FoodKG: A Semantics-Driven Knowledge Graph for Food Recommendation[1]

Steven Haussmann[1], Oshani Seneviratne[1], Yu Chen[1], James Codella[2], Ching-Hua Chen[2], Deborah L. McGuinness[1], Mohammed J. Zaki[1],
[1]Rensselaer Polytechnic Institute, Troy, NY; [2]IBM Research, Yorktown Heights, NY

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 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. We describe a unified food knowledge graph that links the various silos related to food while preserving the provenance information.

1 Introduction

Knowledge graphs (KGs) have an essential role in organizing the information we encounter on a day-to-day basis and making it more broadly available to both humans and machines. However, the elusiveness of standards or 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. Integrating data from many sources leads to several challenges with consistency, accuracy, and completeness. For example, in the domain of food, 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. Furthermore, there are ambiguous entities; many ingredients are difficult to tie to a specific food item due to synonyms, regional names, and use of different languages.

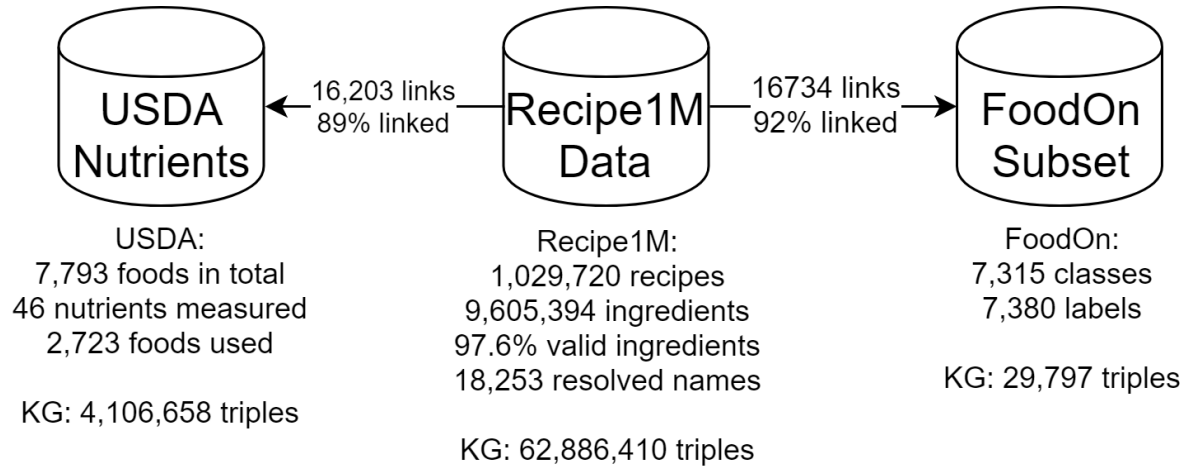

Figure 1: An overview of the food knowledge graph (FoodKG).

2 Knowledge Graph Construction

Given the challenges in the input data in the food domain, we utilized several machine learning techniques to create a knowledge graph in RDF. Our complete food knowledge graph contains several key components, namely: i) Recipes and their ingredients, ii) Nutritional data for individual food items, iii) Additional knowledge about foods, and iv) Linkages between the above concepts.

[1]This paper is a knowledge representation and semantics highlight. This work was originally accepted for presentation at the International Semantic Web Conference 2019.

We extracted the relevant data on food from authoritative sources such as the USDA (`https://www.usda.gov`), online recipes available in the Recipe1M dataset[1], and the FoodOn (`https://foodon.org`). We then applied a semantics based extract-transform-load procedure[2] to structure the food knowledge using our ontology as well as community accepted terminologies, and linked to relevant FoodOn and nutrient resources to support further exploration and augmentation of the FoodKG. The linkages to these resources are done using techniques involving lexical similarity and fuzzy matching to find non-perfect matches between sets of data that frequently lack perfect pairings.

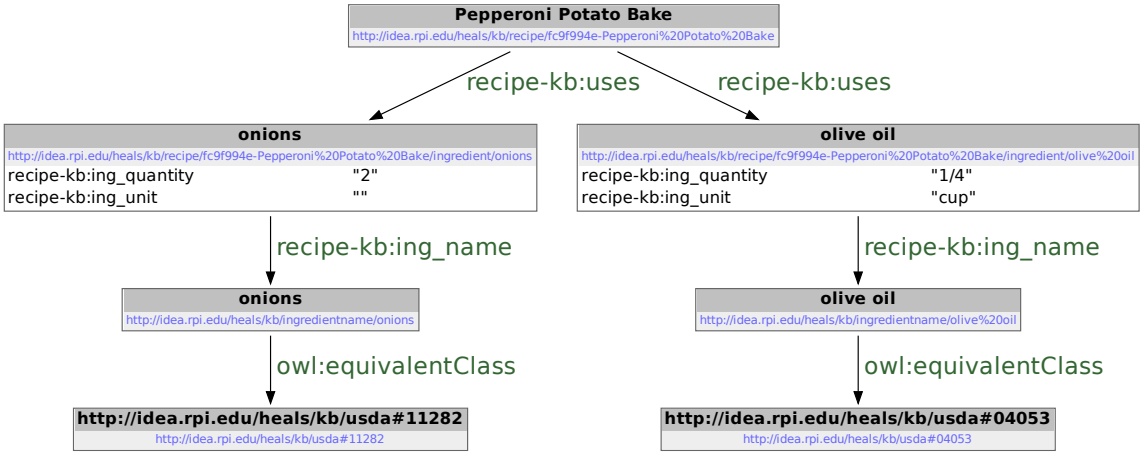

Figure 2: An example of an imported recipe, pruned to show only two ingredients. The ingredients have been linked to USDA records.

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*. In terms of 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 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. 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[3].

3 Conclusion

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. Using this knowledge graph, we can power applications that target healthy lifestyle behaviors to answer complex questions related to recipes, ingredients, nutrition, and food substitutions. For more information on this work, please refer to our website **`https://foodkg.github.io`**.

Acknowledgements

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

References

1. Marin J, Biswas A, Ofli F, Hynes N, Salvador A, Aytar Y, et al. Recipe1M: A dataset for learning cross-modal embeddings for cooking recipes and food images. arXiv preprint arXiv:181006553. 2018;.

2. Rashid SM, Chastain K, Stingone JA, McGuinness DL, McCusker J. The Semantic Data Dictionary Approach to Data Annotation & Integration. In: SemSci@ ISWC; 2017. p. 47–54.

3. Groth P, Gibson A, Velterop J. The anatomy of a nanopublication. Information Services & Use. 2010;30(1-2):51–56.