FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation

Steven Haussmann[1 [0000−0003−0216−1820]], Oshani Seneviratne[1 [0000−0001−8518−917X]], Yu Chen[1 [0000−0003−0966−8026]], Yarden Ne'eman[1 [0000−0002−3017−2722]], James Codella[2 [0000−0002−3430−0429]], Ching-Hua Chen[2 [0000−0002−1020−0861]], Deborah L. McGuinness[1 [0000−0001−7037−4567]], and Mohammed J. Zaki[1 [0000−0003−4711−0234]]

[1] Rensselaer Polytechnic Institute, Troy NY, USA
{hausss,senevo,cheny39,neemay}@rpi.edu, {dlm,zaki}@cs.rpi.edu
[2] IBM Research, USA
{jvcodell,chinghua}@us.ibm.com

Abstract. The proliferation of recipes and other food information on the Web presents an opportunity for discovering and organizing diet-related knowledge into a knowledge graph. Currently, there are several ontologies related to food, but they are specialized in specific domains, e.g., from an agricultural, production, or specific health condition point-of-view. There is a lack of a unified knowledge graph that is oriented towards consumers who want to eat healthily, and who need an integrated food suggestion service that encompasses food and recipes that they encounter on a day-to-day basis, along with the provenance of the information they receive. Our resource contribution is a software toolkit that can be used to create a unified food knowledge graph that links the various silos related to food while preserving the provenance information. We describe the construction process of our knowledge graph, the plan for its maintenance, and how this knowledge graph has been utilized in several applications. These applications include a SPARQL-based service that lets a user determine what recipe to make based on ingredients at hand while taking constraints such as allergies into account, as well as a cognitive agent that can perform natural language question answering on the knowledge graph.

Resource Website: https://foodkg.github.io

1 Introduction

Chronic diseases such as cardiovascular disease, high blood pressure, type 2 diabetes, some cancers, and poor bone health are linked to poor dietary habits [8]. Although much progress has been made in the development and implementation of evidence-based nutrition recommendations in the past few decades [17], that knowledge has not translated into day-to-day dietary practices. One of the barriers to putting recommended dietary guidelines into practice is that

the personalization of the guidelines (e.g., with respect to cultural and lifestyle differences) is largely left to individuals. Much more than just watching one's caloric, fat, salt, and sugar intake, guidelines also advise individuals to eat a variety of nutrient-dense foods. Thus, the number of nutritional parameters that need to be considered can become overwhelming.

A natural solution to this problem is to provide an intelligent and automated method for recommending foods. Trattner et al. [23] provide a comprehensive review of the state-of-the-art in food recommender systems. They highlight a recent but growing focus on not only recommending likable foods but going further and ensuring that they are healthful foods as well. The authors note that, despite its importance, food recommendation, in comparison to other domains, is relatively under-researched. Among the several works they reviewed, only [11] involved the use of semantics, motivating the need for methodologies for constructing a food-focused knowledge graph.

Knowledge graphs (KGs) have an important role in organizing the information we encounter on a day-to-day basis and making it more broadly available to both humans and machines. KGs have been used for a variety of tasks, including relationship prediction, searching for similar items, and question answering [6]. While machine learning algorithms can effectively answer questions, they are notorious for producing answers that are hard to explain, especially automatically. Knowledge graphs make it possible to produce automatic explanations of how answers were derived. Interoperability is another important aspect of knowledge graphs, as they enable understanding and reuse. However, the elusiveness of standards and best practices in this area poses a substantial challenge for knowledge engineers who want to maximize KG discovery and reuse, as dictated by the FAIR (Findable, Accessible, Interoperable, Reusable) principles [24].

In this paper, we discuss our methodology for extracting and maintaining publicly available data about food, and for constructing a knowledge graph that can be consumed by both humans and machines, thus providing useful food recommendations that can in turn promote healthier lifestyles. It is important to note that ours is the first extensive FoodKG resource spanning recipes, ingredients, and nutrients that covers over a million recipes and 67 million triples (see `https://foodkg.github.io`). The novelty and main contribution of our resource is its scope and inclusiveness, not only considering the different datasets it integrates, but the linking with health concepts and the offering of a question-answering service as an application.

1.1 Use Case

Our use case is designed to assist people in personalizing their dietary goals by providing them with information to improve the alignment between their eating behaviors and general nutritional recommendations. For example, consider the American Diabetes Association's (ADA; [1]) recommendation that "Carbohydrate intake from whole grains, vegetables, fruits, legumes, and dairy products, with an emphasis on foods higher in fiber and lower in glycemic load, should be advised over other sources, especially those containing sugars". Unfortunately,

translating this into healthful yet palatable food choices can be a daunting task for many individuals, which is partly due to the fact that knowledge is scattered across multiple sources. Thus, our goal is to assist people in exploring how different modifications to their meals can affect their alignment with guidelines by providing a robust system that can be used to construct a Food Knowledge Graph (FoodKG).

Some of the competency questions (i.e., the questions that help capture the scope, content, and the form of evaluation of the knowledge that is modeled) include questions such as: "What are the ingredients and the total calorie count of a piece of a chocolate cake according to USDA[3] nutritional data?". The answer may include butter, eggs, sugar, flour, milk, and cocoa powder for the ingredients, and a calorie count of 424. For a diabetic who is trying to abide by the ADA guidelines, a question like, "How can I increase the fiber content of this cake?" may be a natural follow-up question to ask. Similarly, a person suffering from lactose intolerance may ask "What can I substitute for milk in chocolate cake?". Answering questions like this is not possible from sources such as DBpedia[4] alone, because the information from those sources is not complete. For example, the *dbo:ingredients*[5] for the resource *dbr:Chocolate_cake*[6] contains only *dbr:Cocoa_powder* and *dbr:Chocolate*. FoodKG contains additional information from online recipe sites, along with the corresponding nutrient information from USDA, that has more relevant information than what is available on DBpedia. Therefore, to answer this question, we can use the semantic structure of our knowledge graph to suggest that whole wheat flour be used instead of white flour, or that soy or almond milk be used instead of cow's milk, or that margarine be used instead of butter.

To address questions like those posed above, we present a methodology that can be used to extract publicly available data on food and construct a semantically meaningful knowledge graph that can power applications to help consumers understand their foods and discover substitutions.

2 Related Work

Ontologies representing food are a well-studied topic. The Food Ontology is a universal "farm to fork" food vocabulary [9] that covers the provenance of food contained within the ontology. However, FoodOn lacks nutrition information and recipes, which is our focus. The Personalized Information Platform for Health and Life Services (PIPS) is a large-scale European Union project dedicated to the development of new ways to deliver healthcare [5]. It describes a food ontology that incorporates nutritional information such that it can be

[3] USDA refers to US Department of Agriculture. `https://www.usda.gov`

[4] DBpedia [2] has structured content from the information created in the Wikipedia.

[5] The *dbo* prefix refers to `http://dbpedia.org/ontology` and *dbo:ingredient* dereferences to `http://dbpedia.org/ontology/ingredient`.

[6] The *dbr* prefix refers to `http://dbpedia.org/resource` and *dbr:Chocolate_cake* dereferences to `http://dbpedia.org/resource/Chocolate\_cake`.

applied to help manage different health conditions like diabetes. A similar ontology is described in [7] for use by hypertensive individuals. The Healthy Life Style (HeLiS) Ontology includes a subportion focused on food, including concepts such as 'BasicFood' and 'Recipe' [10]. It aligns well with our own goal, although it has a somewhat reduced scope. The Food Product Ontology [16] is designed for business purposes. It includes concepts such as price and brand, which is more suitable for food suppliers than end users. The Cooking Ontology [3] comprises four main classes–actions, foods, recipes, and utensils– with supplementary class units, measures, and equivalencies, and the ontology is integrated into a dialogue system to answer the questions. However, they currently do not support a version in English, and have not mapped to comparable classes in other ontologies, which is essential for reuse. Similarly, the BBC Food Ontology (`https://www.bbc.co.uk/ontologies/fo/1.1`) only constructs the important concepts and needs to cooperate with other existing ontologies to work better. The SmartProducts Network of Ontologies (`http://projects.kmi.open.ac.uk/smartproducts/ontology.html`) also contains a food ontology, however our USDA nutrients ontology has more than twice as many food items as in their food_nutrients.owl ontology.

The FOod in Open Data (FOOD) [20] project implements existing ontology designs for foods that are designated as "protected" in the European Union, and then extracts data contained in the Italian agricultural policy documents to produce Linked Open Data (LOD) for public use. However, they focus on characteristics important for policy evaluation and enforcement, rather than for health. Other systems include an information retrieval system that incorporates knowledge from domains of food, health, and nutrition, to recommend food health information based on the users' conditions and preferences is described in [14], and the food search through knowledge graphs [26] focuses on the user's ratings and opinions on tapas/pintxos (small bites/dishes). Finally, the FOODS-Diabetes ontology [22] is meant for medical providers to plan patient meals in terms of caloric intakes, etc., and does not include any recipes or ingredients.

The "internet of food" review [4], and the LOV4iot project [13] (`http://lov4iot.appspot.com/?p=lov4iot-food`) list a number of other food related ontologies. Different food ontologies focus on different aspects of food, such as chemical compositions, supermarket locations, food sources/packaging, and so on. Our focus is on recommendation in the context of personalized health, i.e., suggesting similar or alternative foods and recipes that are more healthy.

3 Data Acquisition

The resource contribution introduced in this paper aims to bridge the gaps between silos of data. However, gathering and integrating data from many sources leads to several challenges with consistency, accuracy, and completeness:

- *Invalid data* - some textual data contains characters that are illegal in an RDF based knowledge graph, requiring escaping. Escaping itself can pose

problems for entity recognition and resolution; it must be applied consistently at all stages of the process.
- *Incomplete data* - recipes may lack quantities for ingredients, or provide non-standard units of measure (e.g. *"to taste"*, *"as needed"*, *"a few shakes"*). Nutrient data might be incomplete, with only some nutrients tabulated.
- *Ambiguous entities* - many ingredients are difficult to tie to a specific food item. This has several root causes, such as local spellings and spelling errors; local names and synonyms; and use of different languages. This can lead to a large number of equivalent names, for example, *corn masa, masa harina, corn flour*.
- *Extraneous information* - ingredients are occasionally listed with complicated units (e.g. *1/3 of a 375g can of beans*) or unnecessary information (e.g. *black beans from the store*).

Our FoodKG relies on three main sources of data: the recipes themselves, the nutritional content of ingredients, and a food ontology to organize the ingredients. We discuss these sources below.

Recipes Online recipe sites allow users to browse and share recipes. Some display content from specific commercial sources; others permit users to upload their own recipes. Each website has specific conventions for how data is presented. In some cases, this includes an effort to provide *machine-readable data*.

There also exist large collections of recipe data produced for research and commercial purposes. An example of the former is the Recipe1M dataset[7], provided by the authors of *Im2Recipe* [18], and consists of over 1 million recipes collected from various internet recipe sharing sites.

Nutrients We chose to use USDA's *National Nutrient Database for Standard Reference* (`https://catalog.data.gov/dataset/food-and-nutrient-database-for-dietary-studies-fndds`), which contains approximately 8,000 records for a variety of types of food and their nutrients. The majority of the foods are generic, rather than coming from a specific brand. Whilst by no means exhaustive, the dataset provides a large variety of foods with extensive nutritional information.

Food Ontologies Lists of recipes and nutritional tables provide bulk information about millions and thousands of entities, respectively, but suffer from a lack of meaning - these components form a strong *knowledge graph*, but lack an *ontology*. To resolve this, we incorporate relevant portions of the FoodOn ontology [9]. FoodOn provides an extensive taxonomy for foods, organizing them by source organism, region of origin, and so forth. This provides much-needed connections between related concepts. For example, gala apples are siblings of red apples, but are further removed from apple pie. However, it was not designed as a nutritional reference, and thus FoodOn lacks detailed information about the nutritional content of items. It also does not directly relate to real-world recipes.

[7] The Recipe1M dataset is available for download after signing up at: `http://im2recipe.csail.mit.edu/dataset`

Since FoodOn is a very large taxonomy, we opted to use only a small subset of it. To accomplish this, we leveraged Ontofox, a tool that extracts terms and axioms from ontologies [25]. Using the tool, we extracted all children of the ***food product by organism*** node, thus capturing a wide variety of food items in a useful hierarchical form (providing a breakdown by *category of organism*, *group of organism*, and finally a specific *organism of origin*).

4 Knowledge Graph Construction

A knowledge graph includes resources with attributes and entities, relationships between such resources, and annotations to express metadata about the resources. Our complete food knowledge graph contains several key components:

i) Recipes and their ingredients, ii) Nutritional data for individual food items, iii) Additional knowledge about foods, and iv) Linkages between the above concepts.

Recipes Each recipe describes the ingredients needed to produce a dish. Each recipe receives a unique identifier, which is accompanied by its name, any provided tags, and a set of ingredients. Each ingredient points to its name, unit, and quantity. An example of the resulting structure is provided in Figure 1.

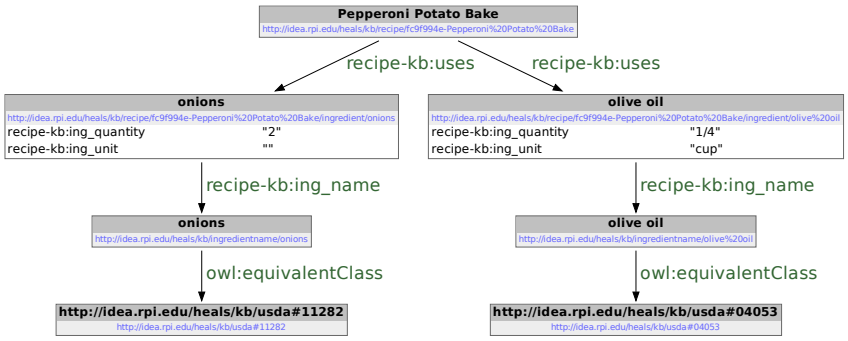

Fig. 1. An example of an imported recipe, pruned to show only two ingredients. The ingredients have been linked to USDA records.

Individual ingredient records usually appear in the form of *(quantity, unit, name)*, such as *2 cups flour* or *1 1/2 lb cabbage, chopped*. Due to the lack of context, parsing these phrases with natural language processors is difficult, and naive parsing methods fail due to minor quirks. To effectively parse such records, we utilize the following steps: 1) Parenthesized statements, such as *(freshly picked)* or *(or chicken)*, are stripped. These provide additional cues to the reader, but are not strictly necessarily to understand components that make up the recipe. Similarly, any text following the first comma is dropped, as it generally describes additional qualities for the ingredient. Whilst these do have meaning, it is less significant than that of the name itself. 2) Numerical values, such as *1/2* or *2.5*,

are removed from the start of the string and saved as the *quantity*. The numerical value for an ingredient is, in almost all cases, found before the unit and name of the ingredient. 3) A list of units is compared against the first word in the string; if one matches, it is removed and stored as the *unit*. As when finding the quantity, this almost always succeeds; it is highly uncommon for the unit to be found anywhere but immediately after the quantity. 4) The remaining text is tokenized with the *Natural Language Toolkit* (NLTK; `https://www.nltk.org`). Adjectives that are not descriptive of color are eliminated. For example, names such as *green bell peppers* and *red onion* are preserved, whilst descriptors like *fresh* are eliminated. Verbs and adverbs are also eliminated, simplifying terms like *diced onion* and *minced garlic*. Text following a conjunction is removed. Finally, NLTK's *WordNetLemmatizer* is used to eliminate plurals. The resulting text is then saved as the *name*. Examples of inputs and results are provided in Table 1. High-quality name recognition significantly improves the quality of later results.

Input	Quantity	Unit	Name
1 cup milk	1	cup	milk
1 tablespoon parsley, chopped	1	tablespoon	parsley
6 tablespoons red currant jelly	6	tablespoons	red currant jelly
1 cup butter, softened	1	cup	butter

Table 1. Examples of processed ingredient data

Nutrients From recipes, we can produce a network of foods and their ingredients. However, without information about the nutritional content of each ingredient, we cannot make meaningful health-related suggestions. We use the USDA public nutrition dataset for this information. The data from USDA exists in a tabular form, describing several dozen nutritional statistics, such as calories, macro-nutrients (protein, carbohydrates, fats), and micro-nutrients (vitamins and minerals). Nutrients are provided per 100 grams of the food item. Two non-mass measurements of the food are also provided, along with the number of grams found in each measure. We make use of the Semantic Data Dictionary approach [21], which produces RDF triples from non-triple data sources. This turns our tabular data into something that can be integrated into our knowledge graph. Some examples of the data converted to the concepts in the knowledge graph can be seen in Table 2.

id	description	water	energy	protein	lipid	carbohydrate
1001	Butter, With Salt	15.87	717	0.85	81.11	0.06
1009	Cheese, Cheddar	37.1	406	24.04	33.82	1.33

Table 2. Example USDA data with a few food items and 5/57 nutrients.

Given this data, we can define the shape of the resulting knowledge graph via semantic relationships, as can be seen in Table 3: **Column** represents the column in the raw data, **Attribute/Entity** represents what *rdf:type* this food

item is, **Unit** refers to the unit of measurement for that nutrient from community accepted terminologies such as DBpedia and the Units Ontology, and **Label** gives a textual description for the data item that can be used in text mining and auto-completion tasks in applications that use the FoodKG.

Notice the interlinking to other ontologies in then **Attribute/Entity** and the **Unit** columns. The various prefixes[8] in the annotations in Table 3 points to the following ontologies:

- *chebi*: Chemical Entities of Biological Interest Ontology (`https://www.ebi.ac.uk/chebi`)
- *dbr*: DBpedia Resource Ontology (`http://dbpedia.org/resource/`)
- *sio*: Semanticscience Integrated Ontology (`http://semanticscience.org/resource/`)
- *envo*: Environment Ontology (`http://purl.obolibrary.org/obo/envo.owl#`)
- *foodon*: Food Ontology (`http://purl.obolibrary.org/obo/foodon.owl#`)
- *schema*: Schema.org mappings (`https://schema.org/`)
- *uo*: Units Ontology (`http://purl.obolibrary.org/obo/uo.owl#`)

Column	Attribute/Entity	Unit	Label
id	chebi:33290, dbr:Food		USDA Id for the food
description	sio:StatusDescriptor		Short description
water	envo:00002006, chebi:15377, dbr:Water	dbr:Gram, uo:0000021	Water (g)
energy	foodon:03510045	dbr:Kcal	Energy (Kcal)
protein	dbr:Protein	dbr:Gram, uo:0000021	Protein (g)
lipid	dbr:Lipid	dbr:Gram, uo:0000021	Lipid Total (g)
carbohydrate	dbr:Carbohydrate, schema:carbohydrateContent	dbr:Gram, uo:0000021	Carbohydrate (g)
sugar	dbr:Sugar	dbr:Gram, uo:0000021	Sugar Total (g)
calcium	dbr:Calcium	uo:0000022	Calcium (mg)

Table 3. Semantic structural representation of a subset of the USDA data.

After these annotations are completed, the semantic data dictionary conversion script is run to convert the tabular USDA data into quads, which are triples grouped into named graphs. A small piece of the high level structure of the resulting graph can be seen in Figure 2.

5 Knowledge Graph Augmentation

With all of the data imported, we are left with a collection of isolated islands of data. Thus, the second phase of the construction of our knowledge graph is *linkage*. We leverage various entity resolution techniques to automatically connect various concepts together. To ensure that the dataset can be practically

[8] Prefixes can be dereferenced via `http://prefix.cc` or `http://www.ontobee.org`.

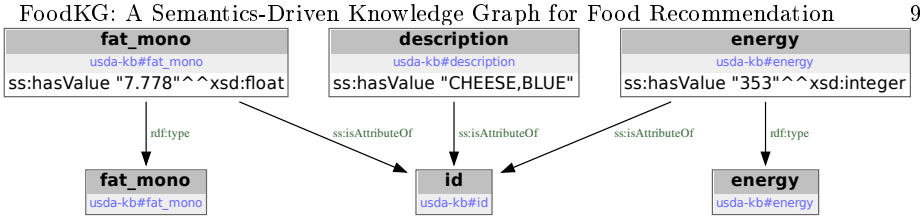

Fig. 2. An example of USDA data, pruned to display a handful of features. The prefixes *usda-kb* and *ss* refers to custom namespaces within our knowledge graph.

expanded and updated, we make use of well-studied linked data techniques to establish provenance of these derived relationships.

Entity Resolution Names are the most obvious shared attributes between our various domains of recipes, nutrients, and foods. For this reason, we have largely focused on entity resolution techniques that work on strings, such as *cosine similarity*, which performs quite well for matching by name, particularly after normalization. We also examined using word embeddings, such as word2vec [19] and FastText [15], with a pretrained model to resolve names. However, results were poor - likely a product of the embedding capturing only the general meaning of a statement.

Entity Selection We found it beneficial to limit the domain of concepts to match against, both for the sake of performance (matching is linearly expensive with respect to the number of entities) and to maximize accuracy (more spurious entities to match against cause more false positives). The exact manner in which this is done depends on the datasets being compared.

For instance, many categories of food are rarely seen as ingredients - but, critically, have names that are similar to kinds of food that *are* relevant. The USDA's Standard Reference contains a large number of entries about baby food, with names such as 'Babyfood, juice, apple' and 'Babyfood, meat, lamb, strained'. We remove such entries, since they cause problems with linkage of ingredients. For example, the former will match the 'apple juice' ingredient in a recipe, but is it unlikely that the recipe is referring to babyfood. We similarly remove categories such as fast food and sweets - although even this is not entirely straightforward. For example, "brown sugar" is lumped in with jelly beans and candy bars, but it is desirable to retain it in the FoodKG. We also ignore text beyond the third comma, as we found that the distinctions between entities becomes insignificant at that point; doing so also speeds up the linkage process.

Other sources of data are significantly broader; as an example, we experimented with linking into the DBpedia knowledge graph. Unfortunately, many entities in the DBpedia dataset are incompletely or inaccurately typed; non-foods have the Food type, and many edible items lack it. Therefore, we opted to use heuristics to select for potential ingredients. All DBpedia resources marked as `ingredientOf` were included, as was anything with a *carbohydrate* value. This tended to produce a subset of actual food items, and whilst it resulted in the loss of some entities, it also eliminated a large number of erroneous choices.

Provenance and Publication Information To provide clear provenance for every claim made in our knowledge graph - including both imported knowledge *and* inferred linkages - we have made extensive and consistent use of the RDF Nanopublication specification [12]. Nanopublications represent atomic units of publishable information, attaching information about *where* it came from and *who/what* published it. They express this knowledge via linked data, using four named graphs:

i) The *assertions* graph contains the claims being made. For a recipe, this includes a title, tags (if any), and ingredients. Ingredients are described by their name, unit, and quantity. ii) The *provenance* graph contains information about where the assertions were derived from. For a recipe, this could be the URL from which data was retrieved, or any other reference that points back to the original data. iii) The *publication info* graph explains who created the nanopublication. For example, the linkages we form are collected into a single nanopublication; the publication info remarks that our linker tool generated the linkages. iv) The *head* graph ties the prior three graphs together, making it possible to find the three components.

6 Application of the Food Knowledge Graph

6.1 Answering Competency Questions in SPARQL

In order to evaluate the knowledge graph, we establish several competency questions that address its possible applications. Our first competency question is "What recipes contain beef?" which would return a list of all recipes in the knowledge graph that are linked to some entity 'beef' in the knowledge graph. This covers the simple case of understanding what ingredients are found in what recipes in the knowledge graph. The second competency question, "What recipes contain beef, carrots, and potatoes?" takes this a step further by asking for recipes that contain multiple ingredients. This type of question mimics the functionality of traditional recipe sharing websites, where users can look for recipes containing certain types of ingredients. Our third competency question is "What recipes contain bananas that do not contain walnuts?" as can be seen in Listing 1.1. This evaluates the ability of our knowledge graph to return recipes that, in addition to containing certain ingredients, do not contain others. This is especially relevant in cases of allergies or dislikes of certain foods. We can further extend this kind of thinking to nutritional information, based on the knowledge graph's health information from the USDA. This brings us to our fourth competency question, "What recipes that have chicken are low in sugar?" as can be seen in Listing 1.2. This and similar kinds of questions address the application of the knowledge graph to assisting with certain health conditions like diabetes or hypertension that place restrictions on nutritional intake. Currently, this question is answered by using glycemic index information which was manually added by hand to certain ingredients. This approach clearly has its limitations, however, since not all ingredients have a glycemic index and ingredient amounts are not considered in this calculation. Our final competency question is "What recipes

are vegan?" Since the knowledge graph structures its knowledge of ingredients in a hierarchical way, it can determine whether certain ingredients fall into certain categories like animal products for vegetarian/vegan diets or pork products for religious restrictions, as a more specific example.

Each of these questions can be answered by querying the underlying ontology using SPARQL, since information like the relationships between recipes and ingredients is encoded directly within the ontology. An example query for the third and fourth competency questions are structured as follows.

```
@PREFIX food: <http://purl.org/heals/food/>
@PREFIX ingredient: <http://purl.org/heals/ingredient/>
SELECT DISTINCT ?recipe
WHERE {
    ?recipe food:hasIngredient ingredient:Banana .
    FILTER NOT EXISTS {
        ?recipe food:hasIngredient ingredient:Walnut .}
}
```

Listing 1.1. SPARQL query for retrieving a food for a person with an allergy

```
@PREFIX food: <http://purl.org/heals/food/>
@PREFIX ingredient: <http://purl.org/heals/ingredient/>
SELECT distinct ?recipe
WHERE {
    ?recipe food:hasIngredient ingredient:Chicken .
    FILTER NOT EXISTS{
        ?recipe food:hasIngredient ?ingredient .
        ?ingredient food:hasGlycemicIndex ?GI .
        FILTER (?GI >= 50)}
}
```

Listing 1.2. SPARQL query for retrieving a food with a low glycemic index

6.2 Answering Competency Questions in Natural Language

We demonstrate another potential use of our FoodKG for answering natural language questions over knowledge graphs, aka, knowledge base question answering (KBQA). Given questions in natural language, our goal here is to automatically find answers from the underlying knowledge graph.

Since there does not exist a Food Q&A dataset related to ingredients, nutrients and recipes, we choose to create a synthetic Q&A dataset based on our FoodKG using a set of manually designed question templates. We create three types of competency questions with increasing levels of complexity using various templates. Table 4 shows the question templates and data statistics of the created dataset. The simple questions, e.g., "How much sugar is in cheese, cream, fat free?", are created based on the USDA data and require only one hop reasoning.

The comparison questions, e.g., "Salt, table or syrups, table blends, pancake, which has less energy?", can be regarded as a composition of two simple questions. The third type of questions we create are those with constraints, e.g., "What Laotian dishes can I make with sugar, water, oranges?"; these queries are based on the Recipe1M data and are similar to those in Section 6.1. To create the dataset, we first sample several subgraphs from the FoodKG. For each subgraph, we then randomly sample a question template from our predefined template pool and fill the slots with KG entities and relations.

Competency Questions	Question Template Examples	Size	Knowledge Source
Simple	How much {nutrient} is in {ingredient}?	12,661	USDA
Comparison	{ingredient1} or {ingredient2}, which has less {nutrient}?	5,565	USDA
Constraint	What {tag} dishes can I make with {ingredient_list}?	6,359	Recipe1M

Table 4. Data statistics of the created synthetic Q&A dataset.

Experiments Our Q&A system consists of three components which are the *question type classifier*, *topic entity predictor* and *KBQA model*. Given a natural language question, e.g., "how much sugar is in Cheese, Blue?", the *question type classifier* is intended to determine the question type which is 'simple' in this case. Then the *topic entity predictor* is applied to detect the topic entity mentioned in the question which is 'Cheese, Blue' and links it to the FoodKG. Finally, the *KBQA model* is called to retrieve answers from the KG subgraph surrounding the topic entity 'Cheese, Blue'.

In our experiments, we only evaluate the *KBQA model* which is the most crucial component in our Q&A system. We compare a simple Bag-of-Word vectors based method (BOW) and our state-of-the-art neural network-based method (BAMnet) [6]. In both methods, we encode the question and each candidate answer within the KG subgraph surrounding the topic entity into the same embedding space, and then compute the cosine similarity score between them using a dot product. Candidate answers whose similarity scores are above a certain threshold are returned as predicted answers. The major difference between the two methods is that in BOW, the question and candidate answers are encoded independently as the average of the pretrained word embeddings, while in BAMnet, a more sophisticated neural network module is used to encode them jointly by considering the two-way flow of interactions between them. For more details, please refer to [6]. We split the dataset into training (50%), development (20%) and test set (30%).

Methods	Simple	Comparison	Constraint	Overall
BOW	13.7	49.6	30.0	26.0
BAMnet	**99.8**	**100.0**	**82.6**	**95.5**

Table 5. Experimental results (F1-scores) on the synthetic Q&A test set.

Table 5 shows the results of two methods on the synthetic Q&A test set where we assume that gold topic entities and question types are known beforehand. As we can observe, even though the questions are created using predefined templates and there is no lexical gap between the questions and the KG (i.e., we use the exact entity and relation names to fill the question templates), the BOW method does not perform well. However, our BAMnet method perform very well on this dataset. Moreover, among the three types of questions, those with constraints are the most challenging. Future directions include creating more realistic and complex questions with more diverse templates and lexical gap.

7 Resource

Our FoodKG resource website at **`https://foodkg.github.io`** links to all the resources, which include the FoodKG knowledge graph, the automated scripts to construct the KG, the whattomake application, the natural language querying application, and accompanying documentation. Using the FoodKG, we can answer complex questions related to recipes, ingredients, nutrition and food substitutions that can power applications that target healthy lifestyle behaviors. The SPARQL queries and the source code for the two applications illustrated in Section 6 are also made available.

Maintenance: The FoodKG is part of the RPI-IBM Health Empowerment through Analytics Learning and Semantics (HEALS) project (`https://idea.tw.rpi.edu/projects/heals`), and we expect to support this project actively for the next 3-7 years. We anticipate that these public tools will be useful for anyone aiming to build an integrated knowledge graph for food. As observed in Fig. 3, our resulting FoodKG spans over 67 million triples (obtained by adding all of the triples comprising the USDA, Recipe1M and FoodOn KG subsets, and the linkages between them). Various other statistics are also shown in the figure.

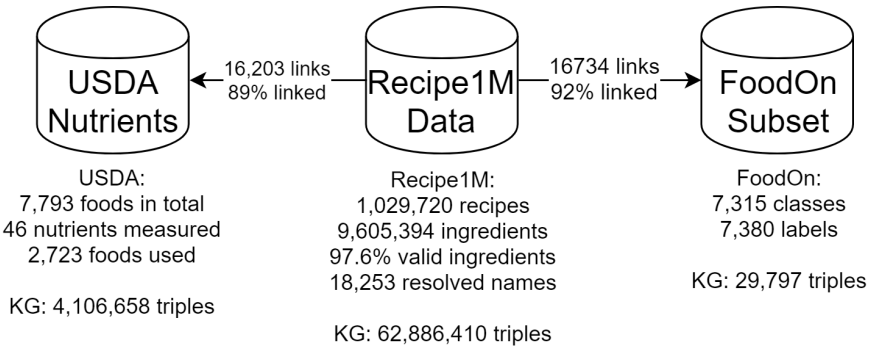

Fig. 3. An overview of the food knowledge graph (FoodKG).

Description: An overview of how to construct the FoodKG with provenance is clearly explained at `https://foodkg.github.io/foodkg.html`. The FoodKG github repository (`https://github.com/foodkg/foodkg.github.io`) contains step-by-step instructions to generate the entire FoodKG, resulting in serialized RDF triples. As outlined in Section 4, the input data is in various formats (e.g., USDA is in CSV, Recipe1M data is in JSON, and other ontologies in RDF/OWL), which we map to RDF. The output of the KG construction is RDF; more specifically, the output is in the NanoPublications format [12], which includes the corresponding assertion, provenance, and publication information, as outlined in Section 5. The output of the scripts include the following serialized RDF files (in .trig format) spanning 67 million triples: i) usda-links.trig, ii) foodon-links.trig, iii) foodkg-core.trig. We do not directly provide the final RDF data, due to the terms of the Recipe1M data. However, our Github code and step-wise instructions can generate the KG exactly as described herein. We believe that generating a KG programmatically for a food knowledge graph has several benefits over supporting a public SPARQL endpoint or a compressed dump of the graph: (1) additional means of enriching the KG programmatically, (2) possibility to tap into various sources of data, (3) clean handling of intellectual property in the ever-changing and complex rights management landscape.

The whattomake app (`https://foodkg.github.io/whattomake.html`), described in Section 6.1, includes comprehensive documentation, sample SPARQL queries, and three food resources: i) `http://purl.org/heals/foodon` (a subset of the FoodOn we used in our mappings), ii) `http://purl.org/heals/food`, and iii) `http://purl.org/heals/ingredient`.

Finally, the KBQA application (`https://foodkg.github.io/kbqa.html`) includes documentation on how to query the FoodKG using natural language questions. We currently support three types of questions, namely simple, comparison and constraint-based as described in Section 6.2.

8 Conclusions and Future Work

It is evident that information on food, while readily available on the Web, requires individuals to combine information from various sources in order to decide what to eat. To address the issue of aggregating all the pertinent information on food in a manner that is consumable by an individual specific to their health and taste preferences, we have created an integrated knowledge graph for food, which can be used to suggest healthier food and restaurant menu item alternatives. We model structured sources in terms of a target ontology, and augment the knowledge graph with other unstructured sources.

More specifically, we extract the relevant data on food from authoritative sources such as the USDA, as well as online recipe sources. We apply a semantics based extract-transform-load procedure to structure the food knowledge using our ontology as well as community accepted terminologies, and link to relevant FoodOn and nutrient resources to support further exploration and augmentation of the FoodKG. The linkages to these resources are done using techniques involv-

ing lexical similarity and string matching to find non-perfect matches between sets of data that frequently lack perfect pairings.

Our FoodKG is a valuable resource for the primary task of food recommendation. At the same time, it can also be used as a benchmark dataset to test various entity resolution and semantic linking methods for recipes, ingredients, units, and so on. In the future, we plan to further leverage the food knowledge graph and relationships between ingredients and recipes to develop novel ingredient and recipe embedding models to produce more meaningful representations for food recommendation. Since our ultimate objective is to provide personalized food recommendations to everyday individuals that consider both their health and lifestyle preferences, we see the need for the food knowledge graph to support competency questions that involve more subjective concepts like 'convenient', 'affordable', 'spicy', and 'refreshing'. We also plan to continue to extend our ontology and knowledge sources, as well as explore novel food embeddings that leverage the relationships captured in the food knowledge graph. In conclusion, we have presented a reusable methodology that integrates information on food into a knowledge graph.

Acknowledgements

This work is partially supported by IBM Research AI through the AI Horizons Network.

References

1. American Diabetes Association: 4. lifestyle management. Diabetes Care **40**(Supplement 1), S33–S43 (2017)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The semantic web. Springer (2007)
3. Batista, F., Pardal, J.P., Mamede, P.V.N., Ribeiro, R.: Ontology construction: cooking domain. Artificial Intelligence: Methodology, Systems, and Applications **41**, 1–30 (2006)
4. Boulos, M., Yassine, A., Shirmohammadi, S., Namahoot, C., Brückner, M.: Towards an "internet of food": food ontologies for the internet of things. Future Internet **7**(4), 372–392 (2015)
5. Cantais, J., Dominguez, D., Gigante, V., Laera, L., Tamma, V.: An example of food ontology for diabetes control. In: ISWC workshop on Ontology Patterns for the Semantic Web (2005)
6. Chen, Y., Wu, L., Zaki, M.J.: Bidirectional attentive memory networks for question answering over knowledge bases. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics (2019)
7. Clunis, J.: Designing an ontology for managing the diets of hypertensive individuals. International Journal on Digital Libraries pp. 1–16 (2018)
8. DeSalvo, K., Olson, R., Casavale, K.: Dietary guidelines for americans. JAMA **315**(5), 457–458 (2016)

9. Dooley, D.M., Griffiths, E.J., Gosal, G.S., Buttigieg, P.L., Hoehndorf, R., Lange, M.C., Schriml, L.M., Brinkman, F.S., Hsiao, W.W.: Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Science of Food **2**(1), 23 (2018)
10. Dragoni, M., Bailoni, T., Maimone, R., Eccher, C.: Helis: an ontology for supporting healthy lifestyles. In: International Semantic Web Conference (2018)
11. El-Dosuky, M., Rashad, M., Hamza, T., El-Bassiouny, A.: Food recommendation using ontology and heuristics. In: International conference on advanced machine learning technologies and applications (2012)
12. Groth, P., Gibson, A., Velterop, J.: The anatomy of a nanopublication. Information Services & Use **30**, 51–56 (2010)
13. Gyrard, A., Bonnet, C., Boudaoud, K., Serrano, M.: Lov4iot: A second life for ontology-based domain knowledge to build semantic web of things applications. In: 4th IEEE International Conference on Future Internet of Things and Cloud (2016)
14. Helmy, T., Al-Nazer, A., Al-Bukhitan, S., Iqbal, A.: Health, food and user's profile ontologies for personalized information retrieval. Procedia Computer Science **52**, 1071–1076 (2015)
15. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: 15th Conference of the European Chapter of the Association for Computational Linguistics (2017)
16. Kolchin, M., Zamula, D.: Food product ontology: Initial implementation of a vocabulary for describing food products. In: 14th Conference of Open Innovations Association (2013)
17. Ley, S.H., Hamdy, O., Mohan, V., Hu, F.B.: Prevention and management of type 2 diabetes: dietary components and nutritional strategies. The Lancet **383**(9933), 1999–2007 (2014)
18. Marin, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., Torralba, A.: Recipe1m: A dataset for learning cross-modal embeddings for cooking recipes and food images. arXiv preprint arXiv:1810.06553 (2018)
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
20. Peroni, S., Lodi, G., Asprino, L., Gangemi, A., Presutti, V.: Food: food in open data. In: International Semantic Web Conference. pp. 168–176. Springer (2016)
21. Rashid, S.M., Chastain, K., Stingone, J.A., McGuinness, D.L., McCusker, J.P.: The Semantic Data Dictionary Approach to Data Annotation & Integration. 1st Workshop on Enabling Open Semantic Science (2017)
22. Snae, C., Bruckner, M.: Foods: a food-oriented ontology-driven system. In: 2nd IEEE International Conference on Digital Ecosystems and Technologies (2008)
23. Trattner, C., Elsweiler, D.: Food recommender systems: important contributions, challenges and future research directions. arXiv preprint arXiv:1711.02760 (2017)
24. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016)
25. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: Ontofox: web-based support for ontology reuse. BMC Research Notes **3**(1), 175 (Jun 2010)
26. Zulaika, U., Gutiérrez, A., López-de Ipiña, D.: Enhancing profile and context aware relevant food search through knowledge graphs. In: 12th International Conference on Ubiquitous Computing and Ambient Intelligence (2018)