FOODpedia: Russian Food Products as a Linked Data Dataset

Maxim Kolchin(✉), Alexander Chistyakov, Maxim Lapaev, and Rezeda Khaydarova

Laboratory ISST, ITMO University, St Petersburg, Russia
kolchinmax@niuitmo.ru, {al.ol.chistyakov,mignolowa}@gmail.com, m.lapaev@telemetria.ru

Abstract. Open and efficient sharing of information about food products and their ingredients is important for all parties of the chain ranging from the manufactures to consumers. There exist a public catalogue of some Russian food products (http://goodsmatrix.ru/) that is used by some manufactures and consumers. Although the information is open, there are many difficulties in using the site, e.g., interoperability, querying and linking that could be mitigated by Semantic Web technologies. This paper presents an approach and a project for extracting and publishing information about food products and also linking it to existing datasets in Linked Open Data Cloud.

Keywords: Knowledge graph · Linked open data · Semantic web

1 Introduction

The goal of this work is to create a 5-star[1] open data dataset about Russian food products and their ingredients. Such work involves *(a)* food ontology development, *(b)* crawling of the existing sources, *(c)* publishing of the information as Linked Data and *(d)* linking to existing LOD datasets, such as *AGROVOC* [1] and *DBpedia* [2].

Based on the dataset that is created using Semantic Web technologies, new applications and services can be built, e.g. manufacturers can uses it to standardise the names for the ingredients, retailers can reuse the information on their e-shops, developers can built applications for customers that help them decide which product to buy based on their health conditions or personal preferences.

2 Dataset Creation

The source of the information for FOODpedia is web site called GoodsMatrix[2] which is manually curated catalogue where information comes mainly from manufacturers.

[1] http://5stardata.info/.
[2] http://goodsmatrix.ru.

© Springer International Publishing Switzerland 2015
F. Gandon et al. (Eds.): ESWC 2015, LNCS 9341, pp. 87–90, 2015.
DOI: 10.1007/978-3-319-25639-9_17

Extraction of food product information from GoodsMatrix goes through a pipeline that includes *(a)* crawling the web site using *Scrapy*[3] framework and set of XPath expressions, *(b)* parsing the resulting data to extract information about energy values, ingredients and E-additives, *(c)* translation of the name and description to English and *(d)* linking ingredients information to resource in *AGROVOC* and *DBpedia* datasets.

The source code of the crawler and other artifacts are available in Github repository[4].

Extraction of Ingredients. Ingredients are crawled as a list of ingredients separated by some character such as comma or semicolon. But there is an unsolved issue, it's rare when different manufacturers use the same names for the same ingredients, some ingredients can have more than dozen alternative names. Usually such names are different only because of word order, missing or extra words, therefore we apply the Ratcli-Obershelp algorithm [3] to measure string similarity and create single resource for similar names.

Extraction of E-additives. E-additives are food additives which have special identifiers called E numbers such as E-100, E-201, etc. and are used in Europe, Russia and other countries. Since the identifiers have well-defined structure, it's quite easy to find them in the ingredient list using regular expressions. The only issue is additives which have E-number, but written on the package without its number, e.g. Curcumin[5].

Multilingual Support. The name and description of food product crawled earlier are translated to English with help of *Yandex.Translate API*[6].

Linking. Extracted E-additives and ingredients are linked to similar resource in *AGROVOC* and *DBpedia* datasets.

AGROVOC is a multilingual agricultural thesaurus consisting of over 32 000 concepts available in 21 languages including Russian, therefore it's a good candidate for linking. Ingredients are mapped to AGROVOC concepts automatically, but it doesn't support E numbers because of that they are mapped manually.

DBpedia is a good source of human readable descriptions of concepts, therefore it's interesting to link E-additives and ingredients to its resources, but it's not so easy, because the ontology is generated semi-automatically. Therefore the mapping is performed manually.

[3] http://scrapy.org/.
[4] https://github.com/ailabitmo/foodpedia.
[5] http://dbpedia.org/resource/Curcumin.
[6] https://api.yandex.com/translate/.

3 Ontologies

To represent food products and their ingredients, Food Product Ontology[7] were developed which extends GoodRelations[8] and Food Ontology[9]. Below you find an example of food product in Turtle:

```
foodpedia:4601242311914 a food:Food;
    fpr:carbohydratesPer100gAsDouble "13.1"^^xsd:double;
    food:containsIngredient foodpedia:E952, foodpedia:E412,
                            foodpedia:E202;
    fpr:energyPer100gAsDouble "52.4"^^xsd:double;
    fpr:fatPer100gAsDouble "0.0"^^xsd:double;
    food:ingredientsListAsText "вода, томатная паста,
                                яблочное пюре, сахар, соль,
                                E412, уксусная кислота,
                                перец красный, E202, укроп,
                                E952"@ru;
    fpr:proteinsPer100gAsDouble "0.0"^^xsd:double;
    gr:description "Кетчуп второй категории с добавлением
                    фруктового пюре"@ru;
    gr:hasEAN_UCC-13 "4601242311914";
    gr:name "КЕТЧУП АРСЕНТЬЕВСКИЙ ОСТРЫЙ 900 Г"@ru,
            "KETCHUP ARSENIEVSKIY ACUTE 900 G"@en.
```

Also an example of ingredient with links to similar resource in *AGROVOC* and *DBpedia* datasets:

```
foodpedia: a food:Ingredient;
    rdfs:label "сахар"@ru, "сахар-песок"@ru, "sugar"@en;
    skos:exactMatch agrovoc:c_7498, dbpedia:Sugar .
```

4 Publishing

The dataset is published using *Pubby*[10]. The interface for human and machine consumption is available at http://foodpedia.tk. Using the SPARQL endpoint[11] provided by the underlying *Virtuoso Triple Store*[12], actors are able to satisfy complex information needs. In addition, actors are able to use another query interface through Linked Data Fragments [4] server[13] for high-availability querying. And last, human can use a simple search interface (see Fig. 1) to find food products by its barcode or name.

[7] @prefix fpr: <http://purl.org/foodontology#>.
[8] @prefix gr: <http://purl.org/goodrelations/v1#>.
[9] @prefix food: <http://data.lirmm.fr/ontologies/food#>.
[10] https://github.com/cygri/pubby.
[11] http://foodpedia.tk/sparql.
[12] http://virtuoso.openlinksw.com.
[13] http://data.foodpedia.tk.

Fig. 1. FOODpedia search interface

Licensing. All published data is openly licensed under Creative Commons Attribution License in accordance with the open definition[14].

Acknowledgements. This work has been partially financially supported by the Government of Russian Federation, Grant #074-U01.

References

1. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., Keizer, J.: The agrovoc linked dataset. Semant. Web **4**(3), 341–348 (2013)
2. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. Semant. Web **6**(2), 167–195 (2015)
3. Ratcliff, J.W., Metzener, D.E.: Pattern-matching-the gestalt approach. Dr. Dobbs J. **13**(7), 46 (1988)
4. Verborgh, R., et al.: Querying datasets on the web with high availability. In: Mika, P., et al. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 180–196. Springer, Heidelberg (2014). http://dx.doi.org/10.1007/978-3-319-11964-9_12

[14] http://opendefinition.org.