] and the native data access interface, plus their linking, and include the mapping specifications to process the input datasets. Each pipeline instance is configured to support specific input dataset types (same format, model and delivery form), and they are created with the following general principles in mind:

- Pipelines must be directly re-executed and re-applied (e.g., extended/updated datasets).
- Pipelines must be easily reusable.
- Pipelines must be easily adapted for new input datasets.
- Pipeline execution should be as automatic as possible. The final target is to fully automated processes.
- Pipelines should support both: (mostly) static data and data streams (e.g., sensor data).

Most of the Linked Data Publication pipeline instances discussed in this chapter perform the transformation and publication of agricultural data as Linked Data; however, there are also some pipelines that are focused on fishery data or on providing access to geospatial datasets metadata as Linked Data. The ultimate target is to query and access different heterogeneous data sources via an integrated layer, in compliance with any privacy and access control needs.

A high-level view of the end-to-end flow of the generic pipeline, aligned with the top-level DataBio generic pipeline, is depicted in Fig. 8.1. Following the best practices and guidelines for Linked Data Publication [3, 4], these pipelines (i) take as input selected datasets that are collected from heterogeneous sources (shapefiles, GeoJSON, CSV, relational databases, RESTful APIs), (ii) curate and/or preprocess the datasets when needed, (iii) select and/or create/extend the vocabularies (e.g., ontologies) for the representation of data in semantic format, (iv) process and transform the datasets into RDF triples according to underlying ontologies, (v) perform any necessary post-processing operations on the RDF data, (vi) identify links with other datasets and (vii) publish the generated datasets as Linked Data and applying required access control mechanisms. The transformation process depends on different aspects of the data like format of the available input data, the purpose (target use case) of the transformation and the volatility of the data (how dynamic is the data). Based on these characteristics, there are two main approaches for making

Fig. 8.1 Generic flow for Linked Data integration and publication pipeline aligned with top-level generic pipeline

the transformation for a dataset: (i) data upgrade or lifting, which consists of generating RDF data from the source dataset according to mapping descriptions and then storing it in semantic triplestore (e.g., Virtuoso) and (ii) on-the-fly query transformation, which allows evaluating SPARQL queries over a virtual RDF dataset, by rewriting those queries into source query language according to the mapping descriptions. In this former scenario, data physically stays at their source and a new layer is provided to enable access to it over the virtual RDF dataset.

In every transformation process, regardless of the method or tools chosen, a mapping specification has to be defined to specify the rules to map the source elements (e.g., table columns, JSON elements, CSV columns, etc.) into target elements (e.g., ontology terms). Generally, this specification is an RDF document itself written in RML¹/R2RML² (and extensions) languages and/or nonstandard extensions of SPARQL, e.g., in the case of the Tarql CSV to RDF transformation tool.³

The resulting datasets can thereafter be exploited through SPARQL queries, or via a plethora of user interfaces. Some examples of these interfaces include:

- SPARQL endpoint interface, to execute queries: <https://www.foodie-cloud.org/sparql>
- Faceted search interface to navigate the linked datasets <http://www.foodie-cloud.org/fct/>
- Map visualization via HS Layer applications, e.g., <http://app.hslayers.org/project-databio/land/>

¹ <https://rml.io/specs/rml/>.

² <https://www.w3.org/TR/r2rml/>.

³ <https://tarql.github.io/>.

Fig. 8.2 Generic Linked Data Publication pipeline component view diagram

- Metaphactory: <http://metaphactory.foodie-cloud.org/resource/Start>.

The following diagram (Fig. 8.2) provides a simplified representation of the generic Linked Data Publication pipeline component view that includes the software components and interfaces involved. More information is available in [5, 6].

The URL link of the generic pipeline in the DataBioHub is <https://mub.me/2f81>.

8.2 Linked Data Pipeline Instantiations in DataBio

The Linked Data Pipeline, as described in the previous section, is a generalization of multiple instantiations, in particular two specific project's pilots and four additional experiments in DataBio. Thus, in order to show how this generic pipeline has been applied in each of these use cases, we present in this section for each of them the pipeline view, previously presented in [5], highlighting the specific methods and components used/applied, along with a description of the task performed and results achieved.

8.2.1 *Linked Data in Agriculture Related to Cereals and Biomass Crops*

This pipeline instance was focused toward publication of INSPIRE-based agricultural Linked Data from the farm data collected from cereals and biomass crop pilots, in

order to query and access different heterogeneous data sources via an integrated layer. The input datasets used for this experiment include:

- Farm data (Rostenice pilot) that holds information about each field name with the associated cereal crop classifications and arranged by year.
- Data about the field boundaries and crop map yield the potential of most of the fields in the Rostenice pilot farm from Czech Republic.
- Yield records from two fields (Pivovarska and Predni) within the pilot farm that were harvested in 2017 and 2018.

The source datasets, collected as shapefiles, were transformed into RDF format and published as Linked Data, using the FOODIE ontology as the underlying model. The resulting linked datasets are available for querying and exploitation through the DataBio SPARQL endpoint deployed at PSNC' HPC facilities. More in detail, the tasks carried out are as follows:

- Definition of the data model to transform the input datasets into RDF. For this step, FOODIE ontology [7], which is based on INSPIRE schema and the ISO 19100 series standards, was used as the base vocabulary and extended as needed (with a Czech pilot extension) in order to represent all the farm and open data from the input datasets. The extension includes data elements and relations from the input datasets that were not covered by the main FOODIE ontology but that were critical for the pilot needs.
- Creation of an RDF mapping file that specifies how to map the contents of a dataset into RDF triples by matching the source dataset schema with FOODIE ontology and its extensions. A generic RML/R2RML definition of the mapping file was generated from the input shapefiles by using applications like GeoTriples and thereafter manually edited as per the data model identified to generate the final mapping definition. GeoTriples was also used to generate the RDF dump from the source data contents. FOODIE ontology and its extension were used extensively in the mapping files to match the source dataset schemas.
- The RDF datasets generated were loaded into DataBio Virtuoso triplestore. A SPARQL endpoint and a faceted search endpoint are available for querying and exploiting the Linked Data in the Virtuoso instance deployed at PSNC infrastructure.
- The final task involved providing an integrated view over the original dataset. As source datasets were particularly large (especially when considering connections with open datasets), and the connections were not of equivalence (i.e., resources are related via some properties but they are not equivalent), it was decided to use queries to access the integrated data as per need rather than using link discovery tools like SILK or LIMES. Hence, cross-querying within the datasets was done in Virtuoso SPARQL endpoint for some use cases to establish possible links between agricultural and related open datasets.
- To visualize and explore the Linked Data in a map, we have created different application/system prototypes. One such map visualization component called HS Layers NG is available at <https://app.hsayers.org/project-databio/land/>.

includes well-known RDF vocabularies (SKOS,⁵ SCOVO,⁶ VOID, FOAF,⁷ Dublin Core⁸).

The SensLog service uses a relational database (PostgreSQL) to store the data. Hence, in the mapping stage, the creation of R2RML/RML definitions required different preprocessing tasks and some on-the-fly assumptions to engineer the alignment between the SensLog database and the ontologies/vocabularies.

Once the mapping file was generated (manually), the RDF data of the dataset was published using a D2RQ server that enables accessing relational database sources as virtual RDF graphs. This on-the-fly approach allows publishing of RDF data from large and/or live databases, and thus the need for replicating the data into a dedicated RDF triple store is not required. The Linked Data from the sensor data from SensLog (version 1) was published in the PSNC infrastructure in a D2RQ server available at <http://senslogrdf.foodie-cloud.org/>. The associated SPARQL endpoint to query the data is available at: <http://senslogrdf.foodie-cloud.org/sparql>.

The figure below (Fig. 8.4) highlights the main components used in this pilot from the generic pipeline components.

Fig. 8.4 Mapping of the generic components into machinery management pilot in the pipeline view

⁵ <https://www.w3.org/TR/skos-reference/>.

⁶ <http://vocab.deri.ie/scovo>.

⁷ <http://www.foaf-project.org/>.

⁸ <https://www.dublincore.org/specifications/dublin-core/dces/>.

8.2.3 *Linked Open EU-Datasets Related to Agriculture and Other Bio Sectors*

This pipeline focuses on EU and national open data from various heterogeneous sources from a wide range of applications in the geospatial domain. The purpose was to experiment on these datasets by transforming them into Linked Data and exploiting them on various technology platforms for integration and visualization. The sources for all of these data contents are widely heterogeneous and in various forms (e.g., in shapefiles, CSV format, JSON and in relational databases), which required extensive work to identify the most suitable mode for their transformation. This included a careful inspection of the input data contents in order to identify available ontologies/vocabularies, and any required extensions, necessary for the representation of such data in RDF format. Additionally, since the source datasets were in different formats, selecting the most suitable tools for their transformation was a key activity in order to create the correct (R2RML/RML) mapping definitions. Some of the input datasets, their formats and the ontologies/vocabularies used for the representation of data in semantic format are described below.

- Input data of land parcel and cadastral data (for Czech Republic and Poland), erosion-endangered soil zones, water buffer and soil type classification are available as shapefiles. The ontologies used for the representation of such data included the INSPIRE-based FOODIE ontology as well as different extensions created to cover all the necessary information (e.g., erosion zones and restricted areas near water bodies).
- The Farm Accountancy Data Network (FADN⁹) data is available as a set of CSV files. The main ontologies used were Data Cube Vocabulary and its SDMX ISO standard extensions that were much more effective in aligning such multidimensional survey data. Data Cube Vocabulary encompasses well-known RDF vocabulary like SKOS, SCOVO, VOID, FOAF, Dublin Core, etc. Preparing the mapping definitions from the input data sources required preprocessing actions to make them reusable for all types of the CSV data sources of FADN. Separate CSV files were manually created for each reusable common class type. Once mapping definitions were generated for each of the created CSV files, they were integrated into one whole mapping file covering all the components from the input data.
- The sample data input from Yelp is available as a set of JSON files. Different ontologies like review,¹⁰ FOAF, schema.org, POI, etc., were used to represent the elements from the input data in semantic format during the creation of the mapping definition.
- Other ontologies from previous efforts for the representation of open geospatial datasets like Corine, Hilucs, OLU, OTM, Urban Atlas, were also used. These ontologies are available in <https://github.com/FOODIE-cloud/ontology>.

⁹ <https://ec.europa.eu/agriculture/rica/>.

¹⁰ <https://vocab.org/review/>.

INSPIRE standards, and it allows to retrieve the metadata in RDF using Geo-DCAT¹² for the representation of geographic metadata compliant with the DCAT application profile for European data portals. Nevertheless, such metadata cannot be queried as Linked Data, and thus the goal was to make it available in this form in order to enable its integration with other datasets, e.g., Open Land Use (OLU). The process for publication, thus, was straightforward: A dump of all the metadata in RDF format was generated from Micka, which was then loaded into DataBio Virtuoso triplestore. Some example SPARQL queries were then generated to identify connection points for integration, e.g., get OLU entries and their metadata given a municipal code and type of area (e.g., agriculture lands). The dataset is accessible via: <https://www.foodie-cloud.org/sparql>.

The second dataset was more challenging. The goal was to make Earth Observation (EO) Collections and EO Products metadata available as Linked Data via a SPARQL compliant endpoint which makes requests to non-SPARQL back ends on the fly. Hence, we wanted to enable querying via SPARQL without harvesting all the metadata and storing the data in a triplestore but access them dynamically via the existing online interfaces. The metadata was accessible via an OpenSearch interface provided by the FedEO¹³ Clearinghouse in Spacebel (<http://geo.spacebel.be/opensearch/readme.html>) that enables retrieving the metadata in different formats, including atom/xml, RDF/xml, turtle, GeoJSON and LD-JSON. We used LD-JSON, which already defines the semantic properties used to represent the metadata elements. These properties comprise terms from different standard and well-known vocabularies/ontologies like Dublin Core, DCAT, SKOS, VOID and OM-Lite-lite, as well as from the OpenSearch specifications. Next, in order to enable access to a REST API via SPARQL queries that would allow linking with other Linked Datasets we used the Metaphactory platform. Metaphactory (<https://www.metaphacts.com/product>) includes a component called Ephedra, which is a SPARQL federation engine aimed at processing hybrid queries. Ephedra provides a flexible declarative mechanism for including hybrid services into a SPARQL federation and implements a number of static and runtime query optimization techniques for improving the hybrid SPARQL query performance [8]. The RDF data is exposed via a SPARQL endpoint provided in the Metaphactory platform (<http://metaphactory.foodie-cloud.org/sparql?repository=ephedra>). A demo interface has also been implemented to visualize the Linked Data in Metaphactory (entry point: <http://metaphactory.foodie-cloud.org/resource/ESA-datasets>).

The figure below (Fig. 8.6) highlights the main components used in this pilot from the generic pipeline components. In the figure, the components related to the first sub-case (Micka) are highlighted in green, while the components related to the second sub-case (FedEO) are highlighted in orange.

¹² <https://ec.europa.eu/jrc/en/publication/geodcat-ap-representing-geographic-metadata-using-dcat-application-profile-data-portals-europe>.

¹³ <http://ceos.org/ourwork/workinggroups/wgiss/access/fedeo/>.

Fig. 8.7 Mapping of the components used in the fishery use case in the pipeline view

- At present, the catch data from five years was transformed and uploaded to the Virtuoso triplestore providing SPARQL and faceted search endpoints for further exploitation.

For the purpose of showcasing the integration and visualization of the dataset, a Web interface using the Metaphactory platform was created, which includes map visualizations and representation of data in the form of charts and graphs. This process is ongoing, and more experimentations are to come by. The interface is presently available at http://metaphactory.foodie-cloud.org/resource/:CatchDataNorway_v2. The resulting linked datasets are accessible via: <https://www.foodie-cloud.org/sparql> and <https://www.foodie-cloud.org/fct>.

The figure below (Fig. 8.7) highlights the main components used in this use case from the generic pipeline components.

8.3 Experiences from DataBio with Linked Data

8.3.1 Usage and Exploitation of Linked Data

The pipelines used in DataBio are part of an ongoing process and yet to be tested on other use cases and input data types. For example, as a result of the pipelines involving the LPIS and Czech field data, it was possible to perform integration experiments of the dataset for various use case scenarios of data integration.

As mentioned above, the datasets are deployed in the Virtuoso triplestore within PSNC and can be accessed via SPARQL and faceted search endpoints. The triplestore

has **over 1 billion triples**, making it one of the largest semantic repositories related to agriculture.

The data in the triplestore is partitioned/organized into *named graphs*, where each named graph describes different contents and is identified by an IRI.

For example, the IRI <<http://w3id.org/foodie/open/africa/GRIP>> is the graph identifier of the African Roads Network dataset, which contains 27,586,675 triples.

Named graphs may be further composed of named subgraphs, as it is the case of the LPIS Poland dataset, which provides information about land-parcel identification in Poland, identified by the graph <<http://w3id.org/foodie/open/pl/LPIS/>>, and contains 727,517,039 triples. This graph contains, for example, the subgraph <<http://w3id.org/foodie/open/pl/LPIS/lubelskie>>, which refers to the data associated with the Lublin Voivodeship.

The table below shows some of the respective graphs produced by all the pipelines previously described and the number of triples contained in them.

Graph URI (note: URIs are not resolvable; they can be used to refer to the specific dataset in the triplestore)	Name of dataset	Number of RDF triples
http://w3id.org/foodie/open/pl/LPIS/{voivodeship} (where voivodeship in Poland = mazowieckie, dolnoslaskie, kujawsko-pomorskie, lodzkie, lubelskie, lubuskie, malopolskie, opolskie, podkarpackie, podlaskie, pomorskie, slaskie, warminsko-mazurskie, wielkopolskie, zachodniopomorskie, swietokrzyskie)	LPIS Poland	727517039
http://w3id.org/foodie/olu agriculture-related lands (hilucs_code < 200) in CZ, PL, ES and for main cities in Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas)	Open land use	127926060
http://w3id.org/foodie/otm CZ, ES, PL; but RoadLinks only for FunctionalRoadClassValue of type: ('mainRoad,' 'firstClass,' 'secondClass,' 'thirdClass,' 'fourthClass') (see http://opentransportmap.info/OSMtoOTM.html)	Open transport map	154340785

(continued)

(continued)

Graph URI (note: URIs are not resolvable; they can be used to refer to the specific dataset in the triplestore)	Name of dataset	Number of RDF triples
http://micka.lesprojekt.cz/catalog/dataset	Open land use metadata	10456676
http://www.sdi4apps.eu/poi.rdf	Smart points of interest (SPOI)	407629170
http://w3id.org/foodie/open/cz/pLPIS_180616_WGS	LPIS Czech Republic	24491282
http://w3id.org/foodie/open/cz/lpis/code/LandUseClassificationValue	LPIS Czech Republic land use classification	83
http://w3id.org/foodie/atlas agriculture-related lands (hilucs_code < 200) and for main cities in Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas)	Urban Atlas	19606088
http://w3id.org/foodie/corine agriculture-related lands (hilucs_code < 200) and for main cities in Czech Republic (centers of NUTS3 regions), Poland (agglomeration areas from Urban Atlas) and Spain (agglomeration areas from Urban Atlas)	Corine land use	16777595
http://w3id.org/foodie/open/cz/Soil_maps_BPEJ_WGSc	Czech soil maps	8746240
http://w3id.org/foodie/open/cz/water_buffer25	Czech water buffers	3978517
http://w3id.org/foodie/core/cz/Pre dni_prostredni_vyfiltrovano_UTM	Yield mass in field crops (CZ Pilot)	1111852
http://w3id.org/foodie/core/cz/Pivovarka_vyfiltrovano	Yield mass in field crops (CZ Pilot)	437404
http://w3id.org/foodie/core/cz/CZpilot_fields	CZ Pilot fields and crop data	20183

(continued)

(continued)

Graph URI (note: URIs are not resolvable; they can be used to refer to the specific dataset in the triplestore)	Name of dataset	Number of RDF triples
http://ec.europa.eu/agriculture/FADN/{FADNcategory} (Where FADN category = year-country, year-country-anc3, year-country-lfa, year-country-organic-tf8, year-country-siz6, year-country-siz6-tf14, year-country-siz6-tf8, year-country-sizc, year-country-tf14, year-country-tf8m, year-country-typology, year-region, year-region-siz6, year-region-siz6-tf8, year-region-sizc, year-region-tf14, year-region-tf8)	FADN	23520756
http://w3id.org/foodie/open/africa/GRIP	African roads network	27586675
http://w3id.org/foodie/open/africa/water_body	African water bodies	11330
http://w3id.org/foodie/open/gadm36/{level} where {level} = level0, level1, level2, level3, level4, level5	GADM dataset	7188715
http://w3id.org/foodie/open/kenya/ke_crops_size	Kenya crop size	85971
http://w3id.org/foodie/open/kenya/soil_maps	Kenya Soil Maps	10168
http://www.fao.org/aims/aos/fi/taxonomic	FAO	318359
http://www.fao.org/aims/aos/fi/water_FAO_areas	FAO	150
http://www.fao.org/aims/aos/fi/water_FAO_areas/inland	FAO	15779
http://www.fao.org/aims/aos/fi/water_FAO_areas/marine	FAO	6768
http://w3id.org/foodie/open/catchrecord/norway/	Catch record norway	192867166
http://standardgraphs.ices.dk/stocks	ICES stocks data	1270280

(continued)

(continued)

Graph URI (note: URIs are not resolvable; they can be used to refer to the specific dataset in the triplestore)	Name of dataset	Number of RDF triples
https://www.omg.org/spec/LCC/Countries/ISO3166-1-CountryCodes/	ISO country codes	8629
https://www.omg.org/spec/LCC/Countries/Regions/ISO3166-2-SubdivisionCodes-NO/	ISO country subdivision codes	391
https://www.omg.org/spec/LCC/Countries/UN-M49-RegionCodes/	ISO region codes	569

The official SPARQL and the faceted search endpoints of the triplestore are: <https://www.foodie-cloud.org/sparql> (Fig. 8.8) and <https://www.foodie-cloud.org/fct>

Fig. 8.8 SPARQL endpoint user interface (query and extract of result)

(Fig. 8.9).

Regarding the sensor data described in Sect. 1.3.2, it is published on the fly which serves the purpose of streaming transformation. This data can be accessed and linked through the following endpoints:

- SPARQL endpoint: <http://senslogrdf.foodie-cloud.org/sparql>
- SNORQL search endpoint: <http://senslogrdf.foodie-cloud.org/snorql/>
- Web-based visualization: <http://senslogrdf.foodie-cloud.org/> (see Fig. 8.10).

8.3.2 Experiences in the Agricultural Domain

RDF links often connect entities from two different sources, with relations which are not necessarily described in either data source. In the agricultural domain, this

Fig. 8.9 Faceted search interface

Fig. 8.10 Web interface entry page and visualization of an observation details of RDF generated on the fly

can be linking fields of specific crop type with the administrative region in which these fields reside, or find whether plots intersect with a buffer zone of water bodies in their vicinity. This is a means to control, e.g., the level and amount of pesticides used in those plots.

Creating such agricultural *knowledge graphs* is important due to environmental, economic and administrative reasons. However, constructing links manually is time and effort intensive, and links between concepts are rather to be *discovered* automatically. The basic idea of link discovery is to find data items within the target dataset

which are logically connected to the source dataset. Formally, this means: Given $\mathcal{S}$ and $\mathcal{T}$, sets of RDF resources, called source and target resources, respectively, and a relation R , the aim of link discovery methods is to find a mapping $M = \{(s,t) \in \mathcal{S} \times \mathcal{T} : R(s,t)\}$. Naive computation of M requires quadratic time to test for each $s \in \mathcal{S}$ and $t \in \mathcal{T}$ whether R holds, which is infeasible for large datasets, and leads to the development of link discovery tools, which address this task.

In the agricultural domain, entities are mostly geospatial objects, and the relations are of a topological nature. Existing tools for link discovery, such as SILK and LINES, are limited when it comes to geospatial data and therefore, as part of the DataBio project, we developed Geo-L, a system designated for discovery of RDF spatial links based on topological relations.

The system provides flexible configuration options to define to-be-linked datasets for SPARQL affine users and employs retrieval and caching mechanisms, resulting in efficient dataset management.

Geo-L uses PostgreSQL, an open-source object-relational DBMS, with PostGIS extension, as the database back end which supports geospatial data processing.

We conducted experiments to evaluate the performance of our proposed system by searching geospatial links based on topological relations between geometries of datasets of the foodie cloud, in particular subsets of OLU, SPOI and NUTS.

The experiments show that Geo-L outperforms the state-of-the-art tools in terms of mapping time, accuracy and flexibility.¹⁵ It also proves to be more robust when it comes to handling errors in the data, as well as with managing large datasets.

We applied Geo-L to several use cases involving datasets from the foodie cloud, e.g.,

- Identifying fields from *Czech LPIS* data with specific soil type, from *Czech open data*
- Identifying all fields in a specific region which grow the same type of crops like the one grown in a specific field over a given period of time
- Identifying plots from *Czech LPIS* data which intersect with *buffer zones* around water bodies.

¹⁵ As shown in a case of searching topological relations with NUTS as target dataset, where geometries are not represented as expected and were transformed on the fly to polygons by our tool.

An example for the last case is depicted in the image below (Fig. 8.11), where an overlap area between a plot and a buffer zone of a water body in its vicinity is colored with orange.

The respective dataset resulting from linking water bodies whose buffer zones are intersected by Czech LPIS plots is available on the DBpedia Databus.¹⁶

8.3.3 Experiences with DBpedia

DBpedia is a crowd-sourced continuous community effort to extract structured information from Wikipedia and to make this information available as a knowledge graph on the Web. DBpedia allows querying against this data and information and linking to other datasets on the Web [9, 10]. Currently, DBpedia is one of the central interlinking hubs in the Linked Open Data (LOD) cloud. With over 28 million of described and localized things, it is one of the largest and open datasets.

As part of the project, we constructed links between satellite entities, available in the European Space Association (ESA) thesaurus,¹⁷ whose recorded images are employed in DataBio pilots and their respective DBpedia resources. These links are beneficial since the data in DBpedia is available in machine readable form for further processing, and in addition there are additional data and external links related to the satellite. We used REST API to retrieve satellite names from the ESA thesaurus and queried for DBpedia resources matching these names, which were then identified as satellites, based on their properties available in DBpedia.

Fig. 8.11 Overlap area between a plot and a buffer zone of a water body in its vicinity, colored with orange

¹⁶ <https://databus.dbpedia.org/amt/geo-L/bufferzones-intersect-lpisPlots/>.

¹⁷ <https://fedeo.spacebel.be/thesaurus/fr/>.

```

<http://dbpedia.org/resource/Aura_(satellite)> <http://www.w3.org/2002/07/owl#sameAs> <https://earth.esa.int/concept/aura> .
<http://dbpedia.org/resource/CASSIOPE> <http://www.w3.org/2002/07/owl#sameAs> <https://earth.esa.int/concept/cassiope> .
<http://dbpedia.org/resource/CryoSat-1> <http://www.w3.org/2002/07/owl#sameAs> <https://earth.esa.int/concept/cryosat> .
<http://dbpedia.org/resource/EarthCARE> <http://www.w3.org/2002/07/owl#sameAs> <https://earth.esa.int/concept/earthcare> .
<http://dbpedia.org/resource/Envisat> <http://www.w3.org/2002/07/owl#sameAs> <https://earth.esa.int/concept/envisat> .
<http://dbpedia.org/resource/Gravity_Field_and_Steady-State_Ocean_Circulation_Explorer> <http://www.w3.org/2002/07/owl#sameAs> <https://earth.esa.int/concept/goce> .

```

Fig. 8.12 Links between ESA-platforms and their respective entities in DBpedia

The listing depicted in Fig. 8.12 presents an excerpt from the link-data result. The links allow, on the one hand, access to other properties of the respective DBpedia resources and, on the other hand, enable other DBpedia users to access the ESA set. This dataset can be found as an artifact¹⁸ on the DBpedia Databus.

DBpedia resources which refer to geographical regions include different important properties about those areas such as temperature amplitudes and monthly precipitation. Such properties may be helpful, e.g., analysis of yields. These resources, however, do not contain the actual geometry of the regions. We used OpenStreetMap to retrieve data about regions and applied Geo-L to link between DBpedia region resources and their geometries.

These geometries can be helpful then not only for the purpose of the DataBio or for agriculture in general, but may be used for locating points of interest, which coordinates are known, within a specific region, a thing which has not been possible so far.

References

1. Wood, D., Lanthaler, M., Cyganiak, R. (2014). *RDF 1.1 Concepts and Abstract Syntax* [W3C Recommendation]. (Technical report, W3C).
2. Harris, S., Seaborne, A. (2013). *SPARQL 1.1 Query Language*. W3C Recommendation. W3C.
3. Hyland, B., Ateamezing, G., Villazón-Terrazas, B. (2014). *Best Practices for Publishing Linked Data*. W3C Working Group Note 09 January 2014. URL: <https://www.w3.org/TR/ld-bp/>.
4. Heath, T., Bizer, C. (2011). *Linked data: Evolving the web into a global data space* (1st Ed.). Synthesis lectures on the semantic web: Theory and technology, 1:1, 1–136. Morgan & Claypool.
5. DATABIO. (2018). D4.i3—Full Platform Description for Trial 2.
6. DATABIO. (2019). D4.4—Technologies for supporting Bio-economy pilots.
7. Palma, R., Reznik, T., Esbri, M., Charvat, K., Mazurek, C. (2015). An INSPIRE-based vocabulary for the publication of agricultural linked data. Proceedings of the OWLED Workshop: Collocated with the ISWC-2015, Bethlehem PA, USA, October 11–15.
8. Nikolov, A. et al. “Ephedra: SPARQL federation over RDF data and services.” International Semantic Web Conference (2017). URL: <https://www.metaphacts.com/images/PDFs/publications/ISWC2017-Ephedra-SPARQL-federation-over-RDF-data-and-services.pdf>.
9. Auer et al. (2007). “DBpedia: A nucleus for a web of open data” https://doi.org/10.1007/978-3-540-76298-0_52.
10. Lehmann et al. (2013). “Integrating NLP using linked data” https://doi.org/10.1007/978-3-642-41338-4_7.

¹⁸ <https://databus.dbpedia.org/amit/esa/links/>.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.