APRICOT: A humAn-comPuteR InteraCtion tool for linking foOd wasTe streams across different semantic resources

Bojan Dimoski
Faculty of Computer Science and Engineering
Ss. Cyril and Methodius, University
Skopje, North Macedonia
bojan.dimoski@students.finki.ukim.mk

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

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

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377744

***Abstract*—In the modern era of data, advanced approaches for extracting information and knowledge from data are required. Moreover, the extracted information needs to be formalised to be usable by information systems. Today, there exist several resources of semantics on food waste, which is a huge environmental problem that need to be fixed as soon as possible. Yet, the problem is that the existing semantic resources are not aligned and therefore needs to be linked. Only in this way, the complementary knowledge from different resources will become of real value. In the paper, an AutoMap algorithm for automated mapping of knowledge on food waste from different semantic resources is presented. By integrating such an algorithm in a web based tool, experts and the general public can get an insight into complex knowledge that is required for inventing new solutions for the food waste valorisation.**

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

I. Introduction

Around 40% of the food is wasted in the United States [17], while this number is around 33% globally [2]. This means that there are approximately 1.3 billion tonnes of food waste generated annually, which has significant environmental impact [8]–[10]. There are many ways of wasting the food, starting from the farms and fishing boats, during processing and distribution, in retail stores, in restaurants and at home. However, the wasted food that still contains valuable components can be valorised, i.e. turned into usable products. For example, pigs could stop 14 million tonnes of surplus food from being wasted[1]. Another solution is the valorization of food waste by producing nutraceuticals or non-food products (like packaging, cups etc.)

The spread of the information about the composition of various waste streams can lead to smarter waste management, processing and reusing. The aim of the recent EU-funded project REFRESH was to increase the use of unavoidable food waste of the most common food products and their associated 1,264 side streams (e.g. lemon peel). This goal has been delivered through the FoodWasteEXplorer [7], which is a web based tool that is intended to be used by the public in order to describe the components in the different side streams of the foods. It is free-of-charge for the general public, and allows investigation about the composition of each food side stream. Currently, its underlying database 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.

There exist numerous wellsprings of food-related information that could be connected with food waste data (for example, food arrangement information, factors identified with valorization handling, heating qualities and pH etc.). All these information sources utilize various principles to depict the information. The linking of the food waste information with other datasources, can provide broader information about the percentage of the nutrients or other constituents (e.g. bioactives) that are being wasted, product that use some of the wasted components, and many other use-cases. However, many of the food-related datasets are created for specific purpose, and represent and structure the concepts based on the specific use-cases. Furthermore, in order to provide inter-domain information, we may often use a general and highly connected datasource, such as DBpedia [14]. In this case, one can face ambiguity, since for instance the food apple may have very close lexical representation with the ”Apple Inc.”

[1] https://eu-refresh.org/press-release-pigs-could-stop-14-million-tonnes-surplus-food-being-wasted

or its products. Therefore, automated mapping is not always feasible.

In this paper, we present a mapping tool that is designed to link the knowledge extracted from the FoodWasteEXplorer and published by the Food Waste Ontology[2] with the other food and general purpose ontologies and taxonomies. In section II we describe available datasources with valuable information that can complement the information on food waste streams and their composition, in Section III we describe the methodology used to create the tool, after which in Section IV and V we discuss about the mapping tool impact, and provide conclusions and directions for future work.

II. Related Work

There exist several relevant food and general purpose ontologies and taxonomies. An ontology encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many or all domains of discourse. More simply, an ontology is a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject.

FooDB [27] is one of the biggest and most thorough assets of information on food constituents. It provides data on both macro-nutrients and micro-nutrients, which includes metadata on the flavor, color, taste, texture and aroma of the food. Every compound section in the FooDB contains in excess of 100 separate information fields covering compositional, biochemical and physiological data (acquired from literature). This covers information for the compound's terminology, its depiction, data on its structure, synthetic class, its physico-substance information, its food source(s), color, fragrance, taste, physiological impact, possible well-being impacts (from distributed studies), and focuses on different nourishments. More importantly, FooDB contains fields which give us confirmed external links to the data from various ontologies.

DBpedia [14] is a project with plans to get information from the data gathered by Wikipedia in a type of organized content. It permits semantic querying features and relationships, and it incorporates food entities, which are characterized as ”any eatable or drinkable substance that is normally consumed by humans” [1].

SNOMED CT or SNOMED Clinical Terms [5] is the most thorough assortment of clinical terms that provide codes, terms, synonyms and definitions utilized in clinical documentation and reporting. It gives the fundamental terms to interoperable electronic well-being records, including terms identified with clinical discoveries, symptoms, diagnoses, procedures, body structures, living beings and different etiologies, substances, drugs, gadgets and specimens. It additionally incorporates food-related information (e.g., information on food allergens) [6].

FoodOn [3] was created not long ago to serve as an ontology that tends to food-related ideas. An ontology is a proper depiction of information as a group of notions within an area and connections that hold between them [11]. It interoperates with the Open Biological and Biomedical Ontology (OBO) Library [4] and presently stands for elements identified as nourishment for people, yet later on it will likewise envelop materials in natural ecosystems and food webs. Its goal is to create semantics for food safety, food security, the agricultural and animal husbandry practices connected to food production, culinary, dietary and chemical ingredients and cycles.

BioPortal is an open repository that comprises biomedical ontologies, services with access to them, and tools that can be utilized to investigate them [23]. Biomedical ontologies are essential for information coordination, data recovery, information comment, natural-language processing and decision assistance. At present, there are *816 distinct ontologies* identified with the biomedical space and can be utilized for connecting information. Some of them incorporate food-related information as well.

Next, we will introduce other significant ontologies and datasets, related to the food domain. HMDB (Human Metabolome Database) [26] is a free database that contains comprehensive data about small molecule metabolites, found in the human body. The database is intended to contain or connect three sorts of information: 1) chemical data, 2) clinical data, and 3) molecular biology/biochemistry data. There are *114,215 metabolite entries* in the database, to which *5,702 protein sequences are linked.* Drug-Bank [25] contains data on drugs and drug targets. DrugBank is both a bioinformatics and a cheminformatics asset that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e.sequence, structure, and pathway) data. Phenol-Explorer [22] is a thorough database on natural phenols and polyphenols, which includes food composition, food processing, and polyphenol metabolites in human and exploratory animals. PubChem [20] is a database of chemical molecules and their exercises against organic measures. One can look for chemicals by name, molecular formula, structure and different identifiers. Data can be noticed about chemical and physical properties, biological activities, safety and toxicity, patents, literature citations and so on. ChEBI [16] is an ontology of molecular entities zeroed in on ’little’ chemical mixes. ChEMBL [12] is a physically curated database of bioactive molecules with drug-like properties. It unites chemical, bioactivity and genomic data to help the interpretation of genomic data into powerful new drugs. KEGG COMPOUND [19] is an assortment of little molecules, biopolymers, and other chemical substances that are pertinent to organic systems. Every section is recognized by the C number, for example, C00047 for L-lysine, and contains chemical structure and related data, just as different links to other KEGG databases and external databases.

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

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

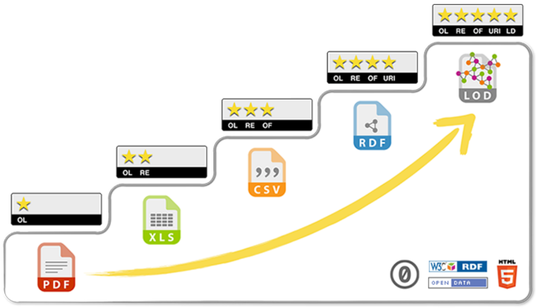


Fig. 1. Steps to publish 5 star Linked Data

A. Food Waste Ontology

In our previous work [24], we presented the Food Waste Ontology[3] that publishes the knowledge available in the FoodWasteEXplorer [4]. The data in this system was manually collected, organized and validated by the domain experts. The FoodWasteEXplorer database originates from peer-reviewed publications and other data sources containing relevant composition data. Most of the data gathered in this database can serve the general good and can support future research that may potentially solve food waste challenges.

The Food Waste Ontology publishes the data stored in a relational database, which was designed according to use-cases of FoodWasteEXplorer. The goal of this ontology is to make the food waste information available as Open Data, which means that the data should be available in a machine-readable format for the purposes of use, reuse, republishing and redistributing, with little or no restrictions [18]. Using the guidelines, techniques and methodologies for interlinking data published by the Linked Data [13] community, the data from the FoodWasteEXplorer is transformed into a five-star Linked Open Data. Figure 1[5] best illustrates the data transformation journey, from unstructured scientific documents to structured and publicly available data published using Resource Description Framework (RDF) [21], that preserves the structure and the relation of the resources. For the purpose of accurate data presentment, the relational database structure was directly mapped into the Food Waste Ontology, and the data is transformed on-the-fly using the D2RQ [15] server. The D2RQ server additionally provides download of the data in RDF, live browsing of the resources using its Faced Browser, and even executing queries towards its SPARQL endpoint[6].

The Food Waste Ontology structure is presented in Figure 2 and the description for each class is shown in Table I. Its main goal is to provide new use-cases and interlinking of the identified waste streams for different foods.

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).

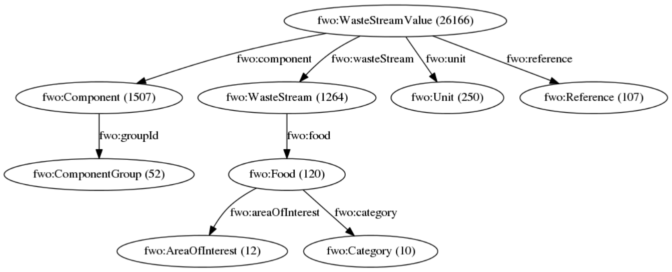


Fig. 2. Food Waste Ontology with the number of instances in each class

Class	**Instances**
db:fwo/ComponentGroup	52
db:fwo/Unit	250
db:fwo/Reference	107
db:fwo/ComponentSubGroup	82
db:fwo/Food	119
db:fwo/Component	1673
db:fwo/WasteStreamValue	26166
db:fwo/AreaOfInterest	12
db:fwo/Category	10
db:fwo/WasteStream	1259

TABLE I
FOOD WASTE ONTOLOGY CLASSES WITH NUMBER OF INSTANCES

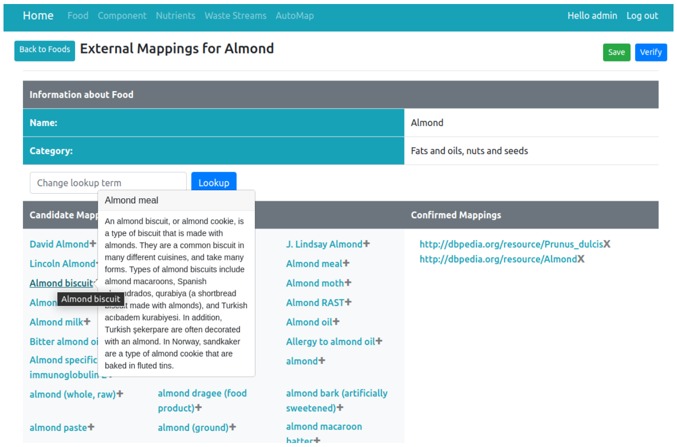
III. METHODOLOGY

The goal of the project presented in this paper is to create a convenient web tool for experts and the general public to access data about food waste streams, as well as food, and food compounds collected from various data sets in one place. This will give an insight on how the food and its components affect food waste.

The mapping tool being presented in this paper is named APRICOT, and is used for linking our data, either to other data sets, ontologies, or to the FooDB database. There are separate interfaces for mapping foods to external links, components to external links, or to compounds from FooDB. This linkage itself would not be any good if it is not checked and verified. Therefore, only users (experts of this domain), are authorized to use these interfaces to check and verify the data, thus achieving greater data quality. The interfaces are straightforward and their purpose is to speed up the whole process of foods and components mapping. Figure 3 illustrates the mapping interface, where the mapped item's original name and category are first displayed, followed by the candidate mappings in the table below the item. The **Candidate Mappings** column in the table gives all suggested links from the AutoMap algorithm (explained below). In the second column of the table, **Confirmed Mappings**, the confirmed links are displayed. As it is shown in Figure 3, there is an option to remove a confirmed mapping, or to add one by clicking on plus icon next to each link from the Candidate Mappings column. This way the experts can easily map the knowledge from the multiple databases.

[3] http://purl.org/fw
[4] https://www.foodwasteexplorer.eu/
[5] Source: https://5stardata.info (Accessed 06 Oct 2020)
[6] http://purl.org/fw/snorql

There is also an interface for authorized experts to edit and map the data to external sources, or verify previously mapped data. An auto-mapping (AutoMap) algorithm is implemented in this tool in order to speed up the mapping process. The purpose of this algorithm is to provide a starting point, with basic, but adequate quality information, which is linked to other sources of information from where the user can browse further and gather more information about the data of interest. The data exposed through the D2RQ server gives anyone the ability to use this project's data and develop complex apps providing analysis or do machine learning (ML) research. Moreover, there is the ability to query the data using the SPARQL query language.



External Mappings for Almond

Information about Food

Name:	Almond
Category:	Fats and oils, nuts and seeds

Change lookup term

Candidate Mapp

Almond meal

An almond biscuit, or almond cookie, is a type of biscuit that is made with almonds. They are a common biscuit in many different cuisines, and take many forms. Types of almond biscuits include almond macaroons, Spanish ndrados, qurabiya (a shortbread t made with almonds), and Turkish acıbadem kurabiyesi. In addition, Turkish şekerpare are often decorated with an almond. In Norway, sandkaker are a type of almond cookie that are baked in fluted tins.

Almond biscuit

David Almond+
Lincoln Almond
Almond biscuit
Almon
Almond milk+
Bitter almond oi
Almond specific immunoglobulin
almond (whole, raw)+
almond paste+
almond dragee (food product)+
almond (ground)+
J. Lindsay Almond+
Almond meal+
Almond moth+
Almond RAST+
Almond oil+
Allergy to almond oil+
almond+
almond bark (artificially sweetened)+
almond macaroon batter+

Confirmed Mappings

http://dbpedia.org/resource/Prunus_dulcisX
http://dbpedia.org/resource/AlmondX

Fig. 3. Web Mapping Tool interface

The unverified mappings are the result of the AutoMap algorithm, which looks up different datasets and ontologies (DBpedia, FoodOn and SNOMED CT in the current version) for the specified search term, and displays all the relevant results from this lookup, if there are any. There can be many results, and to make the user experience better, only the names are shown in the candidate mappings column. If the user needs more information for each of the links, they can hover over the title and additional information for that result will be shown to them. The users can even click on the item and they would be redirected straight to its source.

The AutoMap algorithm can significantly speed up the process when mapping many resources at once. Since the automatic mappings are never 100% accurate, they need to be verified. The AutoMap feature aims to make the mappers job easier. This tool provided the initial mappings for the data in Food Waste Explorer and FooDB resources. There are currently five actions available in this tool:

- Mapping food entities from Food Waste Ontology to foods in FooDB
- Mapping component entities from Food Waste Ontology to compounds in FooDB
- Transitive mapping food entities using the FooDB external owl:sameAs links
- Transitive mapping component entities using the FooDB external owl:sameAs links
- Mappings by querying food and component names in external sources (Currently DBpedia, FoodOn and SNOMED CT ontologies are queried and mapped)

The first two mappings enrich our data and expand it with new features that can improve our mapping and browsing of data, such as the external links available in the FooDB data. The next step is to map our data to external datasets. In this paper, we refer all datasets other that FoodOn as external, since the initial goal of the work is to align the Food Waste Ontology to FooDB. The third and fourth actions from above use the already available external links from FooDB and map them using the owl:sameAs field. This way, the components are automatically mapped to DBpedia, HMDB, DrugBank, Phenol-Explorer, PubChem, ChEBI, ChEMBL and KEGG. The last action extracts possible mappings by querying databases by terms automatically extracted from our data (usually the names). Currently the DBpedia, FoodOn and SNOMED CT ontologies are queried and mapped. We reused the functionality from our work in [24], where the BioPortal API 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, and many others.

One remark is that this tool only looks up possible mappings by the item's name, and this sometimes does not return any results. However, the user may know a synonym that can better represent the mapped concept. In order to support this use-case, we added an option to manually provide the lookup name, which may result in finding the actual result.

IV. Discussion

Besides the user-friendliness of the web mapping tool, multiple Food Waste Ontology (FWO) resources are linked to the outside sources. The main enabler of this linking are the 68 manually mapped FWO's Food resources to resources in FooDB. This enabled us to transitively add the FooDB's resources owl:sameAs links to the FWO's food entities. Additionally, the initial food mappings enabled the linking of 320 FWO components with the best proposed FooDB's compounds by the AutoMap algorithm. During this process, only suggestions for the components from Food Waste Ontology that share manually mapped food with the compounds from FooDB were mapped.

However, depending on the label of the corresponding resource, different number of candidate resources are obtained. Figure 4 shows the number of candidates obtained with this algorithm from external resources for the foods and components (columns 1 and 2), and from the FooDB compounds for components in column 3. The row number in this figure denotes the number of obtained candidates, while the bar chart value presents for how many of the FWO's resources we obtained this number of candidates. This graph shows that we obtained more that 10 external candidates for each of the 119 FWO's food resources, while for 81 of them we got 30 or more candidates. This presents the challenge to select the right mapping out of these many candidates. Therefore, in our future work we plan to introduce a ranking for these candidates, to simplify the job of the person that does the mapping. The high number of candidates obtained for the external sources also holds for the FWO components. The main reason for this high number of resources is the fact that in this process we combine the results from multiple lookup and search calls, all of which return multiple matching candidates. Lastly, we use string matching (exact match or with regular expressions[7]) when searching for compounds in FooDB, and this is the reason why we have **1156** components without match in this database. The **160** FWO components with one compound candidate are mapped automatically. Out of the multiple suggestions for the rest of the resources, additional **160** components were automatically mapped. Therefore, we have **68** manually mapped foods to FooDB resources, **83** food resources automatically mapped to external sources, including FoodOn and Snomed CT. Next, we have **320** components linked to FooDB compounds, and additional **273** external links toward DBpedia, HMDB, DrugBank, Phenol-Explorer, PubChem, ChEBI, ChEMBL and KEGG. In total, the Food Waste Ontology is extended by **383** links toward FooDB, and **1344** external links. And this is just the beginning, since this tool will simplify the mapping of the rest of the foods and components to other resources.

Even though there may be more than 68 FWO foods that can be mapped to FooDB's foods, we chose only the one that do not have any ambiguity, since those are the base for later automatic mapping. Furthermore, all links that are added with the AutoMap algorithm are un-confirmed, and each of them must be validated and confirmed by the domain experts. The Figure 3 shows the valid button that is used to confirm the suggested mappings, together with all other previously selected links. This way, the AutoMap feature speeds up the process, without corrupting the correctness of the linking.

The linking of the Food Waste Ontology with resources such as FooDB can open new use cases, such as ability to answer the question of the amount of micro nutrients wasted in each stream. Additionally, the link with DBPedia opens possibility for extracting products that can be produced using the components from different waste streams, and therefore create new ways of valorisation of the waste.

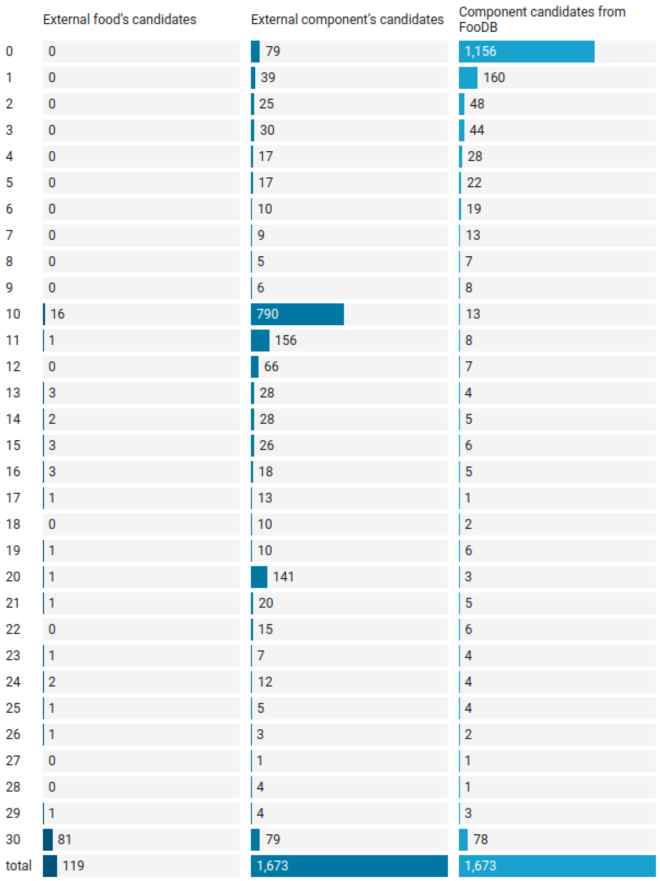


Fig. 4. Number of candidates for food and component

V. Conclusion

Since food waste is a massive problem, the interest in data and information on food waste streams has immensely increased. However, collecting data and information in a manual way is a difficult and time-consuming task, therefore, there is a strong need for advanced approaches for both knowledge extraction and mapping knowledge from different semantic resources. In this paper, an algorithm for automated mapping of knowledge on food waste already formalised in different ways by various resources was introduced. The algorithm, named AutoMap, was integrated in the newly developed web based tool APRICOT, which is primarily aimed for experts and the general public to explore knowledge acquired by the mapping of the Food Waste Ontology with the complementary knowledge from other food as well as general purpose ontologies and taxonomies. The Food Waste Ontology was also created in an automated way from the manually collected data that is provided through the FoodWasteEXplorer. In the paper, the mapping was demonstrated on FooDB as one of the biggest assets of the information on food constituents and few other, so called external resources. In total, the Food Waste Ontology has been extended by **383** links with FooDB, as well as by **1344** links with other resources. This opens new applications on data and information on food waste mapped from different, but complementary resources.

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

[7] The candidate lookup implementation can be found here: https://bit.ly/2IcHoMY