Ontology Based Data Access and Integration for Improving the Effectiveness of Farming in Nepal

Suresh Pokharel*, Mohamed Ahmed Sherif*, Jens Lehmann*
*Universität Leipzig, Institut für Informatik,
AKSW,
{pokharel, sherif, lehmann}@informatik.uni-leipzig.de

***Abstract*—It is widely accepted that food supply and quality are major problems in the 21st century. Due to the growth of the world's population, there is a pressing need to improve the productivity of agricultural crops, which hinges on different factors such as geographical location, soil type, weather condition and particular attributes of the crops to plant. In many regions of the world, information about those factors is not readily accessible and dispersed across a multitude of different sources. One of those regions is Nepal, in which the lack of access to this knowledge poses a significant burden for agricultural planning and decision making. Making such knowledge more accessible can boot up a farmer's living standard and increase their competitiveness on national and global markets. In this article, we show how we converted several available, although not easily accessible, datasets to RDF, thereby lowering the barrier for data re-usage and integration. We describe the conversion, linking, and publication process as well as use cases, which can be implemented using the farming datasets in Nepal.**

I. Introduction

Information and communication technologies (ICT) have gained significant importance in our lives across several domains. Agriculture is no exception and the coining of the term *E-agriculture* roots back to the rather recent *World Summit of the Information Society* in 2003[1]. The key characteristics of E-agriculture are the dissemination, access and exchange of information. ICT can play a vital role to boot up a farmers' living standard by providing relevant information. Nevertheless, in Nepal (a country with agricultural based economy) information such as crop geographical location, properties of soil, climate information and crop production normally are not publicly available. It is difficult for farmers to obtain access to such information and, therefore, they cannot benefit from for planning and decision making.

In the agriculture domain, various aspects have to be integrated to build a fully functioning system with all the information related to agriculture such as weather measurements, soil characteristics, new research results and findings, government policies, market information and inventory. All of such different data are produced by different bodies of the government and all of these departments are working rather independently with limited integration between them.

Taking the rice crop as an example, not only irrigation alone can improve its productivity, there are other factors[2] such as soil status, weather conditions and rice water requirements during each of its sub-seasons. Due to the lack of integration between such heterogeneous data, extraction of information like how much irrigation is required for rice in a particular region on a particular day is difficult to obtain, which in turn leads to reduced farming efficiency.

Recently, many different agriculture related projects were established in Nepal, in particular by the *Ministry of Irrigation*[3] and the *Ministry of Agriculture Development*[4]. For instance, the *Ground Water Irrigation Project*[5] was launched improve the rice productivity in *Chitwan* district, Nepal. While those initiatives provide relevant information, it is not published using established standards. For this reason, we convert information to Linked Open Data (LOD)[1], [2] using the RDF data model and established vocabularies. This allows not only to publish data conforming to W3C standards, but also to establish links between data sources, thereby enabling analysis methods going beyond those possible when using the original data sources in isolation.

In our work, we can draw on existing ontologies. In particular, *AGROVOC*[6] is a controlled RDF vocabulary with around 32.000 concepts covering all of the *Food and Agriculture Organization of the United Nations* (FAO) areas of interest including food, nutrition, agriculture, fisheries, forestry and environment. AGROVOC thesaurus is already mapped to many ontologies such as the *FAO Biotechnology Glossary*, *EUROVOC*, *GEMET*, *Library of Congress Subject Headings* (LCSH), *NAL Thesaurus*, *Thesaurus for Economics* (STW), *Thesaurus for the Social Sciences* (TheSoz), *Geopolitical ontology*, *Dewey Decimal Classification* (DDC), *DBpedia* [3] and *GeoNames*.

The data management efforts performed are the first steps on a larger research agenda, which we publish in the context of *ArgiNepalData* project[7]. In general, this article presents an application of web intelligence methods. A major contribution is the conversion and integration of data from five different sources (cf. Subsection III-A). In addition to providing the farming datasets as RDF, we designed an ontology for representing and aligning those heterogeneous datasets. This

[1]http://www.e-agriculture.org/e-agriculture
[2]http://cals.arizona.edu/pubs/water/az1220/
[3]http://www.doi.gov.np
[4]http://www.doanepal.gov.np
[5]http://www.doi.gov.np/projects/project.php?pid=25
[6]http://aims.fao.org/standards/agrovoc/linked-open-data
[7]http://agrinepaldata.com

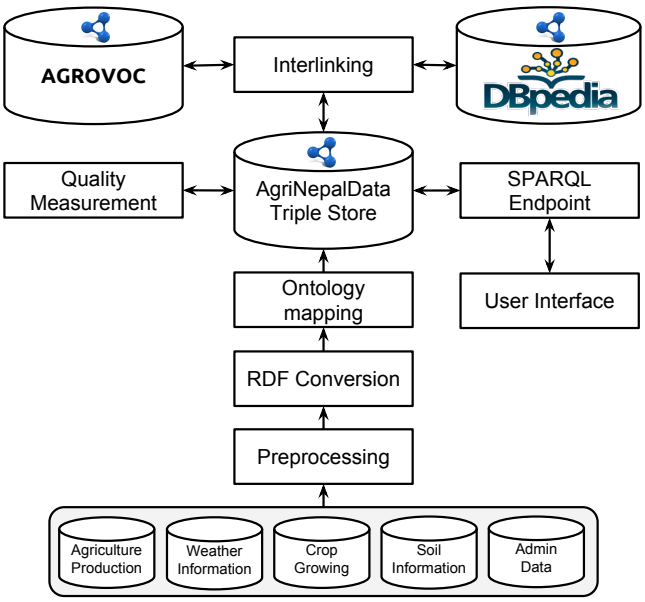

Fig. 1. AgriNepalData Data Management Framework

alignment enables inference of new knowledge from converted data. Moreover, we linked the dataset to *DBpedia* as well as *AGROVOC* and ensured therewith that our dataset abides by all Linked Data principles[8].

The remainder of this application paper is structured as follows: In the subsequent section, we present a detailed description of the the framework used in our data conversion. Then, in Section III, we describe each of the data sources used in our datasets. Moreover, we give an overview of the ontology that forms the background structure of our datasets (Section IV). We present the approach used to link the farming datasets in Nepal with different external datasets in Section V. Based on this, we present several usage scenarios for the datasets at hand (Section VII). Finally, we described related work and give concluding remarks.

II. Methodology

In order to generate the *AgriNepalData* datasets, we have adapted a data management framework (see Figure 1).

For the data management in *AgriNepalData*, we use the Linked Data Life-cycle vision as a base [4][9] (see Figure 2). Below, we discuss each of the 8 lifecycle phases in the context of *AgriNepalData*:

- **Extraction:** The first step is the extraction of RDF from *comma-separated values* (CSV), HTML and shape files. We have used *OpenRefine, TripleGeo and Sparqlify* tools for this process. Detailed descriptions are given in Subsection III-B.
- **Storage and Querying:** For hosting the *agriNepalData* we need a triple store which can handle not only (1) different data types such as strings, numbers, dates and spatial point sets, but also (2) continuously growing size as more datasets converted and added. In order to fulfill the aforementioned requests we chose *Virtuoso*[10]. *Virtuoso* provides backward chaining OWL reasoning, geospatial/text indexing and query functionality through SPARQL endpoint. For querying the *AgriNepalData* user can use the provided endpoint[11].
- **Manual revision and Authoring:** In order to minimize the error rate in the converted data we apply manual test cases. In some cases the manual testing led us to discover some discrepancies either in our conversion framework or in the data itself. In former case we refine our framework and in the later case we refined the data preprocessing phase. For example, when applying a manual test case for checking for the RDF district data, we notice missing some data for fruit production of the *Kabhrepalanchok* District. We surprisingly found this district to have two different names in agriculture production dataset; *Kavre* for wool production (which was already considered as the name used by this dataset to this district) and *Kavrepalanchok* for fruit production (which is missing). Therefore, in the preprocessing phase, we added the *Kavrepalanchok* name to the list of synonyms of *KabhrepalanchokDistrict*.
- **Interlinking:** The AgriNepalData datasets are interlinked with both *DBpedia* and *AGROVOC* datasets using LIMES (for more details see Section V).
- **Classification and Enrichment:** In this phase, we applied the *GeoLift*[12] tool to enrich our dataset with additional geospatial data from other datasets like *DBpedia* and *LinkedGeoData*.
- **Quality and Analysis:** As any dataset is as good as its quality, we applied *RDFUnit*[5] tool to measure the quality of data as well as set of manual verifications (see Section VI for more details).
- **Evolution and Repair:** After applying the manual test cases and the automated data quality tools we discovered a set of discrepancies, which we needed to repair. Once we repaired the discovered errors we re-ran the manual test cases as well as the automatic tools to increase the quality of the datasets.
- **Search and Browsing:**
The *Facete (Faceted Browser)*[6] tool is used to provide a visual searching and browsing interface. More detailed descriptions are given in Subsection VII-C.

III. Dataset Description

In this section, we first describe each of raw sources in detail. Since we obtain the source data from different sources, we cannot expect them to be homogeneous, which leads to challenges in the RDF conversion process. We illustrate the data conversion and those challenges in the second subsection.

[8]http://www.w3.org/DesignIssues/LinkedData.html
[9]http://stack.linkeddata.org/

[10]http://virtuoso.openlinksw.com/rdf-quad-store/
[11]http://agrinepaldata.com/sparql
[12]http://aksw.org/Projects/GeoLift.html

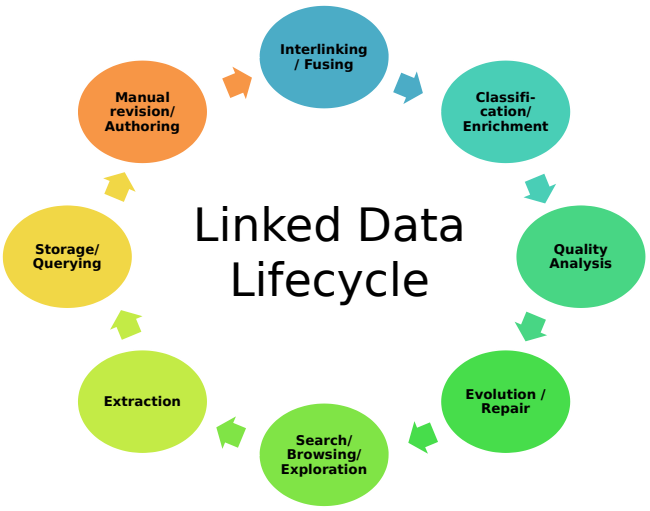

Fig. 2. Linked Data Lifecycle

A. Data sources

The raw datasets have been collected from five semi-structured sources

1) **Agriculture Production Statistics of Nepal:** This dataset is collected from the *Ministry of Agricultural Development*[13] in Nepal. It contains the information about yearly production of different crops since 1990 to 2012. Furthermore, it provides details of the production of crops and livestocks in each of Nepal's districts in 2011/12. The raw dataset is freely available online in PDF format and as CSV file on request to ministry.
2) **Weather Information:** This dataset is collected from the *Department Of Hydrology and Meteorology*[14] in Nepal. It contains information about daily rainfall of stations for *Babai* from 1980 to 2008 and *West Rapti* from 1980 to 2006. Additionally, it includes hourly weather information of *Banepa* from 2011 to 2012. The raw dataset is freely available online in HTML format and in CSV file format on request.
3) **Crop Growing Days Information:** This dataset is collected from FAO website[15]. It contains the information about crop growing days in each stage as well as crop coefficients in each stages. The raw dataset is freely available as HTML file.
4) **Soil Information Of Nepal:** The original dataset was called *SOTER_Nepal*, which is collected from the *ISRIC - World Soil Information*[16]. The SOTER_Nepal database provides generalized information on landform and soil properties at a scale 1:1 million. It consists of 17 SOTER units and characterized by 56 representative and four synthetic profiles for which there are no measured soil data. The raw dataset was in form of shape files. ISRIC encourages the provision and use of all its data for research, education and policy support.
5) **Administrative Data of Nepal :** This dataset is collected from the *International Centre for Integrated Mountain Development (ICIMOD)*[17], Nepal. This is based on topographic zonal map published by department of survey in different dates, which are more then 20 thematic layers covering entire Nepal. It contains the information about each of Nepal's development regions, zones, districts, villages development committees (VDCs), wards, national parks, peaks and roads. The raw dataset was in form of shape files. ICIMOD offers free access for its data for free registered users.

B. Extraction Process

The data was extracted from various sources using three different formats (CSV, Shapefiles, HTML) for each of which we describe the conversion process below.

1) **CSV to RDF Conversion:** The crop statistics, weather information and crop growing days datasets are available in CSV format, but do not have any uniform structure beyond using the same format. First, we did some preprocessing such as removing special characters, unifying measurement units and filling missing data. Afterwards, The CSV files were converted to RDF using *OpenRefine*[18] and *Sparqlify*[7].
Listing 1 shows an example of converting one row of data of a CSV file to RDF format. Originally, the raw CSV data row was: `PaddyTaplejung2011 Taplejung 10477 22167 2116 Paddy 2011/12`, which shows statistics about the *Paddy* produced in from August 2011 to July 2012 in *Taplejung* district.

```
agrd:PaddyTaplejung2011
   a agro:CerealCropProduction, time:TemporalEntity ;
   agro:inDistrict agrd:TaplejungDistrict ;
   agro:produce agrd:Paddy ;
   agro:production "22167"^^dbo:tonne ;
   agro:yield "2116"^^dbo:perHectare ;
   quty:area "10477"^^dbo:hectare ;
   time:hasBeginning "2011-08-01"^^xsd:date ;
   time:hasEnd "2012-07-31"^^xsd:date .
```

Listing 1. RDF Conversion for *Paddy* produced in year 2011/12 in *Taplejung* district

2) **Shape to RDF Conversion:** The original datasets of soil and administrative information are stored as shape files. Shape files hold spatial data information in form of polygon or points as well as some non-spatial information. The spatial information of the shape files are converted to RDF using *TripleGeo*[8], while the non-spatial information of shape files is first extracted to CSV by using *QGIS*[19] and then converted to RDF using *OpenRefine*.
Listing 2 shows an example of the conversion of the information contained in an *ESRI* shape file. The example

[13] http://www.moad.gov.np/
[14] http://www.dhm.gov.np/
[15] http://www.fao.org/docrep/s2022e/s2022e00.htm
[16] http://www.isric.org/data/soil-and-terrain-database-nepal
[17] http://geoportal.icimod.org/downloads/
[18] http://openrefine.org/
[19] http://www.qgis.org/en/site/

TABLE I
AgriNepalData TRIPLES DETAILS.

Source dataset	# Triples	# Subjects
Agriculture Production Statistics of Nepal	27623	2887
Weather Information	404808	42003
Crop Growing Days Information	1030	125
Soil Information Of Nepal	21666	942
Administrative Data of Nepal	978288	281302
Ontology Related	216	216
Total	1433631	327475

shows the conversion of information on the district of *Gorkha* consisting of both spatial (polygon information in WKT format) as well as non spatial (name, area, region, zone, dcode) facts.

```
agrd:GorkhaDistrict
   a agro:District ;
   rdfs:label "Gorkha district"@en ;
   agro:dcode "36" ;
   agro:hasPart agrd:Gandaki ;
   agro:inDistrict agrd:GorkhaDistrict ;
   agro:inZone agrd:Gandaki ;
   agro:region agrd:Hill ;
   quty:area "3645.866"^^dbo:squareKilometre ;
   gsp:hasGeometry agrd:Geom_polygon_GorkhaDistrict .

agrd:Geom_polygon_GorkhaDistrict
   a opgis:Polygon ;
   gsp:asWKT "POLYGON ((85.10174531999999
        28.456713989999997, 85.10162976
        28.454921459999998...))"^^gsp:wktLiteral .
```

Listing 2. Example of spatial and non-spatial RDF conversion of information for *Gorkha* district from an ESRI shapefile.

3) **HTML to RDF Conversion:** The crop growing days Information raw data was in form of HTML files. First we apply some manual selection of interesting pieces of data, which in most cases was in form of tables. Afterworlds, in a manner akin to the one used to convert CSV to RDF, the manually selected tables are converted to RDF using open *OpenRefine*.

After converting all datasets, the resulting RDF files contain more than 1.4 million triples with 327475 distinct subjects. Table I shows the number of triples as well as distinct subjects of the RDF conversion for each of the aforementioned datasets. Table II provides technical details about *AgriNepalData* including the project website, dataset/ontology dumps, and SPARQL endpoint. Also, it includes version and license information.

IV. ONTOLOGY

To integrate the data on schema level, we developed an extensible ontology vocabulary for our dataset. The vocabulary[20] was specified with the aim of supporting any dataset dealing with agricultural aspects. Currently, our ontology includes 38 classes (see Figure 3) covering production, geography and weather aspects. More additional classes can be added to the ontology at hand to cover more aspects if necessary.

[20]http://agrinepal.com/vocab/, also available in ecosystem of LOV http://lov.okfn.org/dataset/lov/index.html

TABLE II
TECHNICAL DETAILS OF THE *AgriNepalData* DATASET.

Dataset name	AgriNepalData
Project website	http://agrinepaldata.com http://aksw.org/Projects/AgriNepalData
SPARQL endpoint	http://agrinepaldata.com/sparql
Dataset dump	http://agrinepaldata.com/download/agrinepal.zip
Ontology	http://agrinepaldata.com/download/agrinepaldataont.owl
Version date	15-03-2014
Version number	1.0
Licensing	(CC BY-NC-SA 3.0)
VoiD file	http://agrinepaldata.com/download/void.ttl
TheDataHub entry	AgriNepalData

The *Production* class is the super class for all other sub-production classes. Currently, there are eight sub-classes of the *Production* class covering different productions types found so far in our dataset. Naturally, extending this part of our ontology is straightforward by adding more *Production* sub-classes for more production types. Each of the production sub-classes contains properies to keep track of its product date, quantity and location (for more details, see the right part of Figure 3). For example, *mandarin* is an instance of class *Fruit*, which its production is handled by the *FruitProduction* class, which is a sub-class of *Production* class. These are modelled in OWL as standard local range restrictions using universal quantifiers.

To keep track of various geographical information, starting from *Country* class our ontology contains a chain of derived classes to represent the hierarchical structure of the administrative regions of the country (Nepal in our case). The *Country* class and all of its sub-classes represent geographical information using Well-known text (WKT) datatypes. Also, this part of the ontology is extensible through inheritance of new classes (for more details, see the left part of Figure 3). For example, *Goldhunga 1* ward is part of the VDC of *Goldhunga*, which is a part of the district of *Kathamandu*, which is a part of *Mid-Western development region*, which is a part of the country *Nepal*.

Finally, our ontology models weather statistics coming from weather stations through the *Station* class which is the super class of three sub-classes dubbed *RainfallStation*, *MeteorologyStation* and *ETOStation* to collect respectively rainfall, meteorology, and evapotranspiration statistics (for more details, see the bottom part of Figure 3). For example, *Kusum407*, located in the *Banke* district, is an instance of the *RainfallStation*, which a subclass of the *Station* class.

V. LINKING

We aimed to link our dataset with as many data sources as possible to ensure maximal reusability and integrability in existing platforms. All links are generated by using the LIMES framework [9]. In this framework, heuristics can be defined for the similarity of RDF resources and all similarity values

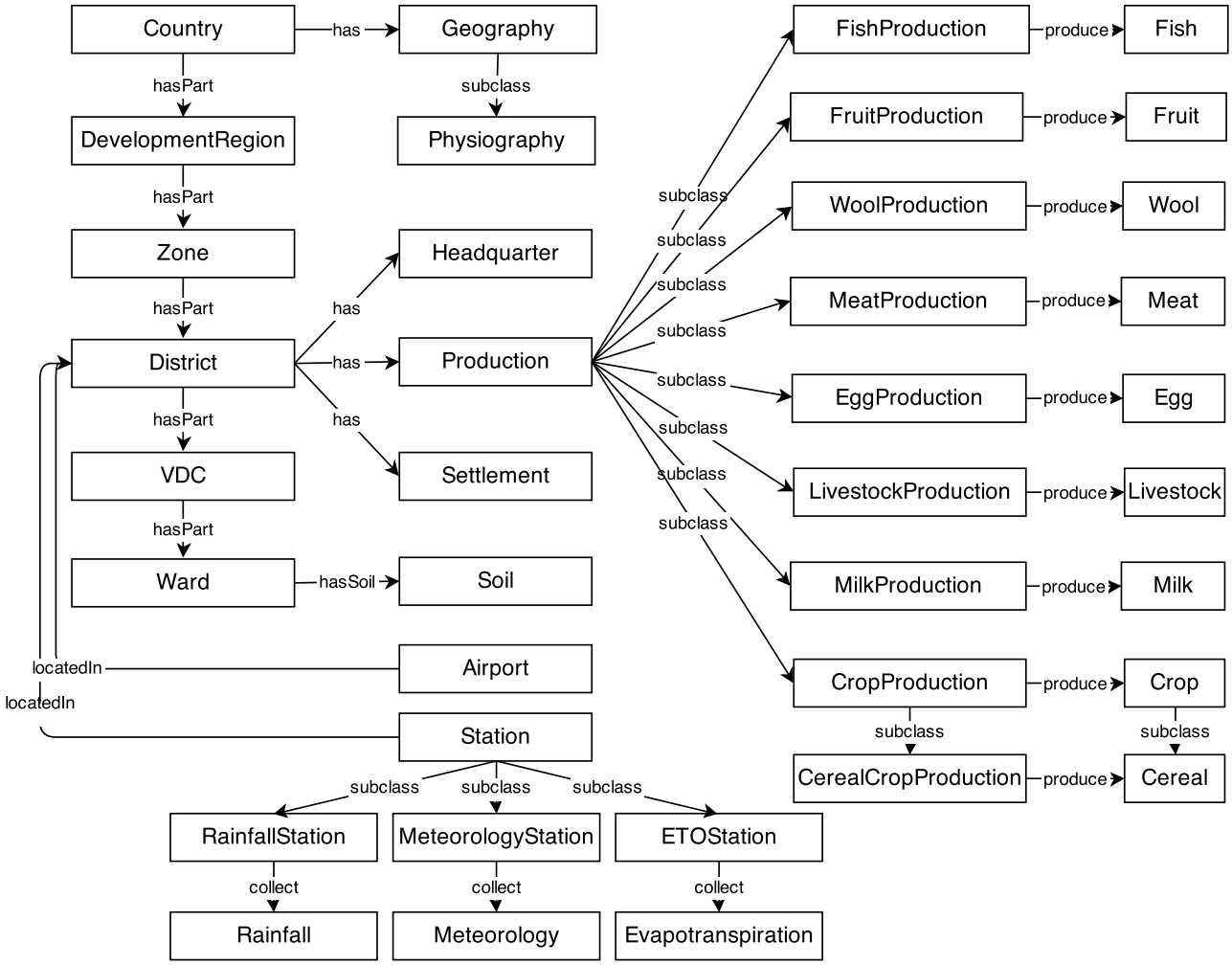

Fig. 3. *AgriNepalData* ontology structure

exceeding a particular threshold are considered links. So far, we have generated links to *DBpedia* as well as *AGROVOC*.

```
<SOURCE>
 <ID>AgriNepalData</ID>
 <ENDPOINT>http://agrinepaldata.com/sparql</ENDPOINT>
 <GRAPH> </GRAPH>
 <VAR>?x</VAR>
 <PAGESIZE>1000</PAGESIZE>
 <RESTRICTION>?x a agro:District</RESTRICTION>
 <PROPERTY>geos:hasGeometry/geos:asWKT RENAME polygon</
      PROPERTY>
 <PROPERTY>rdfs:label AS nolang->lowercase</PROPERTY>
</SOURCE>

<TARGET>
 <ID>DBpedia</ID>
 <ENDPOINT>http://dbpedia.org/sparql</ENDPOINT>
 <VAR>?y</VAR>
 <PAGESIZE>1000</PAGESIZE>
 <RESTRICTION>?y a dbpedia-owl:Settlement</RESTRICTION>
 <PROPERTY>geo:geometry RENAME polygon</PROPERTY>
 <PROPERTY>rdfs:label AS nolang->lowercase</PROPERTY>
</TARGET>

<METRIC>AND(hausdorff(x.polygon,y.polygon)|0.7,
   trigram(x.rdfs:label,y.rdfs:label)|0.7)</METRIC>
```

Listing 3. Fragment of the link specification for linking districts of Nepal between AgriNepalData and DBpedia.

For example, Listing 3 shows a LIMES link specification for linking Nepal districts in *AgriNepalData* to equivalent resources found in DBpedia. The link specifications used for other spatial resources such as zones and VDCs were essentially governed by similar metrics.

Table III shows details of links between *AgriNepalData* and both *DBpedia* and *AGROVOC*.

VI. Quality Measurement

A. Link Verification

For each class that contains links, we evaluated the quality of the links generated by LIMES by manually checking 100 randomly selected links[21]. The manual check was carried out by the first two authors. A link was set to be correct if both authors agreed on it being correct. The results are shown in Table III.

For linking places and airports, the linking achieves a precision between 0.97% and 100%. This high precision value

[21]In cases where there are less than 100 links, we checked all the links

TABLE III
NUMBER OF INTER-LINKS AND PRECISION VALUES OBTAINED BETWEEN *AgriNepalData* AND OTHER DATASETS.

Links between	Link property	Source dataset	Source Instances	Target Dataset	Target Instances	Accepted Links	Verified Links	Precision
Places	owl:sameAs	AgriNepal	37161	DBpedia	754450	524	100	0.97
Species	owl:sameAs	AgriNepal	1265	DBpedia	239194	192	100	0.93
Airport	owl:sameAs	AgriNepal	43	DBpedia	12688	27	27	1
Species	owl:sameAs	AgriNepal	1265	AGROVOC	32294	53	53	0.91

is because we configure LIMES to use a combination of two metrics: (1) the string matching between resources labels, and (2) the geo-spatial matching between resources' *Well-known text* (WKT) using the *Hausdorff*[10] metric (see Listing 3 for an example of combining string and spatial metrics). The recall could not be computed manually due to the absence of ground truth.

For linking species, the linking achieves a precision between 0.91% and 0.93%. In this case, we used exact string matching as otherwise the precision turned out to be too low. As its name implies the exact match gives us only species with identical names to be linked.

B. Dataset Verification

For dataset verification, we used the *RDFUnit*[22] framework. *RDFUnit* generates 956 test cases for all vocabularies used within *AgriNepalData*. Among 956 test cases, the results provided by *RDFUnit* shows that 935 test cases are passed, 3 are failed and 18 time out. Additionally, It shows that 65417 triples contain errors from a total of 1433631 triples, with average error rate of 0.045 per triple. Given that there are 327475 distinct subject in *AgriNepalData*, the average error per distinct subject is 0.199.

All the failed test cases were due to some errors in the raw data. For example, one test case detected that all airports have latitude values out of the valid range $[-90°, 90°]$, which leads us to review the original data, finding a bug in raw data as there were missing floating point symbols (for instance the value of $28.78°$ was saved as $2878° > 90°$). Therefore, we manually fixed this bug. We iterated this process until all of the test cases passed.

VII. USE CASES

A. Irrigation In Field

Example: *Mr. Bhandari* who lives in *Bhairawa*, Nepal wants to know *"How much irrigation water is required for a wheat plant which was planted in November 1 through out the life time of plant (120 days)?"*. To answer this, he first needs to know the weather condition. Therefore, he needs to look for the rainfall of each of the 120 days. Also, he needs to know the maximum and minimum temperature, humidity, wind status and sunshine hours. In addition, wheat as well as any other crops have their own crop specific water requirements. Finally, he needs to know the current water contained in soil which depends on soil type and previous rainfall status. To do so, he has to gather all of these pieces of information manually from different sources and update the information daily. Not only the process of finding all this information is tedious, but also the resulting information storage, update, and integration is hard.

Lacking of timely access to such necessary information may lead to less productivity and constraint him to his traditional methods of farming. In addition, it is possible to develop a farming mobile application so that farmer can access these information from the farm without any prior technical knowledge.

The crop water[11] is calculated as follows:

$$ET_{crop} = ET_o \times K_c \tag{1}$$

where ET_{crop} is the crop evapotranspiration or crop water need (mm/day), K_c is the crop factor and ET_o is reference evapotranspiration (mm/day). Each crop has its own growing stages during its season. In the case of wheat, it has an initial-season of 15 days, development-season of 25 days, mid-season of 50 days and late-season of 30 days. Moreover, each place has its specific ET_o for every month.

Using our dataset, Listing 4 provides a SPARQL query for computing ET_{crop} for each of the 120 days of the wheat season, thereby answering the question of Mr. Bhandari.

```
SELECT ?place AS ?WheatPlace
((0.5 * xsd:float(?int) + 0.5 *
xsd:float(?dev)) * xsd:float(?etoNov)) AS ?WaterPerDayNov
((0.5 * xsd:float(?dev) + 0.66 *
xsd:float(?mid)) * xsd:float(?etoDec)) AS ?WaterPerDayDec
(xsd:float(?mid) * xsd:float(?etoJan)) AS ?WaterPerDayJan
(xsd:float(?lat) * xsd:float(?etoFeb)) AS ?WaterPerDayFeb
WHERE {
agrd:Eto707 bio:place ?place.

cros:cropKcEachSatageWheat agro:kcForInitialStage ?int;
                    agro:kcForDevelopmentStage ?dev;
                    agro:kcForMidSeasonStage ?mid;
                    agro:kcForLateSeasonStage ?lat.

agrd:Eto707 agro:etoOfNepalInNovember ?etoNov;
         agro:etoOfNepalInDecember ?etoDec;
         agro:etoOfNepalInJanuary ?etoJan;
         agro:etoOfNepalInFebruary ?etoFeb.
}
```

Listing 4. How much irrigation water is required for a wheat plant which was planted in November 1 through out the life time of plant (120 days)?

B. Agriculture Planner, Policy Maker

The process of agriculture planning requires a significant amount of diverse knowledge to be available. A part of such

[22]http://aksw.org/Projects/RDFUnit.html

knowledge is related to various *crops' statistics*, e.g. information like how many types of crops are planted in a particular district in a particular year? Also, which district has the maximum agri-production of a particular crop? Furthermore, *temporal information* is essential for long term agriculture planning, such as information related to a particular crop production in the last 10 years. Another part of agriculture planning information has *geographic nature*, such as the size/population/location of each district. Of course, not all those pieces of information can be found in our dataset. Nevertheless, thanks to the Linked Data principles, we can acquire the missing information from other datasets through links. For instance, to answer the a question like *which districts are self dependent in their agri-products?*, Listing 5 provides a SPARQL query to collect the requested information pieces about a crop, paddy, production in each district as ratio of production per person (tonne/person) as well as production per district unit (tonne/km^2) not only from our dataset, but also from *DBpedia* using federated query services provided by SPARQL 1.1.

```
SELECT DISTINCT ?district ?productionyear ?cropProduction
?yieldProduction ?districtAreaKmSq ?population
xsd:float(?cropProduction)/xsd:float(?districtAreaKmSq)
                          AS ?cropProdPerSqDistrict
xsd:float(?cropProduction)*1000.00/xsd:float(?population)
                          AS ?cropProdPerPerson
WHERE{
 SERVICE <http://dbpedia.org/sparql> {
   SELECT ?districtAreaKmSq ?population ?dbpediauri
   FROM <http://dbpedia.org>
   WHERE{
      ?dbpediauri dbp:area ?districtAreaKmSq;
                  dbp:population ?population;
                  dbp:title "Districts of Nepal"@en .
   }
 }
 ?districturi owl:sameAs ?dbpediauri.

 ?s ?p agro:CerealCropProduction;
   gnd:dateOfProduction ?productionyear
   qty:area ?cropingArea;
   agro:produce agrd:Paddy;
   agro:production ?cropProduction;
   agro:yield ?yieldProduction;
   agro:inDistrict ?districturi.

 ?districturi rdfs:label ?district.
   FILTER (lang(?district) = "en")
}ORDER BY ASC(?cropProdPerPerson)
```

Listing 5. Which districts are self dependent in their agri-products?

C. Agriculture Spatial Data Visualization

In order to understand the data, the spatial part of agriculture data like rainfall stations, airports and district are visualized using the *Facete*[6] tool. Facete is a web-based exploration and visualization application enabling the spatial faceted browsing of data with a spatial dimension. Figure 4 demonstrates the information about the rainfall station location. The left section of the figure contains the selection field where we selected station properties and below the facet value can be seen. The middle section of the figure contain the information about data which is displayed according to the selection from left section. For example, a agriculture planner wants to know the numbers of stations and their location in a specific area for collecting weather information. A planner may also interested for finding the nearby locations and visualized them. In the figure, the data `0407, 0408, 0413 etc.` are rainfall station numbers. The right part of the figure shows the location of different spatial location and the details can be seen by clicking the marker. The location of the *Lumbini* station is clicked which is marked by blue color.

VIII. Related Work

A survey of the current status of agricultural linked open data can be found in [12]. Moreover, [13] introduces *Agricultural Ontology Services* (AOS) as well as ontology engineering techniques for the agriculture domain. AOS is a web-based multilingual vocabulary editing and work flow tool, which transforms thesauri, authority lists and glossaries into *SKOS* concept schemes for use in a linked data environment.

The work introduced by [14] in addition to declaring the *AGROVOC* Conceptual Scheme, it explains the integration between AGROVOC and both the *Web Ontology Language* (OWL) and the *Simple Knowledge Organization System* (SKOS) model . Using string similarity link discovery algorithms, [15] link AGROVOC thesaurus with other datasets such as EUROVOC, NALT, GEMET, STW, LCSH and RAMEAU. The *Agricultural Science and Technology* (AGRIS)[16] is an important public domain database with currently around five million bibliographical records on agricultural science and technology. These records are classified using AGROVOC. The *multi-source environmental knowledge framework* (i-EKbase) proposed by [17] provides large-scale availability of relevant sensor-model data. Furthermore, they present lightweight ontologies based on extracted meta-data from heterogeneous data sources. The *Kirby Smart Farm* [18] is a prototypical livestock smart farm system with an architecture for rapid development, controlling quality of data and integrating them with things in the farm using geospatial analysis and providing Linked Data Cube for semantic analysis and visualization.

For schema matching of agricultural ontologies, [19] introduced a data-driven approach for discovering matches between different classification schemas. The approach is based on content analysis and linguistic processing in order to extract information in the form of relation tuples, use the extracted information to associate the content of different repositories and match their underlying classification schemas based on the degree of content similarity.

IX. Conclusion and Future Work

By providing the AgriNepalData data sources as Linked Data and combining them with other datasets, it is now possible to obtain a variety of related agricultural information from one structured dataset. We have done an initial demonstration via SPARQL queries or tool deployments that the resulting data enables several relevant use cases. This is the first step on a larger research agenda aiming at an increase of productivity and efficiency of farming in Nepal. While our study is limited to Nepal, it can also be generalised to other countries in

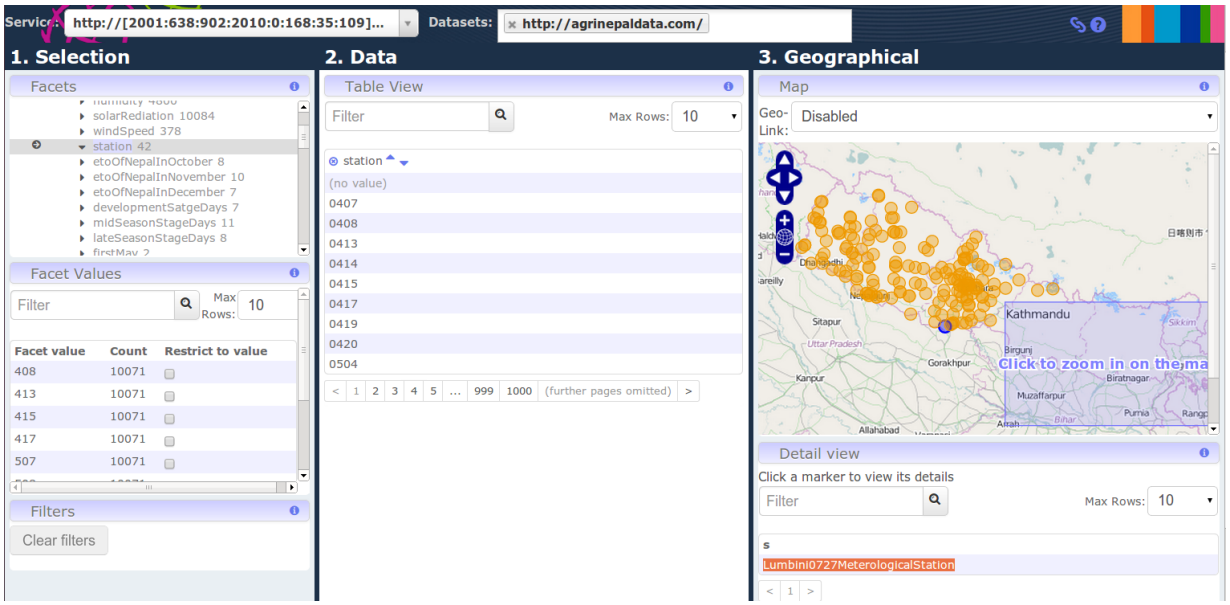

Fig. 4. Facete visualization of the *Lumbini* rainfall station.

the mid and long term. In future work, we plan to convert more data from previous years to enable large-scale temporal analysis. Furthermore, we intend to include other data from other domains that can influence the agricultural process like transportation and trade. As an important extension to our work, we aim to implement question answering techniques to enable non-experts to access our dataset from the project web site and mobile applications. Developing automatised solutions for each of the manual data conversion/verification tasks is one of the remaining general research challenges given the heterogeneous nature of data published by a number of different bodies.

ACKNOWLEDGMENT

We would like to thank Mr. *Chandra Prasad Ghimire*, VU University Amsterdam, Netherlands for providing weather data, Mr.*Badri Khanal*, Agri-Economist at Ministry of Agricultural Development, Nepal for providing crop statistics data and Mr. *Bikesh Shrestha*, Hydrologist/Water Resources Engineer at Consolidated Management Services Nepal Pvt. Ltd, Nepal for providing weather data. This work was supported by a grant from the European Unions 7th Framework Programme provided for the project GeoKnow (GA no. 318159).

REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International journal on semantic web and information systems*, vol. 5, no. 3, pp. 1–22, 2009.

[2] S. Auer, J. Lehmann, A.-C. N. Ngomo, and A. Zaveri, "Introduction to linked data and its lifecycle on the web," in *Reasoning Web*, 2013. [Online]. Available: http://jens-lehmann.org/files/2013/reasoning_web_linked_data.pdf

[3] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web Journal*, 2014.

[4] S. Auer and J. Lehmann, "Making the web a data washing machine - creating knowledge out of interlinked data," *Semantic Web Journal*, 2010. [Online]. Available: http://www.jens-lehmann.org/files/2010/washing_machine_swj.pdf

[5] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. J. Zaveri, "Test-driven evaluation of linked data quality," in *Proceedings of the 23rd international conference on World Wide Web*, 2014, to appear. [Online]. Available: http://svn.aksw.org/papers/2014/WWW_Databugger/public.pdf

[6] C. Stadler, M. Martin, and S. Auer, "Exploring the Web of Spatial Data with Facete," in *Proceedings of 23rd International World Wide Web Conference (WWW)*, 2014.

[7] I. Ermilov, S. Auer, and C. Stadler, "Csv2rdf: User-driven csv to rdf mass conversion framework," in *Proceedings of the ISEM '13, September 04 - 06 2013, Graz, Austria*, 2013. [Online]. Available: http://svn.aksw.org/papers/2013/ISemantics_CSV2RDF/public.pdf

[8] K. Patroumpas, M. Alexakis, G. Giannopoulos, and S. Athanasiou, "Triplegeo: an etl tool for transforming geospatial data into rdf triples."

[9] A.-C. Ngonga Ngomo and S. Auer, "Limes - a time-efficient approach for large-scale link discovery on the web of data," in *Proceedings of IJCAI*, 2011.

[10] A.-C. Ngonga Ngomo, "Orchid - reduction-ratio-optimal computation of geo-spatial distances for link discovery," in *Proceedings of ISWC 2013*, 2013.

[11] R. G. Allen, L. S. Pereira, D. Raes, M. Smith *et al.*, "Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage paper 56," *FAO, Rome*, vol. 300, p. 6541, 1998.

[12] D. Lukose, "World-wide semantic web of agriculture knowledge," *Journal of Integrative Agriculture*, vol. 11, no. 5, pp. 769 – 774, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2095311912600665

[13] A. Kawtrakul, "Ontology engineering and knowledge services for agriculture domain," *Journal of Integrative Agriculture*, vol. 11, no. 5, pp. 741 – 751, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S209531191260063X

[14] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, and J. Keizer, "The agrovoc linked dataset," *Semantic Web*, vol. 4, no. 3, pp. 341–348, 2013.

[15] A. Morshed, C. Caracciolo, G. Johannsen, and J. Keizer, "Thesaurus alignment for linked data publishing," 2011.

[16] A. Fogarolli, D. Brickley, S. Anibaldi, and J. Keizer, "Agris-from a bibliographical database to a web data service on agricultural research information." *Agricultural Information Worldwide*, vol. 4, no. 1, 2011.

[17] R. Dutta, A. Morshed, J. Aryal, C. D'Este, and A. Das, "Development of an intelligent environmental knowledge system for sustainable agricul-

tural decision support," *Environmental Modelling & Software*, vol. 52, no. 0, pp. 264 – 272, 2014.

[18] R. Gaire, L. Lefort, M. Compton, G. Falzon, D. Lamb, and K. Taylor, "Semantic web enabled smart farming," in *Proceedings of the 1st International Workshop on Semantic Machine Learning and Linked Open Data (SML2OD) for Agricultural and Environmental Informatics, ISWC*, 2013.

[19] A. Koukourikos, G. Stoitsis, and P. Karampiperis, "Data-driven schema matching in agricultural learning object repositories," in *Metadata and*