

Building FKG.in: a Knowledge Graph for Indian Food

Saransh Kumar Gupta^{1,*†}, Lipika Dey^{1,*†}, Partha Pratim Das^{1,*†} and Ramesh Jain^{2†}

¹Ashoka University, India

²Institute of Future Health, UC Irvine, USA

Abstract

This paper presents an ontology design along with knowledge engineering, and multilingual semantic reasoning techniques to build an automated system for assimilating culinary information for Indian food in the form of a knowledge graph. The main focus is on designing intelligent methods to derive ontology designs and capture all-encompassing knowledge about food, recipes, ingredients, cooking characteristics, and most importantly nutrition, at scale. We present our ongoing work in this workshop paper, describe in some detail the relevant challenges in curating knowledge of Indian food, and propose our high-level ontology design. We also present a novel workflow that uses AI, LLM, and language technology to curate information from recipe blog sites in the public domain to build knowledge graphs for Indian food. The methods for knowledge curation proposed in this paper are generic and can be replicated for any domain. The design is application-agnostic and can be used for AI-driven smart analysis, building recommendation systems for Personalized Digital Health, and complementing the knowledge graph for Indian food with contextual information such as user information, food biochemistry, geographic information, agricultural information, etc.

Keywords

Food Computing, Ontology Design, Knowledge Engineering, Semantic Reasoning, Nutrition Informatics, Large Language Models

1. Introduction

Food is playing an increasingly central role in health and sustainability discourses as the preservation of diverse cultures, food security, precision nutrition, personal and public health, agricultural practices, climate impact, and supply chains become focal points of discussion. However, any definitive effort in such a scientific pursuit requires well-founded applications to be designed around food and food-related data, with access to knowledge representation and reasoning systems for food. In this light, food knowledge graphs are crucial and reusable digital resources that can capture various nuances of food including but not limited to recipes, ingredients, flavor, texture, cooking techniques, cuisine, nutritional information, and mealtimes. They can be used for various applications like food recommendation, recipe recommendation, diet planning, health tracking, food quality control, managing food supply chains, and so on. While several countries including the US, parts of Europe (like Latvia, Norway, Spain, the UK,


Formal Ontology in Information Systems Conference (FOIS) '24: Integrated Food Ontology Workshop (IFOW)
July 15–19, 2024, University of Twente, Enschede, Netherlands

*Corresponding author.

†These authors contributed equally.

✉ saransh.gupta@ashoka.edu.in (S. K. Gupta); lipika.dey@ashoka.edu.in (L. Dey); partha.das@ashoka.edu.in (P. P. Das); jain49@gmail.com (R. Jain)

ORCID 0000-0003-3831-5545 (L. Dey); 0000-0003-1435-6051 (P. P. Das); 0000-0003-2373-4966 (R. Jain)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Italy, and Portugal), China, and Japan are working on building such knowledge bases for specific regions, there appears to be a vacuum when it comes to Indian food. We intend to bridge this gap to aid food computing initiatives for Indian food.

In this paper, we present our work on building a knowledge graph for Indian food, named FKG.in, which aims to exhaustively cover the panorama of Indian food and act as a digital resource for building subsequent food computing applications over it. The proposed ontology adapts from earlier food ontologies along with modifications and extensions to capture unique aspects of Indian food and is designed in an application-agnostic way. We also propose a novel AI-based semi-automated approach to curate culinary information from multiple websites in the public domain to populate the knowledge graph, along with a human-in-the-loop intervention to ensure the soundness of information.

The rest of the paper is organized as follows. Section 2 presents an overview of the work done in the areas of designing food ontology and building food knowledge graphs, along with earlier efforts in building Indian food knowledge graphs. Section 3 presents the unique challenges associated with the task of consolidating knowledge about Indian food. Section 4 presents the ontology design and its connections to existing food knowledge graphs. Section 5 presents the AI-based technologies adopted to populate a knowledge base for Indian food. Section 6 presents some results. Finally, we conclude with plans to extend the knowledge graph and integrate it with other food computing applications.

2. Related work

Several food knowledge graphs have been constructed based on these ontologies and their extensions, to support food computing applications. We mention only a few representative ones here. For a more comprehensive review of food ontology, food knowledge graphs, and food computing applications, one may refer to the works cited in [1, 2, 3]. In [3], authors have categorized existing food knowledge graphs into four different types - (1) knowledge graphs about recipes, (2) knowledge graphs about nutrients and health, (3) knowledge graphs about food safety, and (4) general food knowledge graphs. We follow the same pattern to group both ontology and knowledge graphs, though there exist many overlaps among the groups.

The first group of ontology mostly focuses on concepts of food related to recipes and cooking. Notable in this group are Table [4], Cooking ontology [5], BBC food ontology¹, and so on. There are a few ontologies dedicated to special categories of food, like Open Food Facts² that model information about packaged foods, Seafood ontology [6], and so on.

Recipe knowledge graphs are built to store recipe entities that are extracted from crowd-sourced consumer review sites, recipe-sharing websites, and social media to primarily support food recommendation systems and build social networks around food. Foodbar knowledge graph [7] is one such system that extracts consumer opinions ratings etc. from different sources and augments this information with information about users, points of interest, cultural facts, and so on. RcpKG [8] is a multimodal and hierarchical food knowledge graph that curates information from popular recipe websites like Yummly and AllRecipes as well as semi-structured datasets like Recipe1M+ [9]. RcpKG also incorporates social relationships into the food knowledge graph

¹<https://www.bbc.co.uk/ontologies/food-ontology>

²<https://world.openfoodfacts.org/data>

for generating food recommendations that can take care of both personal preferences and social relationships. In a unique experiment, cooking is viewed as a uniquely human endeavor for transforming raw ingredients into delicious dishes [10], and it is proposed that recipes can be viewed as cultural capsules that capture culinary protocols. This work is focused on learning the valid protocols for a given set of constraints and thereby generating recipes without violating the cooking principles. This work is also envisaged to generate recipes following the culinary grammar that can be leveraged to improve public health through dietary interventions. The underlying knowledge base is RecipeDB³ [11], which is a structured compilation of recipes, and ingredients along with their nutrition and flavor profiles and health associations.

Several ontologies have been built to cater to the concepts of food, nutrition, and health. Personalized Information Platform for Health and Life Services (PIPS) [12] consists of an abstract model of different types of food along with health and nutrition concepts, targeted at providing nutritional advice for diabetic patients. FOODS [13] also focused on storing information about food and nutrition with an eye toward food or menu planning for people with diabetes. Edamam food ontology⁴ provides concepts related to food, recipes, and nutrition to promote healthy eating through various applications like cooking robots. FoodOn [14] contains a fairly exhaustive list of properties that can be used to describe basic ingredient types coming from animal, plant, or fungal origins, agricultural and animal husbandry practices linked to their growth, also lists of common processed food items, and chemical ingredients along with the processes used to make them and terminology to describe nutritional values. The ontology aims to provide a shared vocabulary that can be used for knowledge exchange across domains like environment, agriculture, animal husbandry, food processing, etc. to ensure food safety and security.

FoodKG [15] is a large-scale and unified food knowledge graph that brings together FoodOn and WhatToMake ontology and contains recipe and nutrient instances extracted from Recipe1M+ as well as nutrient records from the US Department of Agriculture (USDA). This knowledge graph can support a multitude of applications, like recipe recommendations, ingredient substitutions, and Question Answering about nutrition. The Chinese Food Knowledge Graph [16] containing information about Chinese dietary cultural elements and Traditional Chinese Medicine was built to enable knowledge retrieval about health and balanced diets.

Other special-purpose ontologies include AGROVOC [17], which is dedicated to storing agriculture, fisheries, and forestry terminologies related to food. Food Track and Trace Ontology [18], which models knowledge related to the food supply chain, has been designed specially for the food safety domain to help in food traceability. Supply Chain Traceability (SCT) ontology [19] also supports information about critical tracking events (CTEs) to provide unified support to food traceability from logistics to production lines. The Meat Supply Chain Ontology (MESCO) is specialized to support the meat supply chain area. Knowledge graphs built along these lines include the Food safety knowledge graph [20] and the Food spot-check knowledge graph [21] which are mainly concerned with food safety issues. The Food Safety Knowledge Graph supports a Question Answering application to answer user queries about unqualified foods, based on official information released on the Internet. Food spot-check supports a similar application based on data released on the Internet about spot-checks.

³<https://cosylab.iiitd.edu.in/recipeDB/>

⁴<https://www.edamam.com/>

In the Indian context, a framework for knowledge acquisition, conceptualization, formalization, implementation, and evaluation for a knowledge base is presented in [22]. This knowledge base contained many Indian food items, but the focus was on methodology. A digital resource of 528 key Indian food ingredients along with their nutritional information is curated and presented in [23]. Most of the above ontologies and knowledge graphs were built from semi-structured recipe cards for specific applications. [24] proposed a formal but generic methodology for gathering information and building a food ontology in an automated fashion. The proposed work for building an Indian Food Ontology extends this pipeline.

3. Unique and Complex Challenges in building FKG.in: Significance and Relevance of Indian food

The diverse Indian cuisine reflects a history of over 8,000 years, during which the history of various ethnic groups and cultures have interacted with each other in the Indian subcontinent, resulting in a vast variety of cooking techniques, flavors, and regional cuisines found in modern-day India. While this diversity has resulted in a rich repertoire of recipes, it has also introduced some unique challenges towards automating the task of building a food knowledge base. We note these and a few other challenges below: p

1. **Lack of a comprehensive vocabulary** of food items has necessitated that this work start from almost scratch. This spans all concepts related to food like ingredients, cuisines, styles, cooking processes, cookware, and so on. Indian food habits are largely society-driven or family-driven, which has led to unique styles and practices that are crucial to Indian food, but not properly documented. Additionally, there doesn't exist any comprehensive source of nutritional information about Indian food which makes it difficult to envision well-grounded applications in health around Indian food.
2. There exists a **multiplicity of recipes** with the same name but different compositions from different regions. To a large extent, this occurs due to regional variations in climate, culture, and availability of ingredients. The most common example of this is *dal* (lentil soup) which, though an integral part of almost all regional meals, has huge variations across the country. Additionally, as Indian food recipes are often quite complex, capturing the nuances of similar recipes is often a very difficult task.
3. The **multilingual** nature of India poses a challenge exactly opposite to the previous one. The same food items have various vernacular names across the country. For example, *haldi* (Hindi), *holud* (Bengali), *halad* (Marathi), *pasupu* (Telugu) and *manjal* (Tamil), all refer to turmeric in different Indian languages. Building a common and inclusive dictionary of food items for India needs multilingual capabilities to address this diversity.
4. **Food homonyms** present a challenge in the form of confusing granularities where ingredients and recipes may be known by the same name. A typical example is *chawal* (rice) which refers to both raw rice i.e. an ingredient and steamed rice i.e. the final dish.
5. Another challenge stems from the fact that Indian food is **not about precision cooking**. Measurements are often expressed in terms of common kitchen containers like "a cup" or "a *katori* (bowl)", for which there are no specific standards. Use of linguistic variables like "a little", "some", and "a handful of" are also encountered quite often.

6. **Sociocultural association of food items with festivals, religious celebrations, and spiritual motivations** is a worldwide phenomenon. These notions have to be captured to generate contextually relevant recommendations. For example, *kheer* or *payasam* (milk pudding), and *Hyderabadi haleem* are almost always associated with different religious celebrations, whereas abstinence food is typically expected to be without garlic or onion. Appropriate vocabularies for capturing the Indian context must be built.

The above challenges only highlight some of the issues that are faced while building a knowledge base of recipes. Unless addressed, it is only likely that these challenges will amplify and introduce more difficulties while building food computing applications on top of the knowledge base. While we have tried to put in adequate measures to mitigate most of these challenges in our design, we do not claim that all the challenges have been addressed completely. We believe that the design has enough flexibility to accommodate newer dimensions of food without requiring a whole lot of rework.

4. Proposed Ontology Design for Indian Food

As mentioned earlier, due to the absence of computationally accessible comprehensive digital documentation, we had to start the work of designing a foundational ontology for Indian food from scratch. Documentation about the domain itself is quite sparse and scattered. Though various controlled vocabularies, food ontologies, and other sources of rich food-related information exist for selected aspects of food and in some cases, even for Indian food, however, either they largely fail to capture the wide-ranging attributes and subtleties of Indian food, or they serve niche applications and services.

We now present the details of FKG.in, the Indian Food Ontology. We have taken design inspiration for our core ontology from FoodOn [14] and FoodKG [15] and wherever needed, adapted them to suit the Indian context. FKG.in attempts to capture important properties of Indian food in the purview of culinary language, cooking variations, and precision nutrition. In doing so, we have also attempted to make the ontology modular and flexible to incorporate changes in the knowledge curation stage if required.

Figure 1 presents the proposed ontology design. **Food** is an abstract superclass and **Ingredients** and **Recipes** are its most important subclasses. These two concepts are described in more detail below:

1. **Recipe** class: At the heart of the proposed food ontology lies the **Recipe** class. A **Recipe** instance (or recipe) is composed of measured ingredients, physical and conceptual properties, cooking characteristics, and a set of cooking instructions. Different sets of properties are associated with the **Recipe** class. The first set comprises those that have values represented as simple strings like **name**, **cuisine**, **serving size**, **calories**, etc. Secondly, a recipe is characterized by detailed **cooking instructions**, which we store as long text. The third set of properties is composite. For example, a recipe has **Cooking Characteristics**, which is an aggregate class, designed along the lines of a similar class available in FoodKG [15]. This contains **cooking techniques**, **cookware**, **cooking temperature**, etc. Vocabularies to cater to Indian cooking processes, practices, kitchen utensils, etc. have been curated. For a **Recipe** instance, while many of these

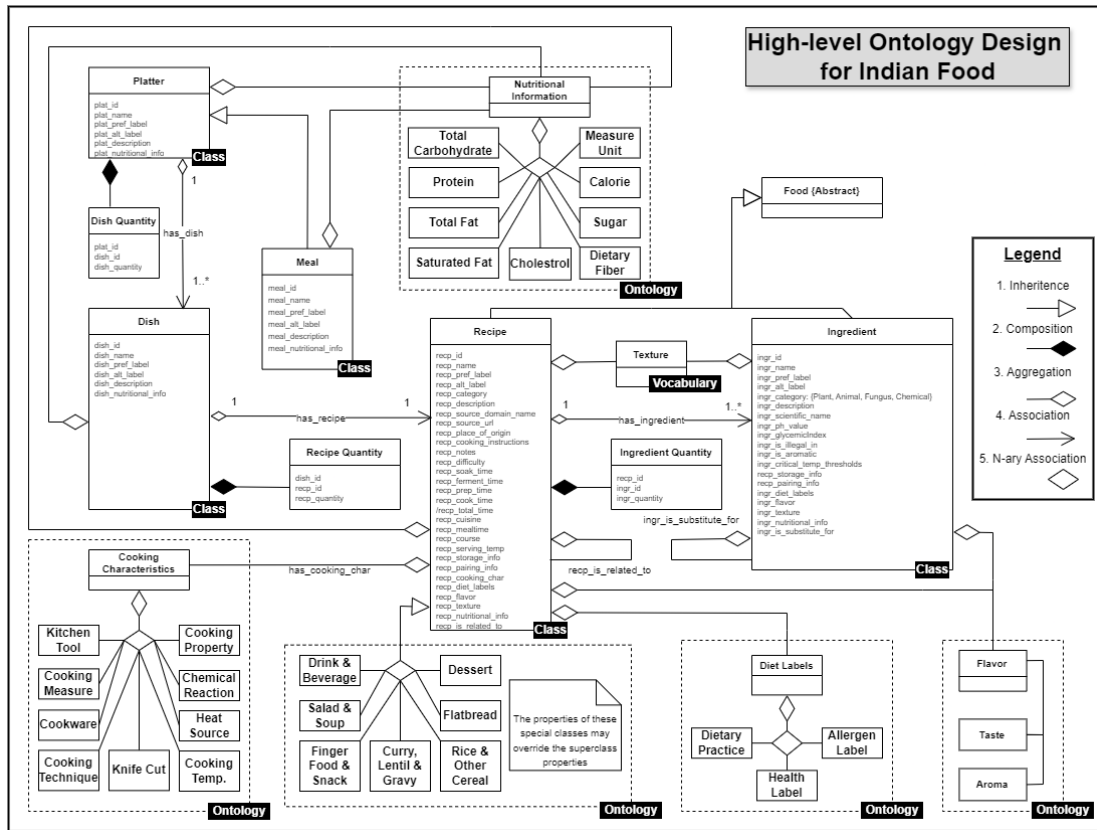


Figure 1: High-level Ontology Design for Indian Food (incl. Cooking Characteristics and Nutrition)

property values may be available from data, some are derived using predefined functions. For example, often recipes do not have nutritional information or diet labels directly mentioned. These can be derived from the ingredients, their measurements, and cooking characteristics using dedicated functions. The property **pairing information** stores names of other recipes that are usually taken with it. For example, <chawal, dal>, <idli, (chutney, sambar)>, <biryani, raita/salan> are common pairs in Indian food.

2. **Recipe** subclasses: There are several subclasses of **Recipe** like **Flatbread**, **Dessert**, **Beverage & Drink**, etc. which have typical properties associated with them like **texture**, **servicing temperature**, and so on. This set of subclasses has been adapted from FoodKG [15] and extended to accommodate Indian food.
3. **Ingredient** class: A **Recipe** instance uses **ingredients**, which is the second most significant concept in our ontology. Any food item that can contribute to a combination of other ingredients to make a particular recipe is an instance of the **Ingredient** class. This class has a long list of properties that define the **origin**, **flavor**, **glycemic index**, **nutritional information**, **pH value**, etc. This list of properties is also adapted from the class **Food Category** of FoodKG [15] and extended to accommodate Indian food concepts. According to [15], there are four basic ingredient categories based on origin - **plant**,

animal, fungus, and chemical. Plant-origin ingredients themselves may be used either in primary or processed forms. Further sub-categorization may be based on **fruits, herbs, legumes, milled cereal products,** and so on. There are many more sub-categories highlighting texture, flavor, processing technique, etc. Based on these, we can categorize Indian spice ingredients under different subheadings, based on the descriptions used in the recipes. For example, coriander is used in recipes very frequently either as a **fresh herb, dried seeds, powder,** or **paste.** These are often not interchangeable. Provisions to accommodate all of these are provided in the ontology.

A few examples of the relations supported in the ontology are $\langle r, \text{has_ingredient}, i \rangle$, $\langle r, \text{has_cooking_char}, c \rangle$, $\langle r1, \text{is_a}, r \rangle$, where **r**, **i**, **c** and **r1** are objects belonging to **Recipe**, **Ingredient** and **Cooking Characteristics** classes and **Recipe** subclasses respectively. The **has_ingredient** relation is a labeled one that stores the measured quantity of ingredient **i** to be used for recipe **r** as well.

Different instances of the same recipe may exist with variations in terms of ingredients, instructions, etc. Each of them is considered a unique object of the class recipe. In the Indian context, the same ingredient may be referred to by different regional names. For each such ingredient, a single instance of it is created in the knowledge graph, while storing the different names within it for resolution later. Additionally, since Indian recipes are more compositional in nature in which the **texture** and **flavor** of the food item are described more in terms of the end product rather than those of the ingredients alone, we have added texture and flavor as properties of the recipe as well, thus adapting [15] to the Indian context. A detailed **Cuisine** hierarchy has been built for Indian sub-continental food along with associated functions to determine the cuisine label from the recipe ingredients, origin, etc. Variations of the same recipe often exist in different Indian cuisines and it is important to capture the granular cuisines within the Indian subcontinent. For example, *Lucknowi Chicken Biryani* is associated with the **Awadhi cuisine** of North India whereas *Mutton Donne Biryani* is of South Indian origin and is associated with the **Karnataka cuisine**. Similarly, mealtime vocabulary has been extended to accommodate Indian festival meals like *Iftaar*, *Navaratri* specials, etc. A long and multidimensional list of **diet labels** has been also curated from multiple sources to accommodate concepts like **<Dietary Practice: Jain-vegetarianism>**, **<Health Label: Keto-friendly>**, and **<Allergen Label: Dairy-free>**. For example, within **diet labels**, lists of 14 **dietary practices**, 21 **allergen labels**, and 22 **health labels** applicable to the Indian context have been created so far.

4. **Dish** class: Any **Recipe** instance that is qualified by a measurement unit inherited from the recipe with appropriate scaling and describing the serving size of the recipe is an instance of the **Dish** class.
5. **Platter** class: A platter is a composition of dishes, along with their respective quantities specified, to be always viewed as a single entity.
6. **Meal** subclass: Meal is a subclass of platter, which is usually associated with specific occasions or times of day.

Now we provide a complete example of these concepts. *Chicken Chettinad* is a popular South Indian recipe of chicken curry. A recipe of *Chicken Chettinad* is mentioned to serve 4 people and contains 1 kg of chicken and various other ingredients like 6 pieces of cardamoms, 2 onions,

etc. 1 dish of *Chicken Chettinad* serving 1 person may be a scaled version of the recipe which contains $\frac{1}{4}$ (or 250 gm of) Chicken. A typical meal called a *South Indian nonvegetarian thali* may be composed of 1 plate of steamed rice, 1 glass of *tomato rasam*, 1 cup of curd, 1 bowl of *Chicken Chettinad*, and 1 *papadum*. Bowl, glass, cup, and plate are typical measures of Indian dishes belonging to different recipe categories which themselves may result in measurement variations. To avoid the ambiguity associated with these terms, we will store a dictionary of these terms along with their precise definitions such as 1 bowl equals 250 gm. A meal can also constitute a single dish alone. For example, *Chicken dum biryani* is a dish that is also a complete meal by itself, prepared using *dum-cook*, a typical Indian cooking process, in a *Handi* which is also a typical Indian cooking vessel.

Some special relations are described to capture the essence of a **Recipe** instance better:

1. **ingr_is_substitute_for** Ingredient property: A pair of ingredients **i1** and **i2** are stored as **<i1, ingr_is_substitute_for, i2>**, if they are substitutes of each other, but may have different **nutritional properties, diet labels**, etc. For example, Iodized salt and Himalayan pink salt are substitutes, and depending on which one is used, recipe nutrition may vary.
2. **recp_is_relate_to** Recipe property: This relation captures the semantic similarity between a pair of recipes. For example, Aloo samosa and Mutton samosa are similar.

In an ontology, it will be important to specify restrictions, rules, logic, or constraints on concepts and relations that must be satisfied by an object in the ontology for performing consistency checks. The following examples show how restrictions have been used in the current system to validate properties:

1. For all recipes **r**, if **r** has the property label “Non-vegetarian” for **cuisine**, then there exists at least one ingredient **g** for **r** that has **ingredient_category** “meat” or “egg” from **animal_origin**.
2. For all recipes **r**, if **r** has the property label “Vegetarian” for **cuisine**, then there does not exist any ingredient **g** for **r** that has **ingredient_category** “meat” or “egg” from **animal_origin**. This defines Indian vegetarian cuisine which is primarily lacto-vegetarian.
3. For all recipes **r**, if **r** has the property label “Jain” for **cuisine**, then there does not exist any ingredient **g** for **r** that has **ingredient_category** (“meat” from **animal_origin**) OR (“root_vegetable” from **plant_origin**)

In the next section, we define how a knowledge graph of Indian recipes has been built using the proposed ontology design.

5. Knowledge Curation Workflow for building FKG.in

Figure 2 presents the core tasks of the AI-driven semi-automated workflow for building a food knowledge graph using structured and unstructured information curated from multiple websites. The details of each task are presented in this section.

- **Task 1** - Creation of Foundational Ontology An instance of the foundational ontology for Indian food is created using Web Ontology Language (OWL) in the RDF/XML format by extending reliable and recognized ontologies and dictionaries. Detailed hierarchical designs for storing the labels of **cuisines, diet labels, mealtimes, cooking characteristics**,

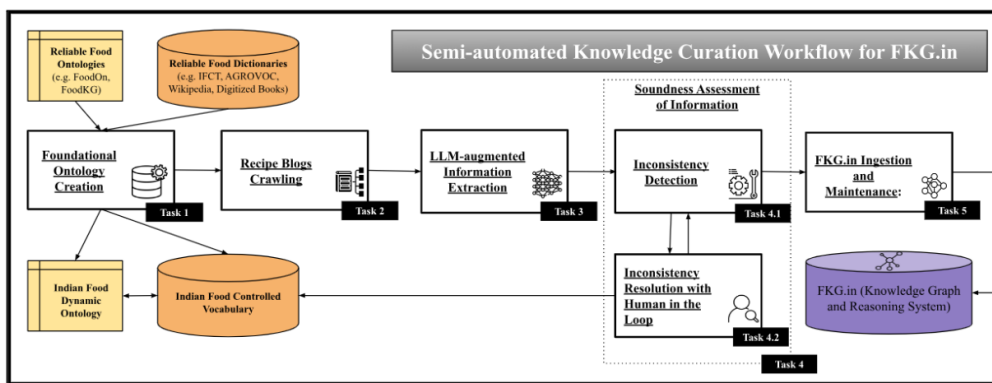


Figure 2: Semi-automated Knowledge Curation Workflow for FKG.in

etc. described earlier have also been created in the form of a dictionary. Initial vocabulary for each of these structures was populated using various sources like Wikipedia, digitized books [25], FoodKG [15], Indian Food Composition Table [23], and other sources for Indian food [22], though the vocabulary itself is not restricted to Indian food alone. The vocabulary was further augmented using large language models. For example, a list of bulb or stem vegetables commonly found in India was generated using OpenAI's GPT-3.5 Turbo⁵ and added to the list of ingredients. The initialization step is a one-time process that involves careful manual curation with several sanity checks. The vocabulary is also updated periodically as explained later. We are using the SKOS (Simple Knowledge Organization System) W3C recommendation⁶ to represent and organize the structured controlled vocabulary as the principal element categories of SKOS such as concepts, labels, notations, documentation, semantic relations, mapping properties, and collections suit the needs of storing food, culinary and nutritional knowledge quite well.

- **Task 2 - Crawling of Recipe Blogs:** We have identified 40 recipe blogs and websites with rich information about Indian food recipes, their nutritional information, and other culinary information. To begin with, we have crawled 5 recipe blogs viz. **archanaskitchen**, **hookedonheat**, **indianhealthyrecipes**, **masalakorb** and **vegrecipesofindia**, each of which has several recipe websites along with a detailed recipe card for each. The crawler gathers content from each recipe blog page and stores it locally as an HTML file along with metadata like its **source URL**, **recipe name**, **recipe category**, **blogpost timestamp**, and **scraping timestamp** for reproducibility and parsing.
- **Task 3 - LLM-augmented Information Extraction:** The HTML files are then cleaned and parsed to extract the recipe details such as ingredients, cooking characteristics, nutritional information, etc. from both structured and unstructured parts. This was done by setting up a pipeline using Langchain and GPT-3.5 Turbo, a large language model, to process the recipe webpage content and generate semi-structured output using zero-shot and few-shot prompts. These would be used later to populate the knowledge graph after the

⁵<https://platform.openai.com/docs/models>

⁶<https://www.w3.org/2009/08/skos-reference/skos.html>

soundness check. This process is executed for all the recipe URLs before moving on to the next steps.

- **Task 4** - Soundness Assessment of Information: After generating food entities and relations from each recipe, this step runs automated checks to validate the information against existing vocabularies, performs entity resolution, if needed, and then flags inconsistencies to humans for correction. This step is to ensure that the information added to the knowledge base is correct.

- **Task 4.1** - Several language processing tools along are employed to detect inconsistencies. Prompt engineering is employed to pose a set of contextual questions to the LLM with the recipe as a context and force the output to be in a structured form. Figure 3 shows a sample output. The extracted information is used to improve the accuracy and precision of entity recognition. Information extracted from multiple tools is cross-checked to ensure correctness, and flag information that needs to be verified through human intervention. For example, if an entity *kadahi* (*wok*) is incorrectly identified as a recipe ingredient, instead of a vessel to cook as listed in the vocabulary, then the system flags an inconsistency. Similarly, Indian spice names, if already present in the vocabulary, are mapped to the unique identifier and if not present, then flagged for human inspection to be included in the vocabulary appropriately. Restriction-based checks are also applied at this stage.

- **Task 4.2** - All inconsistencies identified in the earlier step are presented to human curators for validation and correction if needed. For example, a common mistake made by language tools, including LLMs, is the failure to detect multi-word named entities correctly. For example, for a recipe to make "*pudina chutney sandwich*", the key ingredient is "*pudina chutney*" and not "*pudina*" which is a basic ingredient. A human can correct the entity and help in appropriate incorporation of "*Pudina chutney*" as an ingredient, which is also a recipe, and may appear as such in other recipes also. Several instances of incorrect and incomplete information extraction are observed for unstructured portions.

An easy-to-use interface has been built to aid the correction process. We observed that in the initial phases, the system was more error-prone, and therefore needed more human interventions to resolve conflicts. However, as the size of the vocabulary increases, with the addition of more content, the task reduces. Further, insights obtained from human intervention were used to improve the LLM prompts, which also helped reduce the error of the extraction process. The human feedback is then incorporated into the vocabulary by augmenting it as per the ontology in an atomic, reliable, and consistent manner in light of the new information obtained from the recipe web pages.

The above methods ensure the soundness of the knowledge graph, i.e. information added to the knowledge base is correct. It however does not ensure completeness of information in situations where the large language model fails to extract a piece of information altogether. Such issues will be addressed in the future while working on the completeness of the knowledge graph for Indian food.

- **Task 5** - FKG.in Ingestion and Maintenance: The verified and validated information

components are ingested into the knowledge graph. While some of them may result in vocabulary extensions, some are added as instances of classes and relationships.

The algorithmic details of the knowledge curation workflow are presented below:

1. **Initialization:** Make a list of target information to be extracted from recipe URLs based on the data/class properties associated with ingredients and recipes as per the ontology.
2. **Crawling and Extraction:** Fetch and store the recipe dump locally for all the recipe URLs. Use the **requests** library in Python to parse and extract target information from the recipe card and store it in an XML file.
3. **Semantic Resolution:** Use semantic resolution to map property names across recipe domains. For example, recipe blogs may use the term **region** or **style** to refer to **cuisine**. While the dataset is curated mostly with manual intervention in the initial phases, the lists of property names and values are automated and learned over time.
4. **LLM-enabled Entity Recognition:** Use fine-tuned prompts and LLMs to extract information and recognize entities from the unstructured recipe webpage content which contains ingredient details along with cooking instructions, cooking characteristics, etc in long text. Ingredient measures may also be included in this information. Prompt engineering is used to store the extracted information in a structured format to enable comparison with the recipe card information that was stored in the XML format earlier.
5. **Soundness Assessment:** Compare the XML output with the LLM output to obtain a match score, where a score of +1 indicates a match between the two tuples and a score of -1, whenever a mismatch occurs between a recipe card tuple and an LLM output. All LLM tuples that do not have a corresponding match in the recipe card are matched against the vocabulary terms of the corresponding property list. For each match found, a score of +1 is awarded and -1 for matches not found. For each recipe parsed, the total positive score is an indicator of the soundness of the information as it is double-checked against recipe cards and vocabulary lists. All negative scores are flagged for human validation in the next step. All XML tuples and LLM tuples with a score of +1 are candidate elements for the knowledge graph. The total positive score provides an assessment of the underlying LLM-based information extraction system, which will be the only way to extract information from totally unstructured websites without recipe cards. This will be explored in future.
6. **Human Validation and Updation:** An easy-to-use interface is used by human curator to assess all flagged information. Vocabulary updates, if any, are also enabled through an interactive platform. Tuples are also assessed and corrected, if necessary. All actions, resolved and not resolved, are documented. The information was used heavily to finalize the ontology design and is stored for any future needs.
7. **Ingestion of Tuples into FKG.in:** After updating the vocabularies, the knowledge graph is ingested with the new and validated information as per the latest ontology. All unique tuples from the candidate set of step 5 and human-approved tuples from step 6 are used to extend the knowledge graph to include new instances of objects and relations. The tuples in the form of RDF/XML triple-stores are stored in an OWL file which stores both the Indian food ontology and the associated vocabulary. We are currently using Ontotext's GraphDB⁷ to build the knowledge graph.

⁷<https://graphdb.ontotext.com/>

Prompt: Give brief answers to the following questions marked as "Q" from the given context

Context: Scraped Website Content (Right)

Q: A list of Questions (Below)

Response:

Q: What is the main dish being mentioned?
A: Carrot Porial.

Q: Identify each ingredient and their measures used for cooking and print it as ingredient : measure.
A:

- Carrots : 4
- Mustard seeds : 1/2 teaspoon
- White Urad Dal (Split) : 1 teaspoon
- Dry Red Chilli : 1
- Curry leaves : 1 sprig
- Asafoetida : 1/4 teaspoon
- Turmeric powder : 1/4 teaspoon
- Fresh coconut : 1/4 cup
- Oil : 1 teaspoon
- Salt : to taste

Q: What are the cooking processes mentioned?
A: Steaming the carrots, seasoning, and stir-frying.

Q: What are the cooking vessels mentioned?
A: Pressure cooker and a wok.

Q: How many steps are mentioned for cooking?
A: Four steps.

Q: What is the preparation time in minutes or hours?
A: 10 minutes.

Q: What is the cooking time in minutes or hours?
A: 30 minutes.

Figure 3: Sample response by ChatGPT based on Contextual Q&A prompt for an *archanaskitchen* recipe

Though not implemented currently, in the future the knowledge graph will undergo systematic checks to perform evaluations and optimizations by using quality metrics and optimized organization principles based on the SKOS recommendation.

6. Current Status of FKG.in

The size of FKG.in is presently around 50 MB. It has information about 9628 unique recipe instances gathered from the five recipe blogs mentioned earlier. After consolidating the meta-data, these recipes belong to 39 distinct categories such as **breakfast, cakes, vegetarian, Hyderabadi, Indian sweets**, etc. The total number of ingredient nodes in the knowledge graph is currently 38819. We have observed some nodes have Hindi names in **Devanagari** fonts, indicating that more resolution rules will need to be added to address code-mixing. It has also been observed that complex ingredients, which are recipes themselves, are duplicated in the knowledge graph, as both a recipe node and an ingredient node. This needs to be resolved with an associative relationship, which is currently not a part of the design.

7. Conclusions and Future Work

In this paper, we have presented the work initiated towards building FKG.in. Due to the lack of reference resources, almost everything had to be initiated from scratch. Unlike earlier methods, which have focused on building knowledge graphs from semi-structured data and in application-specific ways, our focus is on using AI-enabled methods for extracting relevant information from all kinds of recipe blogs to populate the knowledge graph. Zero-shot and few-shot methods that exploit the large language model GPT-3.5 Turbo have been used extensively to build initial vocabularies and subsequently to extract entities and relations to populate the knowledge base. Methods to ensure soundness of information are also incorporated into the pipeline. We have

presented the current status of the knowledge graph called FKG.in.

Further work on the refinement of ontology design as well as on knowledge engineering techniques is underway. NLP tools for multilingual semantic reasoning are one of the primary areas identified for future research. Another area of focus is on quantitative assessments of the soundness and completeness of the knowledge graph.

In the future, we expect work to spread across many other directions that involve reasoning over Indian food concepts as well. Such knowledge graphs can address several questions of historical, social, and cultural aspects of food and food habits, enable several applications including but not limited to food recommendation systems, personal health navigation systems, recipe generation, and recipe recommendation systems, and also aid knowledge discovery from underlying data. The methods for knowledge curation proposed in this paper are generic and can be replicated for any domain.

8. Acknowledgments

This research was supported by the Ashoka Mphasis Lab - a collaboration between Ashoka University and Mphasis Limited.

References

- [1] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," *ACM Comput. Surveys*, vol. 52, no. 5, pp. 1–36, Sep. 2019.
- [2] W. Min, S. Jiang, and R. Jain, "Food recommendation: Framework, existing solutions, and challenges," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2659–2671, Oct. 2020.
- [3] W. Min, C. Liu, L. Xu, and S. Jiang, "Applications of knowledge graphs for food science and industry," *Patterns*, vol. 3, no. 5, May 2022.
- [4] A. Cordier, V. Dufour-Lussier, J. Lieber, E. Nauer, F. Badra, J. Cojan, E. Gaillard, L. Infante-Blanco, P. Molli, A. Napoli, and H. Skaf-Molli, "Taaable: a case-based system for personalized cooking," in *Successful Case-based Reasoning Applications-2*, S. Montani and L. C. Jain, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, Jan 2014, pp. 121–162.
- [5] R. Ribeiro, F. Batista, J. P. Pardal, N. J. Mamede, and H. S. Pinto, "Cooking an ontology," in *Artificial Intelligence: Methodology, Systems, and Applications*, J. Euzenat and J. Domingue, Eds. Berlin: Springer, 2006, pp. 213–221.
- [6] V. Sherimon, S. P.C., A. Ismaeel, W. Varkey, and N. B., "Modeling of seafood domain using ontology," *Intl Jr. of Open Info. Technologies*, vol. 9, no. 2, 2021.
- [7] U. Zulaika, A. Gutiérrez, and D. López-de Ipiña, "Enhancing profile and context aware relevant food search through knowledge graphs," *Proceedings*, vol. 2, no. 19, 2018.
- [8] Z. Lei, A. U. Haq, A. Zeb, M. Suzauddola, and D. Zhang, "Is the suggested food your desired?: Multi-modal recipe recommendation with demand-based knowledge graph," *Expert Systems with Applications*, vol. 186, no. 30, Dec. 2021.
- [9] J. Marín, A. Biswas, F. Ofli, N. Hynes, A. Salvador, Y. Aytar, I. Weber, and A. Torralba, "Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images," *IEEE Trans. on PAMI*, vol. 43, no. 01, pp. 187–203, Jan 2021.
- [10] G. Bagler, "A generative grammar of cooking," *arXiv*, Oct. 2020.

- [11] D. Batra, N. Diwan, U. Upadhyay, J. S. Kalra, T. Sharma, A. K. Sharma, D. Khanna, J. S. Marwah, S. Kalathil, N. Singh, R. Tuwani, and G. Bagler, "RecipeDB: a resource for exploring recipes," Database, vol. 2020, Nov. 2020.
- [12] J. Cantais, D. Dominguez, V. Gigante, L. Laera, and V. Tamma, "An example of food ontology for diabetes control," 2005, pp. 1–9.
- [13] C. Snae and M. Bruckner, "Foods: A food-oriented ontology-driven system," in 2008 2nd IEEE Int'l Conf. on Digital Ecosystems & Technologies, 2008, pp. 168–176.
- [14] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. L. Brinkman, and W. W. L. Hsiao, "FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration," npj Science of Food, vol. 2, Dec. 2018.
- [15] S. Haussmann, O. Seneviratne, Y. Chen, Y. Ne'eman, J. Codella, C.-H. Chen, D. L. McGuinness, and M. J. Zaki, "FoodKG: A semantics-driven knowledge graph for food recommendation," in The Semantic Web – ISWC 2019, C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, and F. Gandon, Eds. Cham: Springer International Publishing, 2019, pp. 146–162, resource Website: <https://foodkg.github.io>.
- [16] Y. Chi, C. Yu, X. Qi, and H. Xu, "Knowledge management in healthcare sustainability: A smart healthy diet assistant in traditional chinese medicine culture," Sustainability, vol. 10, no. 11, 2018.
- [17] C. Caracciolo, A. Stellato, S. Rajbahndari, A. Morshed, G. Johannsen, Y. Jaques, and J. Keizer, "Thesaurus maintenance, alignment and publication as linked data: The agroovoc use case," in Metadata and Semantic Research, E. García-Barriocanal, Z. Cebeci, M. C. Okur, and A. Öztürk, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 489–499.
- [18] T. Pizzuti, G. Mirabelli, M. A. Sanz-Bobi, and F. Gómez-Gonzaléz, "Food track & trace ontology for helping the food traceability control," Journal of Food Engineering, vol. 120, pp. 17–30, Jan. 2014.
- [19] F. Ameri, E. Wallace, R. Yoder, and F. Riddick, "Enabling traceability in agri-food supply chains using an ontological approach," Jr. of Computing and Info. Sc. in Engg., vol. 22, no. 5, pp. 17–30, Oct. 2022.
- [20] L. Qin, Z. Hao, and L. Zhao, "Food safety knowledge graph and question answering system," in ICIT '19: Proc. of the 2019 7th Intl Conf. on Info. Tech.: IoT and Smart City. New York, United States: ACM, Dec. 2019, pp. 559–564.
- [21] L. Qin, Z. Hao, and L. Zhao, "Question answering system based on food spot-check knowledge graph," in ICCDE '20: Proceedings of 2020 6th International Conference on Computing and Data Engineering. New York, United States: Association for Computing Machinery, Jan. 2020, pp. 168–172.
- [22] T. Padmavathi and M. Krishnamurthy, "Ontology for the domain of food science," Journal of Information and Knowledge, vol. 53, no. 5, pp. 409–417, Oct. 2016.
- [23] S. Sahu, "ifct2017/compositions: Detailed nutrient composition of 528 key foods in india," Sep. 2022.
- [24] D. P. Madalli, U. Chatterjee, and B. Dutta, "An analytical approach to building a core ontology for food," Journal of Documentation, vol. 70, no. 1, pp. 123–144, 2017.
- [25] K. Ashok, Masala Lab: The Science of Indian Cooking, Kindle ed. Penguin, Dec. 2020.