Food Waste Ontology: A Formal Description of Knowledge from the Domain of Food Waste

Riste Stojanov
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius, University
Skopje, North Macedonia
riste.stojanov@finki.ukim.mk

Tome Eftimov
Computer Systems Dept.
Jožef Stefan Institute
Ljubljana, Slovenia
tome.eftimov@ijs.si

Hannah Pinchen
Quadram Institute
Norwich Research Park
Norwich, Norfolk UK
hannah.pinchen@quadram.ac.uk

Maria Traka
Quadram Institute
Norwich Research Park
Norwich, Norfolk UK
maria.traka@quadram.ac.uk

Paul Finglas
Quadram Institute
Norwich Research Park
Norwich, Norfolk UK
paul.finglas@quadram.ac.uk

Drago Torkar
Computer Systems Dept.
Jožef Stefan Institute
Ljubljana, Slovenia
drago.torkar@ijs.si

Barbara Koroušić Seljak
Computer Systems Dept.
Jožef Stefan Institute
Ljubljana, Slovenia
barbara.korousic@ijs.si

***Abstract*—Recently, as a part of an EU-funded project called REFRESH, a new web-based tool named FoodWasteEXplorer was developed. It provides an easy access to valuable data on unavoidable food waste that can be explored by researchers, industry, governmental agencies and the public to find ways of its valorization. The food waste data was manually collected and stored in a relational database. To enrich and make best use of it, we automatically transform the collected information into a new food waste ontology. The created Food Waste Ontology provides a formal description of knowledge from the food waste domain. Examples of its application are: (i) database querying based on natural language questions, and (ii) finding new or missing data from other datasets.**

***Index Terms*—food waste standardization, data normalization and linkage**

I. Introduction

Approximately one-third of foods fit for human consumption are wasted globally (ca. 1.3 billion tonnes annually) [2]. The recently finished EU-funded project REFRESH [5] aimed to help towards reducing food waste across Europe and beyond and to maximise the value of unavoidable food waste and packaging materials. One element of the REFRESH goal was improved use of unavoidable food waste (e.g. peel, pomace etc.). Thus, some of the most common food products (ca. 120) (e.g. lemon) and their associated 1,264 side streams (e.g. lemon peel) were identified, based on European consumption statistics [10] and the environmental impact of its production [9], [11], [12]. This information was used to develop a new web-based tool named FoodWasteEXplorer [8], which supports productive use of these natural resources.

A. FoodWasteExplorer

FoodWasteEXplorer is free-of-charge for researchers, government agencies, industry including SMEs, and the general public. Filters can be applied to retrieve selected subsets of data, such as side streams (e.g. peel, stalks, seeds) and component groups (vitamins, minerals, proximates, bioactives, toxicants and other waste related components), and search results can be exported for further offline analysis. It is a tool for those exploring how food waste might be better used, e.g. citrus peel limonene can be used to make medical plastic. Potentially, a fruit juice producer could use FoodWasteEXplorer to identify this and start the process towards alternative applications.

Currently, the FoodWasteEXplorer contains **27,069 data points**, representing **587 nutrients**, **698 bioactives** and **49 toxicants**, collected from a variety of data sources, including scientific (peer-reviewed) papers, manufacturers' data (grey literature) and other data sources. This work is on-going, and more data will be added with time.

B. Food waste data normalization and its linkage with other data sets

There exist many sources of food-related data (such as food composition data, factors related to valorization processing such as heating values and pH, market value of recovered compounds from agricultural wastes, waste geographical location and volume of waste available), which could be linked with food waste data. All these data sources use different standards to describe the data. To link them data normalization process should be applied. There are other reasons for data normalization, among those are: (i) the amount of redundant data is reduced, (ii) anomalies when maintaining data can be avoided, and (iii) relations between data can be identified.

In this paper, we use a methodology to transform the food waste data into a new food waste ontology together with its linkage to other food-related data. To the best of our knowledge, this is the first attempt to model and represent the food waste domain. Section II provides an overview of the related work. In Section III, the methodology for developing a food waste ontology and its linkage with other relevant food ontologies is described in details. Sections IV and V, provide a discussion, conclusions and directions for future work.

II. Related Work

Several knowledge bases and ontologies have been already developed as open data sets to support machine learning (ML) research. Some of them are DBpedia and large number of biomedical ontologies avilable as a part of the BioPortal repository.

DBpedia is a project that aims to extract knowledge from the information collected by Wikipedia in a form of structured content [15]. DBpedia allows users to semantically query features and relationships, including links to other related ontologies. It also includes *Food* entities, which are defined as "any eatable or drinkable substance that is normally consumed by humans" [1].

SNOMED CT or SNOMED Clinical Terms [6] is the most comprehensive systematically organized collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. It provides the core terminology for inter-operable electronic health records, including terms related to clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other etiologies, substances, pharmaceuticals, devices and specimens. It also includes food-related data (e.g., data on food allergens) [7].

FoodOn [3] is a recently developed ontology that addresses food-related concepts. An ontology is a formal description of knowledge as a set of concepts within a domain and relationships that hold between them [13]. It interoperates with the Open Biological and Biomedical Ontology (OBO) Library [4] and currently represents entities related to food for humans, but in the future it will also encompass materials in natural ecosystems and food webs. Its aim is to develop semantics for food safety, food security, the agricultural and animal husbandry practices linked to food production, culinary, nutritional and chemical ingredients and processes.

BioPortal is an open repository that consists of biomedical ontologies, services to access to those ontologies, and tools that can be used to explore them [20]. Biomedical ontologies are crucial for data integration, information retrieval, data annotation, natural-language processing and decision support. Currently, there are 816 different ontologies related to the biomedical domain and can be used for data linking. Some of them also include food-related data.

In all above-mentioned food data resources, the food information is limited and does not include any information about food waste.

III. Methodology

The food waste data that is available in the FoodWasteExplorer was collected from peer-reviewed publications and other data sources containing relevant composition data. The data was manually collected by domain experts and stored in a structured way. A basic categorisation of the data was manually done by domain experts, by grouping within food type and then waste stream. One important feature is also the information about potential valorization approaches for each waste stream. However, this was only partially completed and not all side streams have relevant valorization information. Because it can happen that different data sources consist of information about same compounds in same side streams, the values from all data sources were stored together with a reference. In such cases, the user is able to follow the data source from where the data was collected and have more information about the data quality.

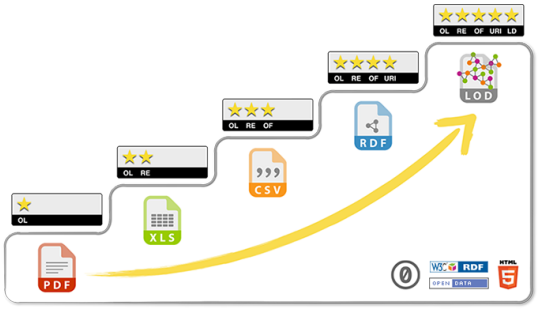

Fig. 1. Steps to publish 5 star Linked Data

The data presented by FoodWasteEXplorer is currently structured in a relational database, which was designed according to use-cases of FoodWasteEXplorer. Most of the data gathered in this database can serve the general good and can support future research that may potentially solve the food waste challenges. FoodWasteEXplorer is designed according to Open Data principles and enables browsing, searching and data downloading for offline analysis. Besides the HTML website presentations, PDF and spreadsheet data formats are currently supported. While the PDF format is intended only for humans and is not of a much value for the applications, the spreadsheet format provides structured data representation, which may be used for further data analysis. Even though the spreadsheet format seems advanced and application friendly, it does not fully capture the relationships among the data, and it is hard to relate the content to its real life concept. For this reasons, the goal of our study is to make the food waste information available as Open Data, which means that the data should be available in a raw and machine-readable format, for the purposes of use, reuse, republishing and redistributing, with little or no restrictions [18]. Therefore, the open format datasets can enable building of useful applications, which leverage their value and offer different use-cases for the interested parties. Additionally, to integrate this data with another data source will require suitable alignments and links.

The Linked Data [14] community has developed tools and methodologies for interlinking data from the Web by their meaning. The Linked Data techniques rely on identifying resources with URIs, providing data about these resources, and connecting them to other resources on the Web, by using standards such as the Resource Description Framework (RDF) [19]. By following these techniques, the FoodWasteExplorer data was transformed into five-star Linked Open Data. Figure 1[1] shows different quality ratings that can be assigned to

[1]Source: https://5stardata.info (Accessed 14 Nov 2019)

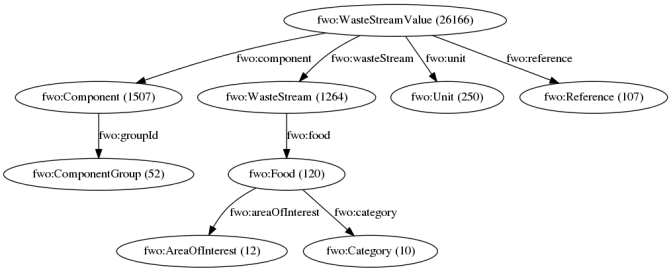

Fig. 2. Food Waste Ontology with the number of instances in each class

the data. The FoodWasteExplorer enables exporting data into non-proprietary computer readable format (Comma Separated Values - CSV), which is 3 star data. We continue by explaining how this data is further enriched to become identifiable via Unified Resource Identifiers (URI) and connected to other Linked Data knowledge bases. In this process, the well-established platform for exposing relational data in a RDF format, D2RQ [16], was used. This platform enables mapping of the relational database's tables to RDF types, the columns into RDF properties, and their values into RDF resources and literals. The D2RQ platform comes with multiple tools, but only the mapping generation tool and the D2RQ server were used. The mapping generation tool was used to initialize the mapping configuration; extracting the mapping format from the database schema. In this process, only the types that relate to the food and food waste domain were selected, while the security and maintenance related tables remained private. The endpoint exposed by the D2RQ server enables browsing of the exposed data, obtaining the data in the RDF format, and querying it using the SPARQL endpoint[2].

A. Food Waste Ontology

Figure 2 represents the food waste ontology[3]. As it was already explained, it is a result of experts' analysis of scientific (peer-reviewed) papers, manufacturers' data (grey literature) and other data sources. Its goal is to identify the waste streams for different foods, and to identify the amounts of different components present in these side streams. Currently, the ontology exposes **120** *Food* instances, which may belong in multiple categories and area of interests, represented with the classes *Category* and *AreaOfInterest*, respectively. For each food data, its waste streams (the *WasteStream* class) are provided, such as pomace, peel, etc. The food's waste streams data currently includes compositional values for **1507 components**. The *Component* class includes instances like proximates (e.g., fat, protein), inorganics (e.g., sodium), vitamins (e.g., niacin), bioactives (e.g., anthocyanin) and undesirable components (e.g., bacteria and toxins). We should also mention that any other components that are useful, but do not fit under the above-mentioned classification, are classified under *"waste compounds"*.

[2]http://purl.org/fw/snorql

[3]http://purl.org/fw

The main class of the ontology is the *WasteStreamValue* class, which represents the amount of waste per *Component* for each *Food*'s *WasteStream*. Each value is also associated with a unit, stored by the *Units* class, and a reference that provides the source that has published the corresponding value (denoted as *Reference* class in the ontology).

Figure 2 also shows that the Food Waste Ontology contains 9 classes, where the numbers of instances per class is displayed in the brackets. Additionally, the ontology contains 28 properties (including rdf:type and rdfs:label).

B. Linking to other open data sets

Since the data published by the presented ontology is manually gathered and represented according to the referencing source, it can not be easily connected to the available open knowledge available elsewhere. However, some of the classes such as *Food*, *Component*, *Category*, *AreaOfInterest* and *Unit*, can be linked with other ontologies and knowledge bases. To transform the isolated four star ontology into a five star linked data, the tools provided by the Linked Data community were used. For this reason, the *owl:sameAs*[4] property was used, in order to link the FoodWasteEXplorer's resources to *DBpedia*. The sum of in-links toward DBpedia in this moment is **39,007,478**[5], which allows use of the data in many different use-cases. The *Food* and *Component* instances are also linked to *SNOMED CT* terminology, in order to unify the FoodWasteEXplorer terminology with the commonly accepted clinical terms. The BioPortal API [22] is used for obtaining the terms from the *SNOMED CT* terminology. Moreover, this API additionally enables searching through all ontologies registered to this repository. Therefore, the mapping tool enables linking not only to the *SNOMED CT* terminology, but to many other ontologies, including *FoodOn*, *ISO-FOOD* [17], and many others.

The main goal of the FoodWasteEXplorer and its ontology is to achieve a good data quality. For this reason, the linkage to the other ontologies should be checked by domain experts, in order to provide a validated knowledge base that can be further used for more advanced data analysis. Figure 3 shows the interface that was developed to speed up the experts checking process. Using this tool, the expert can search each term in **DBpeda** database[6] and *BioPortal* registered databases[7], and select only the relevant results. The user interface is optimized for displaying multiple query results by presenting only abbreviated information, while more details are presented when the expert will place the mouse over the particular resource. This tool brings all relevant results from multiple sources in one place, which significantly speeds up the process of resource interlinking. Additionally, the mapping tool enables the expert to change the query term if they think the resulting resource is represented differently in the linking knowledge bases.

[4]https://www.w3.org/TR/owl-ref/sameAs-def

[5]https://wiki.dbpedia.org/services-resources/interlinking (Accessed: 6 Oct 2019)

[6]The DBpedia lookup API is used for extracting candidate resources.

[7]The BioPortal API is used for terms querying.

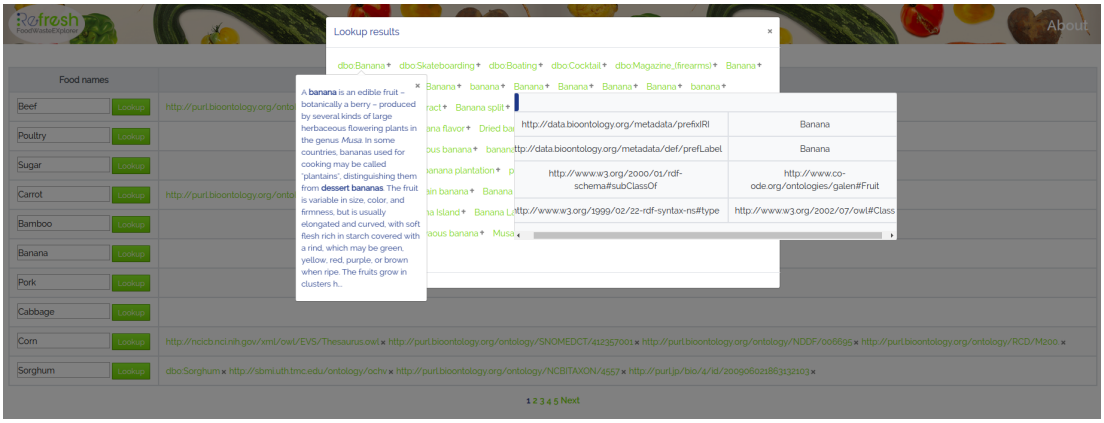

Fig. 3. Mapping interface

IV. Discussion

The data collection process is always a challenging task. Data may be published in papers or reports in different formats, which requires advanced techniques for data normalization and linkage. Besides that, new data is emerging very quickly and finding a way of upgrading databases with new or missing data is urgent. Moreover, users of information systems are becoming reliant on better information sources in the sense that the information they are looking for needs to be provided as quick as possible and in a complete form. To satisfy these needs, new approaches need to be developed and the Food Waste Ontology presented in this paper is an example of such a solution.

The ontology presented in this paper is created and published following the Linked Data guidelines. All its data is manually gathered by domain experts, using relevant sources. Additionally, it is mapped to DBpedia and large number of ontologies available at the BioPortal repository. This makes the Food Waste Ontology a valid dataset that may be used as gold standard for many future information extraction tasks in this field.

There are multiple ways for accessing this expert knowledge: (1) through FoodWasteEXplorer web application and its PDF and spreadsheet exports, (2) using the facet browser provided by the D2RQ server, (3) the RDF data enables applications to access the resources and discover their interconnections, and (4) the SPARQL endpoint enables pragmatically querying of the data using the W3C's query language for the semantic web [21]. All these access methods support different use-cases, such as data searching (1 and 4), resource navigation and exploration (2 and 3) etc. The variants 1 and 2 enable the human users to explore and understand this knowledge, while 3 and 4 are intended to be machine readable, in order to enable other applications with richer features. Additionally, the SPARQL query language is flexible enough to fit the natural language questions and to enable querying the FoodWasteEXplorer database using natural language questions that are automatically translated into database queries.

Food Waste Ontology improves the data stored in the relational database by providing a globally identifiable resources, that are linked with any available dataset on the Web. This way, new use-cases can emerge with the combination of the inter-linked data, and new inter-disciplinary questions can be answered due to the wider available knowledge. Therefore, the *Components* can be potentially linked to other ontologies that describe products that contains them, which leads toward the greather goal of exploring the valorisations of the foods' waste streams.

V. Conclusion

The Food Waste Ontology presented in this paper enables further development of research in the field of food-related data science. The way of developing the Food Waste Ontology from the relational database is interesting as it enables quick and efficient transformation of data into knowledge. Its linkage with other open data sets is also supported, thus enabling the usage of food waste data in different contexts.

In the near future, we aim to evaluate the Food Waste Ontology by the help of human experts from the food domain. The evaluation will be performed in two ways. First, the human experts will evaluate the new way of exploring data in the FoodWasteEXplorer. They will formulate several open food waste science questions, whose answers they will try to find by the help of the FoodWasteEXplorer. Second, they will perform an evaluation of new food waste data acquired by the FoodWasteEXplorer through the Food Waste Ontology.

Acknowledgment

This work was supported by the project REFRESH, which was funded by the Horizon 2020 Framework Programme of the European Union under Grant Agreement no. 641933 (2014-2018). Another financial support that this project received was from the Slovenian Research Agency (research core funding No. P2-0098). The authors acknowledge the financial support from both funding parties.