POS tagging probability weighted method for matching the Internet recipe ingredients with food composition data

KDIR 2015 – 7th International Conference on Knowledge Discovery and Information Retrieval

Tome Eftimov

Barbara Koroušić Seljak

{tome.eftimov, barbara.korousic}@ijs.si

EuroFIR European Food Information Resource

ISO Fd Food ERA Chair Safety Quality Traceability

Computer Systems Jožef Stefan Institute

Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

Overview

- Motivation
- Introduction
- Related work
- Problem definition
- Evaluation and results
- Conclusion

Motivation


Finally settling down to my vegan, gluten free, soy free, antibiotics free, raw, non GMO, organic, fat free, low carb meal!

Introduction

- Food composition databases (FCDBs)
- Internet recipes
- Information retrieval method

Related work

- Matching text concepts to an entry in a knowledge base has been addressed in many ways

- Muller et al. (2012) presented a system that automatically calculates the nutritional content of recipes sourced on Internet
  - 6 human assessors manually evaluate list of ingredients for ambiguous ingredient names
  - 1,515 positively classified instances to witch they added the same number of negatively classified instances
  - features extraction and penalized regression model
  - 91.1% of the recipes they used were matched completely

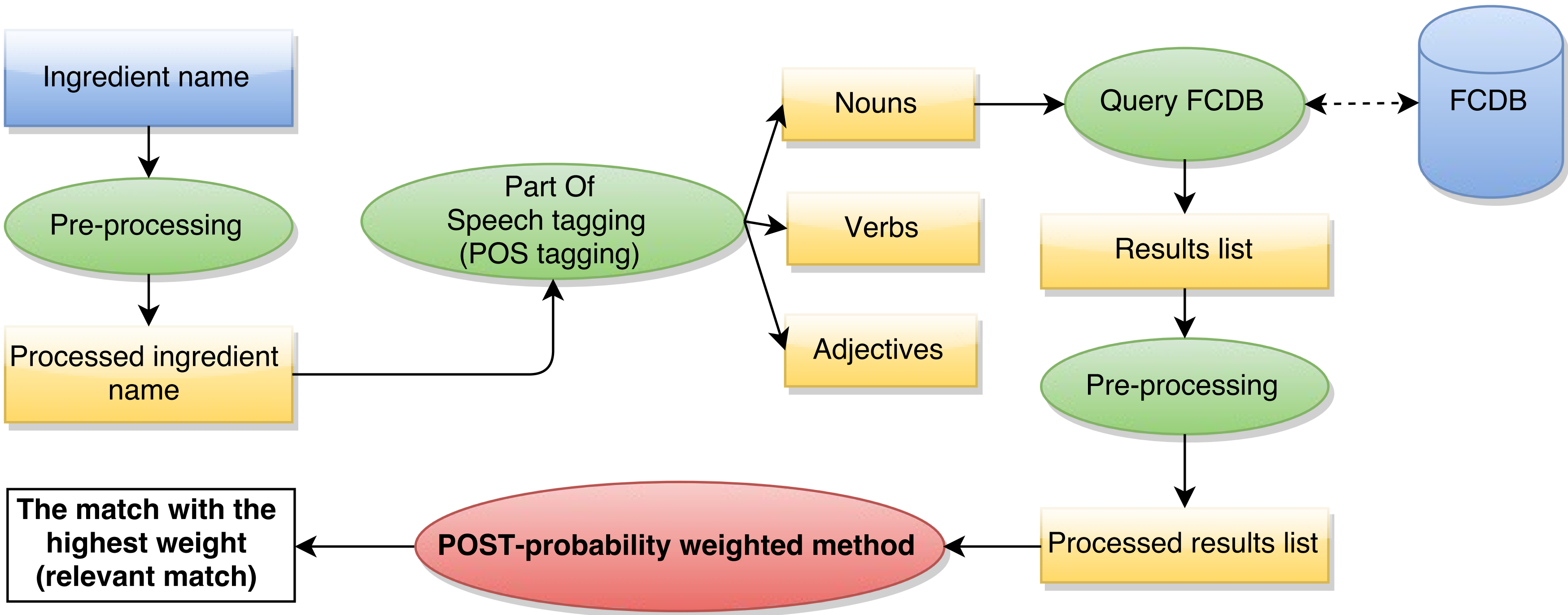
Problem definition

- People use human language to write the names of the used ingredients
  - “salt - iodised”, “iodised salt”, “salt, idodised”
- Ingredient synonymy problem
- The form of the ingredient and the cooking process

Proposed method (1/2)

- POS tagging (NN*,JJ*,VB*)
- String similarity - Jaccard index
- Laplace probability estimate

Proposed method (2/2)



Evaluation and results

- Data – collection of 721 recipes written in English [1]
- 1,615 different names of ingredients
- EuroFIR FCDB [2]
  - food table
  - ENGFDNM
- 44,033 English names of food analyses

[1] http://allrecipes.com/

[2] EuroFIR – non-profit Association under the Belgian law, http://www.eurofir.org/

Data pre-processing

- Remove punctuations
- Convert each name in lower case letters
- Whitespace tokenization
- Lemmatization, only for nouns
- Manually created rules (without skin; skinless), (with salt; salted)

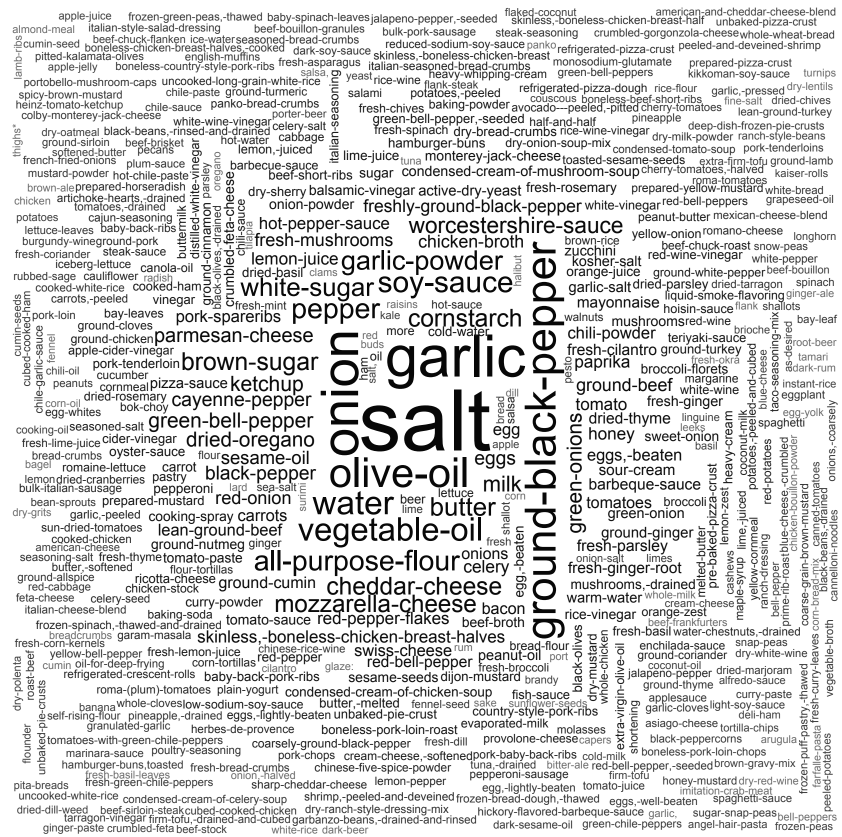
Experiment 1

Recipe ingredients	FCDB matching
Sweet italian sausage	Sausage, italian, sweet, links
Lean ground	Beef, ground, lean, raw
Onion	Onion
Tomatoes	Tomatoes
Tomato paste	Tomato paste
Canned tomato sauce	Tomato, sauce, canned
Water	Water
White sugar	Sugar, white
Dried basil leaves	Basil, dreid
Fennel seeds	Fennel, seeds
Italian seasoning	
Salt	Salt
Ground black pepper	Black peeper, ground Pepper, black, ground
Fresh parsley	Parsley, fresh
Lasagna noodels	Egg noodles Noodles egg-free
Ricotta cheese	Ricotta cheese Cheese, Riccota
Egg	Egg
Mozzarella cheese	Cheese, mozzarella Mozzarela cheese
Parmesan cheese	Cheese, parmesan Parmesan cheese

Recipe	FCDB	Code
gibanica	prekmurska gibanica	RECMEM000162
omaka sojina	sojina omaka tamari (iz soje)	16124
pomaranče	pomaranča	P0402
melona	melona casaba melona honeydew	9183 9184

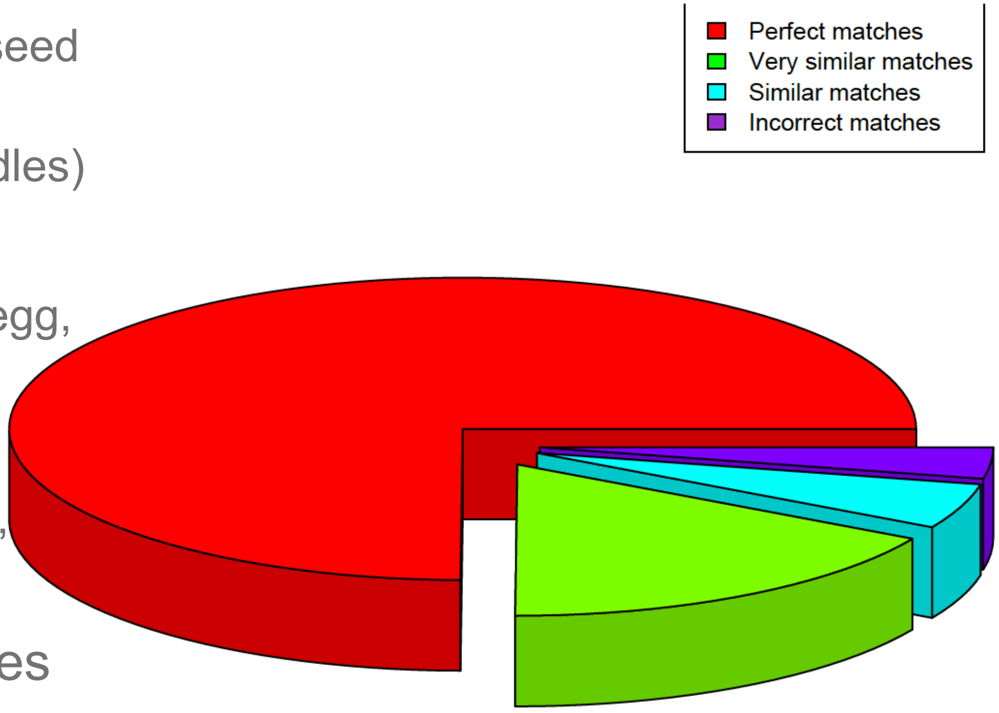
Experiment 2 (1/2)

- We divided the matches on 4 groups
- ***Perfect***, ***Very similar***, ***Similar*** and ***Incorrect*** matches
- Perfect - 1,210 (74.92%)
- Very similar - 273 (16.90 %)
- Similar - 78 (4.84%)
- Incorrect - 54 (3.34%)



Experiment 2 (2/2)

- Perfect matches
  - (black olives; olives, black)
  - (fresh ginger; ginger, fresh)
- Very similar matches
  - (fresh cilantro; spices, coriander seed (cilantro))
  - (uncooked egg noodles; egg noodles)
- Similar matches
  - (dry penne pasta; pasta, without egg, dry)
- Incorrect matches
  - (angel hair pasta; cake angelfood, commercially prepared)
- Perfect and very similar matches are 91.82% together



Conclusion

- Method for matching recipe ingredients with food composition data
- Can be use to explore what is missing in the FCDBs
- No need of labeled data
- The result can be used for feature selection and then solving classification problems

Quisper Ontology Learning from Personalized Dietary Web Services

Tome Eftimov, Barbara Koroušić Seljak, Peter Korošec

{tome.eftimov, barbara.korousic,peter.korosec}@ijs.si

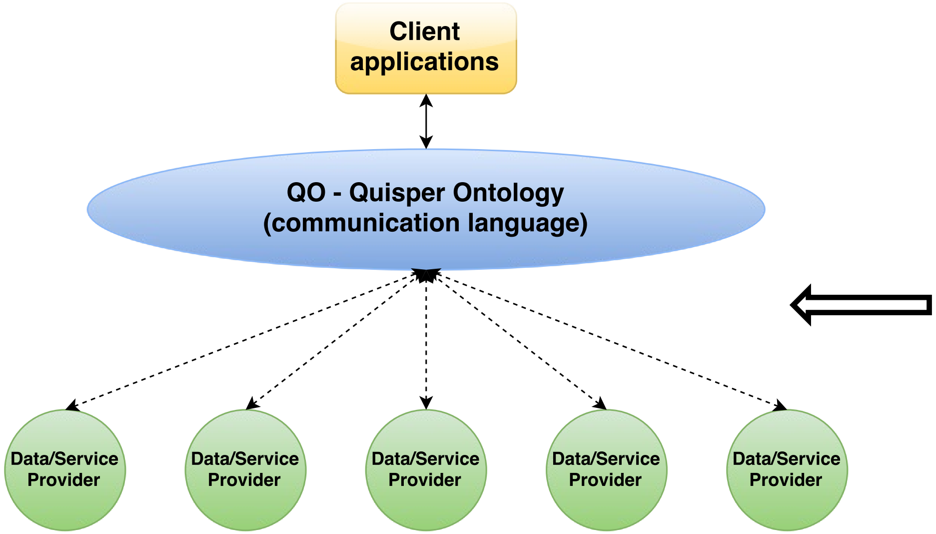
Paul Finglas

{paul.finglas}@ifr.ac.uk

Computer Systems Department, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia

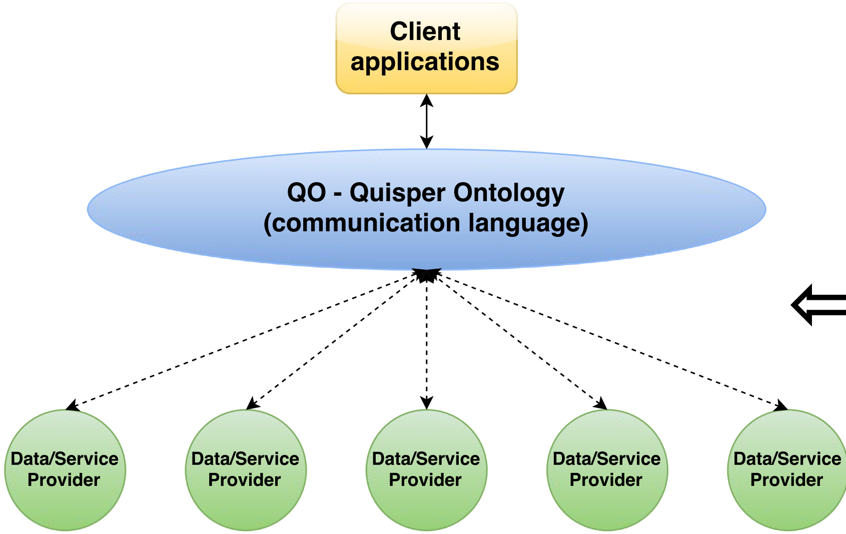
Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia

Quisper Ontology Learning



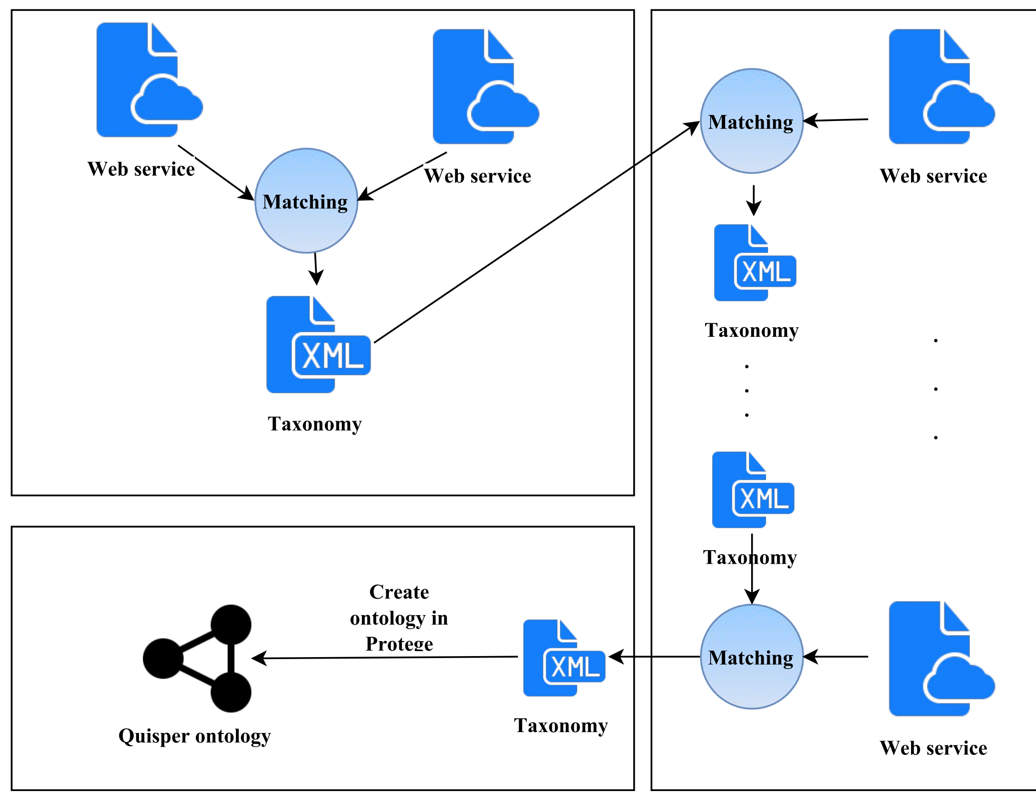
How to extract the knowledge from the data sources using different ways of describing and classifying the data ???

Quisper Ontology Learning



How to extract the knowledge from the data sources using different ways of describing and classifying the data ???

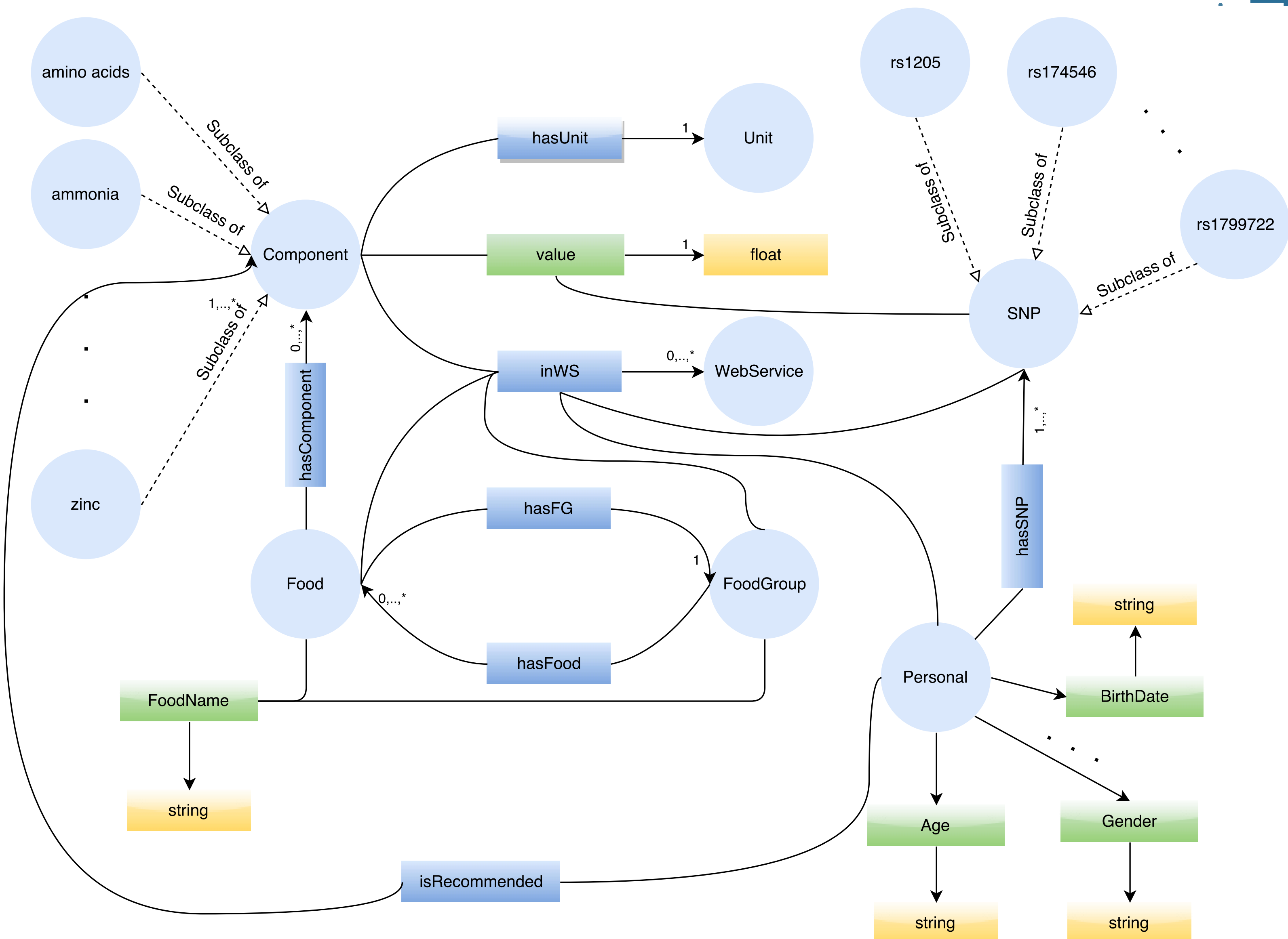
- Taxonomy
- Ontology



Quisper taxonomy

```xml
<Taxonomy>
  <WSproviders>
   <WSprovider name="EuroFIRFoodTransport"
   version="1.0.4"/>
   <WSprovider name="Food4me" version="1.0.0"/>
   <WSprovider name="SafeCape" version="1.0.0"/>
  </WSproviders>
  <Matching class="Component">
    <Descriptions>
       <Description source="EuroFIR"
       description="alpha-carotene,carotene,alpha-"
       code="CARTA"/>
       <Description source="Food4me"
       description="Alpha-carotene"
       code="413483001"/>
    </Descriptions>
    <MatchScore>
       <MatchPair>
         <WSprovider name="EuroFIRFoodTransport"/>
         <WSprovider name="Food4me"/>
         <matchScore>0.1875</matchScore>
       </MatchPair>
     </MatchScore>
 </Matching>
       .
       .
       .
</Taxonomy>
```

Quisper ontology



Thank you for your attention!!!