Building a Knowledge Graph for Indian Food

Saransh Kumar Gupta [1] Lipika Dey [1] Partha Pratim Das [1] Ramesh Jain [2]

[1]Centre for Health Analytics, Research and Trends, Ashoka University, India [2]Institute of Future Health, UC Irvine, USA

Ashoka Mphasis Lab

Abstract

- **Concern:** Rapidly increasing diet-related diseases **(obesity, hypertension, diabetes)** in India.
- **Context:** ISO/TC 215/WG 11 (Health Informatics/Personalized Digital Health) is working on a draft for **Personalized Health Navigation** to guide **food, healthcare, and lifestyle choices.**
- **Challenge:** Any definitive effort in the scientific pursuit and all well-founded applications around food that can be envisioned can only be **as good as the underlying food-related data.**

Figure 1. Major gap between **what we enjoy eating (taste) and what our body needs (health)**

- **Need of the hour:** To leverage data science, software engineering, and machine learning to understand Indian food and associated culinary & nutritional information with a scientific lens.
- **Opportunity:** To build a **reliable, comprehensive, and granular knowledge representation and reasoning system of Indian food** as an indispensable step for **personalized digital health** applications and use technology as a means to improve the quality of food, health, & life.

Background

Our research group is working on the **ontology design, knowledge engineering, and multilingual semantic reasoning** problems while building a **semi-automated system for assimilating culinary information in the form of a knowledge graph for Indian food (FKG.in)** with associated ontologies, data dictionaries, and intelligence at scale to **digitally capture all-encompassing knowledge about food, recipe, ingredients, processes, and most importantly nutrition**.

- **Objective:** To perform **personalized AI-driven smart analysis**, build **recommendation systems** for personalized health navigation, and complement the knowledge graph for Indian food with **contextual information related to users, food biochemistry, geographies, agriculture, etc.**
- **Related Work:** Our work is inspired by and builds upon **IFCT, AGROVOC, FoodOn, FoodKG, Recipe1M+, RecipeDB, and Edamam** to accommodate Indian food (References not included).

Method

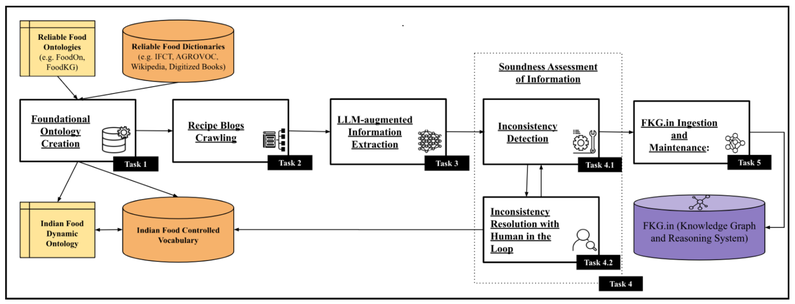

Figure 2. Semi-automated knowledge curation workflow for FKG.in

We propose a **novel AI-based semi-automated approach to curate culinary & nutrition information from public domain websites** to populate the knowledge graph, **with a human-in-the-loop intervention and algorithmic methods to ensure the soundness of information**.

Challenges: Uniqueness and Complexities of Indian food

The **diversity of Indian cuisine** reflects over 8,000 years of history of various groups and cultures interacting with the Indian subcontinent, historical influences of other cuisines, recent interactions across the world, and the contemporary rise of various fusion dishes.

1. A comprehensive Indian food vocabulary is lacking and reliable nutritional information is scarce.
2. Multiplicity and diversity of recipes across geographies, cultures, and households
   - *Dal* (lentil soup) has the same ingredients but different compositions in different regions.
3. Multilinguality: aliases and spelling variations of dishes, ingredients, and measurement units
   - E.g. Turmeric is known as *haldi* (Hindi), *holud* (Bengali), *halad* (Marathi), *pasupu* (Telugu), and *manjal* (Tamil) locally.
4. Food homonyms
   - E.g. *Chawal* (rice) refers to both raw rice i.e. an ingredient and steamed rice i.e. the final dish.
5. Lack of a precision cooking practice: traditional and inconsistent measurement units.
   - E.g. No specific standard for terms like “a cup”, “a *katori* (bowl)”, ”a glass” or “a handful.”
6. Sociocultural associations of food with festivals, religious celebrations, and spiritual motivations
7. Invisibilized, demeaned, and unclassified food items.
8. Complex and composite recipes with elaborate and delicate cooking steps.
9. Seasonal, regional, and varied ingredients and multi-tiered spices.
10. Tricky serving patterns of dishes and meals in plates.
11. Limited pre-existing food ontologies, schemas, standards, and regulations.
12. Weak notions of tracking personal health and culture-based nutrition systems.

Food recommendation in India is much more **society-driven and family-driven** than in many other countries where food recommendation is more **focused, personal, and choice-driven.**

Proposed Ontology Design

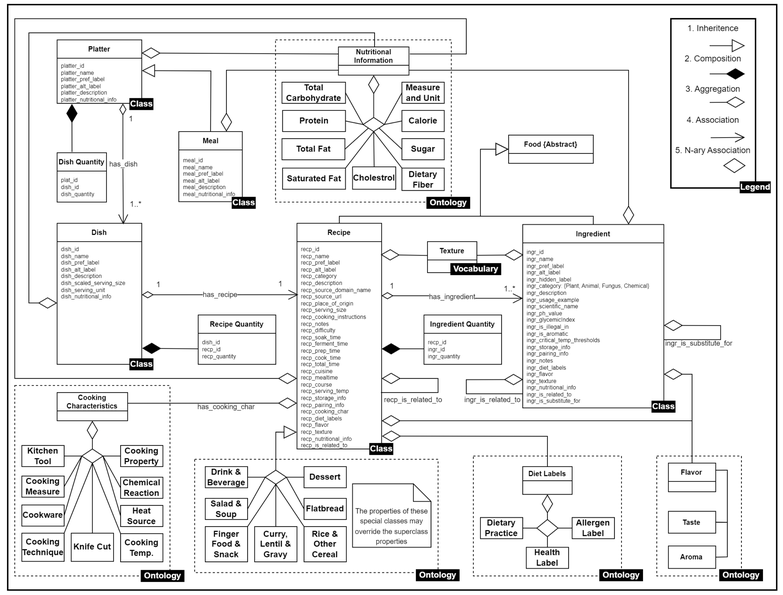

Figure 3. High-level ontology design for Indian food and FKG.in

We have attempted to address some of the challenges mentioned above as depicted in this design. However, **the ontology design still requires a lot of work** and we hope to **continuously and iteratively improve it in the next few phases of designing, building, and testing FKG.in**.

Results

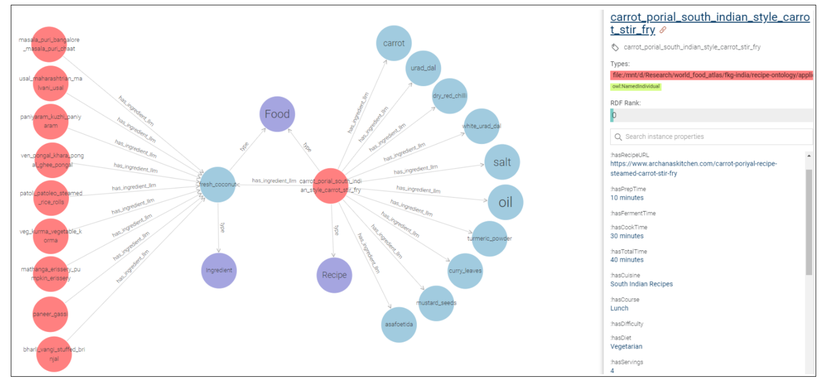

Figure 4. An FKG.in instance highlighting some of the ingredients of **carrot *porial* (stir-fry)** - an *archanaskitchen* recipe

Current Status of FKG.in: Summary

- The size of FKG.in (OWL file) is presently around **50 MB**.
  - An initial vocabulary was populated from Wikipedia, Books, FoodKG & IFCT and **augmented with LLMs**.
- Number of recipe blog domains crawled: **5**, identified: **40**.
  - **archanaskitchen, hookedonheat, indianhealthyrecipes, masalakorb, vegrecipesofindia**
- Number of unique recipe URLs analyzed: **9628**.
  - These belong to **39 distinct categories** such as breakfast, cakes, vegetarian, Hyderabadi, Indian sweets, etc.
- Number of ingredient nodes: **38819**.
  - Includes noise, word sense ambiguity, & redundancies **(code-mixing, spelling errors, & complex ingredients)**.

Sample SPARQL query results as performed on FKG.in (in natural language for readability)

1. Query: Find recipes that **contain chicken, carrots, and potatoes** as ingredients.
   - **Results:** goan chicken stew with vegetables, chicken barley soup, chicken steak with pan-roasted vegetables
2. Query: Which recipes **contain walnuts but do not contain bananas**?
   - **Results:** walnut mushroom bolognese, kale pesto pasta, pear walnut salad, walnut burfi, mix dry fruit laddu
3. Query: What recipes from **Chinese cuisine can be cooked in under 30 minutes**?
   - **Results:** egg hakka noodles, chicken schezwan momo, chicken and egg soup

Conclusion

- Due to a lack of reference resources along many axes, **most of this work started from scratch.**
- Unlike earlier methods, which focused on building knowledge graphs from semi-structured data and in application-specific ways, our focus is on **using AI-enabled methods to extract relevant information from recipe blogs to populate application-agnostic food knowledge graphs.**
- Zero-shot & few-shot methods exploiting large language model **(GPT-3.5 Turbo)** are extensively used to **build initial vocabularies and to extract entities and relations for FKG.in population.**

Future Work:

1. Refinement of **ontology design** as well as **knowledge engineering techniques.**
2. NLP tools for **better entity recognition**, **reliable dictionary population**, and **multilingual semantic reasoning**.
3. Detailed description, design, and development of a **Health and Nutrition** framework.
4. Qualitative and quantitative **assessments of soundness and completeness** of the knowledge graph.
5. **Redundancy and inconsistency resolutions** in FKG.in using organization principles and quality metrics (SKOS).

Acknowledgement

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

FKG.in aims to **address questions** around Indian food knowledge, dietary practices, and health informatics, **enable applications** such as recommendation, personal health navigation, recipe generation, & recipe recommendation, and **aid knowledge discovery** from underlying data.