Domänen-spezifisches Vokabular *cueML*

Anforderungen
cueML

Zunächst werden hier die Anforderungen an ein Vokabular in der Koch-Domäne dargelegt, welche sich aus unserem Kochbuch ergeben. Da die zuvor vorgestellten Auszeichnugssprachen (AuszeichnenRezepte.html) diesen Anforderungen nicht genügen, stellen wir anschließend unser selbst entwickeltes *cueML*-Vokabular vor.

Anforderungen

Folgend erklären wir, was für uns ein Rezept ausmacht und was das Vokabular dementsprechend erfassen können muss. Die einzelnen Punkte werden zusammenfassend am Ende noch einmal aufgelistet.

Ein Rezept besteht für uns nicht nur aus reinem Text. Aus unserer Sicht **ist ein Rezept in Themenbereiche unterteilt**. Das Beispiel-Rezept aus Abb. 1 ist eindeutig in einen Bereich für ein Bild, eine Überschrift/Namen einen einleitenden Text, usw. aufgeteilt. Ein Rezept hat für uns im Allgemeinen folgende klar strukturierte und voneinander separierte Bereiche:

- Den Namen als Überschrift
- Optional einen einleitenden, möglichst ansprechenden Text wie zum Beispiel „Meine absoluten Lieblings-Pfannekuchen nach dem geheimen Rezept meiner Oma!“ (welches wohl spätestens nach der Veröffentlichung des Rezeptes nicht mehr ganz so geheim ist)
- Optional und sehr empfehlenswert ein Bild vom Gericht
- Geschätzte Zubereitungszeit
- Nährwertangaben pro Person
- Anzahl der produzierten Portionen
- Eine Zutatenliste
- Zubereitungs-Anweisungen

Ein Vokabular sollte diese Struktur erfassen können. Sofern in einem Rezept die Struktur noch nicht gegeben ist, sollten die im Rezept vorhandenen Informationen nach einer Auszeichnung mit dem Vokabular automatisch in so eine Struktur umgewandelt werden können.

Abb. 1: Beispiel Rezept (Nach: 06onkel von Chefkoch.de, 2017) (Quellen.html#ChefkochPfannekuchen)

In den Rezepten aus unserem Kochbuch ist keine Zutatenliste vorhanden. Die Zutaten sind allerdings in den Zubereitungs-Anweisungen zu finden. Dementsprechend sollte nach einer Auszeichnung mit dem Vokabular **eine Zutatenliste aus den Zubereitungs-Anweisungen eines Rezeptes extrahierbar sein**. Daraus ergeben sich weitere Anforderungen, wie folgende zwei Beispiel-Sätze verdeutlichen:

1. *„Der [Englische] Soja macht die Suppe gewürzreicher, kann jedoch gut wegbleiben, und statt Madeira kann man weißen Franzwein und etwas Rum nehmen.“* (Davidis, 1849, S. 33 f.) (Quellen.html#DavidisKochbuch)
2. *„Man kocht nach No. 1 eine gute Bouillon [...]“* (Davidis, 1849, S. 32) (Quellen.html#DavidisKochbuch)

In Erstens *kann* Soja zum würzen verwendet werden. Da Soja jedoch sehr geschmacksintensiv ist, hat das Gericht, je nachdem ob Soja verwendet wird, eine ganz andere Geschmacksrichtung. Dementsprechend wäre eine Festlegung, ob Soja in die Zutatenliste kommt oder nicht, eine Reduzierung des Rezeptes. Daher soll das Vokabular **optionale Zutaten** von Obligatorischen unterscheiden können.

Des Weiteren gehören in das Rezept nicht Madeira, weißer Franzwein *und* Rum, sondern Madeira *oder* weißer Franzwein und Rum. Also muss das Vokabular **alternative Zutaten** abbilden können.

Erstens enhält somit genau genommen vier mögliche, unterschiedliche Rezepte. Je zwei ob die Soja verwendet wird oder nicht, sowie je zwei ob Madeira oder weißer Franzwein und Rum verwendet wird. Nach der Auszeichnung sollen daher aus den Rezepten alle möglichen Rezepte abgeleitet werden können.

In Zweitens soll *nach No. 1* eine Bouillon gekocht werden. Rezept No. 1 ist jedoch nicht ein Rezept für Bouillon, sondern ein Rezept für Rindfleischsuppe, in welches das Rezept für Bouillon integriert ist (Davidis, 1849, S. 28) (Quellen.html#DavidisKochbuch) Dementsprechend soll das Vokabular **Verweise auf (Teil-) Rezepte** ermöglichen.

Für eine kulinarische Analyse sind offensichtlich nicht nur die Zutaten wichtig, sondern auch die Mengenangaben der Zutat, wobei eine Mengenangabe ohne Einheit wertlos ist. Daher wird im Vokabular zusätzlich eine Möglichkeit zur Auszeichnung von **Mengenangaben** und **Mengeneinheiten** benötigt.

Im Sinne des semantischen Webs ist es zudem wünschenswert, dass die **Zutaten und Mengeneinheiten als weltweit frei verfügbare und eindeutige Ressourcen** (mittels URIs) verstanden werden. Eine gute und anschauliche Erklärung des Begriffes *semantisches Web* ist in (Berners-Lee et al., 2001) (Quellen.html#SemanticWeb) zu finden. Zu der URI einer Zutat können Nährwertangaben und Einordnungen in Taxonomien wie „ist-vegetarisch“ hinterlegt werden. Die Mengeneinheiten sind erst dann sinnvoll verwendbar, wenn Informationen zur Umrechnung in standardisierte Einheiten vorliegen; z. B. die Mengenangabe *„für 8 Pfennig [...] Weißbrod“* (Davidis, 1849, S. 35 f.) (Quellen.html#DavidisKochbuch) ist ohne die Information, wie viel Gramm/Scheiben Weißbrot man nach Meinung von Frau Davidis für 8 Pfennig kriegt, wertlos. Erst die Verknüpfung beider Ressourcen ermöglicht eine transparente Extraktion von Nährwertangaben aus einer Zutatenliste.

Zusammengefasst ergeben sich daraus für uns folgende Anforderungen:

- Auszeichnung von Themenbereichen
- Unterscheidung von obligatorischen Zutaten, optionalen Zutaten und alternativen Zutaten
- Extraktion einer Zutatenliste aus den Zubereitungs-Anweisungen
- Verweise zwischen Rezepten
- Zutaten und Mengeneinheiten als frei verfügbare und eindeutige Ressourcen

cueML

Die zuvor vorgestellten Auszeichnungssprachen (AuszeichnenRezepte.html) erfüllen die oben genannten Anforderungen an das benötigte Vokabular nicht, wie folgend knapp an je einem Punkt erläutert wird: *TEI* erhält zwar die Struktur des Rezeptes, stellt allerdings keine Vokabeln zur Verfügung, um die einzelnen Themenbereiche wie Zutatenliste oder Zubereitungs-Anweisung als solche auszuzeichnen. *Schema.org/Recipe* bietet zwar die Möglichkeit Zutaten als solche zu markieren, aus dem eingeschlossenen Text der Markierungen kann der Computer jedoch weder die konkrete Zutat noch die Mengenangabe mit dazugehörender Einheit ableiten. Die eigenen Vokabulare von *Chefkoch.de* und *Cooking.nytimes.com* zeichnen die bestehenden Zutatenlisten aus, welche bei unserem Kochbuch nicht vorhanden sind. Des Weiteren sind ihre Vokabulare nicht frei verfügbar und somit von uns sowieso nicht verwendbar.

Daher haben wir **culinary editions Markup Language (cueML)** entwickelt, was netterweise wie Kümmel auszusprechen ist. Eine beispielhafte Auszeichnung von einem Rezept aus unserem Kochbuch mit cueML ist in Abb. 2 zu sehen.

Abb. 2:

Abb. 2a: Unsere Transkription eines Rezeptes aus dem Kochbuch

```
<cue:recipe type="Suppen." rcp-id="B-16">
   <head>Mock Turtle Suppe.</head>

   <p>Es wird hierzu für 24–30 Personen eine kräftige Bouillon von 8–10 Pfund Rindfleisch
      mit Wurzelwerk gekocht. Zugleich bringt man einen großen Kalbskopf, eine
      Schweineschnauze und Ohren, einen Ochsengaumen und eine geräucherte Ochsenzunge zu
      Feuer und kocht dies Alles gahr, aber nicht zu weich. Kalt, schneidet man es in
      kleine, länglich viereckige Stückchen, gibt das Fleisch in die Bouillon, nebst
      braunem Gewürz, ein Paar Messerspitzen Cayenne-Pfeffer, einige Kalbsmidder in
      Stückchen geschnitten (siehe Vorbereitungsregeln), kleine Saucissen, so viel
      Kalbskopfbrühe, daß man hinreichend Suppe hat, und macht dies mit in Butter braun
      gemachtem Mehl gebunden. Nachdem dies Alles ¼ Stunde gekocht hat, kommen noch Klöße
      von Kalbfleisch, einige hart gekochte Eier in Würfel geschnitten, ein Paar Eßlöffel
      Engl. Soja hinzu, und wenn die Klößchen einige Minuten gekocht haben, ½ Flasche
      Madeira und auch Austern, wenn man sie haben kann. Dann wird die Suppe sogleich
      angerichtet. </p>

   <note>Anmerk. Der Soja macht die Suppe gewürzreicher, kann jedoch gut wegbleiben, und
      statt Madeira kann man weißen Franzwein und etwas Rum nehmen. Sowohl die Bouillon als
      Kalbskopf können schon am vorhergehenden Tage, ohne Nachtheil der Suppe, gekocht
      werden. </note>
</cue:recipe>
```

Abb. 2b: Gleiches Rezept mit cueML ausgezeichnet

```
<cue:recipe type="Suppen." rcp-id="B-16">
   <head>Mock Turtle Suppe.</head>

   <p>Es wird hierzu für <cue:recipeYield atLeast="24" atMost="30" unit="people">24–30
         Personen</cue:recipeYield> eine kräftige <cue:recipeIngredient target="#Bouillon"
         >Bouillon</cue:recipeIngredient> von 8–10 Pfund <cue:recipeIngredient
         ref="#Rindkochfleisch" atLeast="8" atMost="10" unit="Pfund"
         >Rindfleisch</cue:recipeIngredient> mit <cue:recipeIngredient ref="#Wurzelwerk"
         >Wurzelwerk</cue:recipeIngredient> gekocht. Zugleich bringt man einen großen
         <cue:recipeIngredient ref="#Kalbskopf" quantity="1"
         >Kalbskopf</cue:recipeIngredient>, [...]</p>

   <note>Anmerk. Der <cue:recipeIngredient ref="#Englische_Soja" optional="True"
         >Soja</cue:recipeIngredient> macht die Suppe gewürzreicher, kann jedoch gut
      wegbleiben, und statt <cue:recipeIngredient ref="#Madeira" altGrp="1"
         >Madeira</cue:recipeIngredient> kann man <cue:recipeIngredient
         ref="weißer_Franzwein" altGrp="2">weißen Franzwein</cue:recipeIngredient> und
      etwas <cue:recipeIngredient ref="#Rum" altGrp="2" quantity="etwas"
         >Rum</cue:recipeIngredient> nehmen<cue:alt target="1 2"/>. [...]</note>
</cue:recipe>
```

Es kombiniert und erweitert *TEI* und *Schema.org/Recipe*. Da die Transkription bereits in TEI vorliegt, ist es naheliegend TEI zu erweitern. Das Vokabular von Schema.org/Recipe übernehmen wir. So kann aus cueML für Suchmaschinen leicht zusätzlich eine Auszeichnung mit Schema.org/Recipe abgeleitet werden. Zusätzlich erweitern wir es um Attribute für Mengenangaben (*quantity*, bzw. *atLeast* und *atMost*) und Mengeneinheiten (*unit*), sowie Möglichkeiten um optionale (*optional="True"*) und alternative Zutaten (*altGrp="id"* und *alt="id1 id2 [...]"*) auszuzeichnen. Des Weiteren führen wir Möglichkeiten ein, um eine Zutat auf Bereiche des Kochbuches verweisen zu lassen (*target="#id"*) sowie allgemeine Verweise innerhalb eines Rezeptes kenntlich zu machen (ein in Abb. 2b nicht vorhandenes *ref*-Element).

Einen Ressourcenbestand von Zutaten, wie in den Anforderungen beschrieben, konnten wir leider nicht finden. Daher behelfen wir uns mit dem Bundeslebensmittelschlüssel (BLS) (https://www.blsdb.de/). Der BLS enthält zu knapp 15.000 Zutaten und Gerichten 38 Nährwert-Informationen wie z. B. Ballaststoffe pro 100g. An sich ist er nicht frei verfügbar, wir dürfen jedoch mit ihm im Rahmen einer akademischen Lizenz arbeiten.

Abb. 3 zeigt, wie wir den BLS in cueML integriert haben. Jede Zutat wird auf ein *ingredient*-Element abgebildet und das *BLSref*-Attribut ist ein Verweis auf den eindeutigen BLS-Schlüssel. Zusätzlich geben wir eine Liste von möglichen Lemmata für die Zutaten sowie optional ein erläuterndes *note*-Element an.

Abb. 3: Anbindung an den BLS

```
<cue:ingredient xml:id="Cayennepfeffer" BLSref="R252000">
    <cue:prefBasicForm>Cayennepfeffer</cue:prefBasicForm>
    <cue:altBasicForm>Cayenne-Pfeffer</cue:altBasicForm>
</cue:ingredient>
<cue:ingredient xml:id="Midder" BLSref="V582100">
    <cue:prefBasicForm>Midder</cue:prefBasicForm>
    <cue:altBasicForm>Kalbsmidder</cue:altBasicForm>
    <cue:altBasicForm>Bries</cue:altBasicForm>
    <cue:altBasicForm>Kalbsmilch</cue:altBasicForm>
    <cue:note>"Kalbsmidder ist auch unter dem Synonym: Bries, Kalbsmilch bekannt. Kalbsmilch ist
        die Thymusdrüse des Kalbes. 100 g frische Kalbsmilch enthalten 99,8 kcal, 3,4 g Fett,
        17,2 g Eiweiß, 77,8 g Wasser, 1,91 mg Eisen, 268 mg Cholesterin und 0,42 g Purine.
        Dieses Organ ist bei Jungtieren voll entwickelt. Bei erwachsenen Tieren bildet sich das
        Organ zurück. Kalbsmilch wird zur Herstellung von Spezialitäten, wie Suppen, Klöße und
        Ragout, verwendet." (http://www.cosmiq.de/qa/show/70827/was-ist-kalbsmidder/)</cue:note>
</cue:ingredient>
<cue:ingredient xml:id="Saucisse" BLSref="W000000">
    <cue:prefBasicForm>Saucisse</cue:prefBasicForm>
    <cue:note>Franz. Name für bestimmte Würste (s.
        https://fr.wikipedia.org/wiki/Saucisse)</cue:note>
</cue:ingredient>
<cue:ingredient xml:id="Butter" BLSref="Q610000">
    <cue:prefBasicForm>Butter</cue:prefBasicForm>
</cue:ingredient>
```

Einen Ressourcenbestand für Mengenangaben konnten wir ebenfalls nicht finden. Auf einen Ressourcenbestand, der Frau Davidis' Einheiten wie *1 Maß*, oder *für 8 Pfennig Weißbrod* umrechnet, wird nicht weiter eingegangen, da das nicht Teil dieser Informatik-Arbeit ist. Unabhängig davon sei erwähnt, dass es ganz im Sinne der Programmier-Prinzipien *Seperation of Concerns* und *DRY* keine gute Idee ist, die Umrechnung in der Auszeichnung vorzunehmen. Sollte sich beispielsweise bei späteren Recherchen ergeben, dass 1 Maß nach Frau Davids doch nicht 1l sondern 0,8 oder doch 1,2l entsprechen, müssten sämtliche ausgezeichnete Umrechnungen erneut vorgenommen werden. Bei einem Verweis auf eine URI, welche zu der Mengeneinheit 1 Maß von Frau Davidis gehört, muss nur der Eintrag bei der entsprechenden URI geändert werden. Daher übernehmen wir Frau Davidis' originalen Mengeneinheiten.

Im Gegensatz zu Schema.org/Recipe haben wir **cueML durch eine RELAX NG-Grammatik wohldefiniert**. Schema.org hat die Prämisse „some data is better than none“ (Schema.org, 2017) (Quellen.html#Schema.orgNoGrammar) und validiert daher nicht gegen eine Grammatik. Für Suchmaschinen, die möglichst viele Daten erfassen wollen und die Zielgruppe der Schema.org-Vokabulare sind, ist das eine vertretbare Prämisse. Unsere Arbeitsgruppe (https://comsys.informatik.uni-kiel.de/) ist dagegen der Überzeugung, dass die Validierung gegen eine Grammatik nicht schwer ist und einfache Fehler, wie beispielsweise Tippfehler verhindert. Darüber hinaus kann eine Grammatik gut als Dokumentation verwendet werden. Des Weiteren verstärkt sie die Idee des Vokabulares als Ontology. basisCueML.rng

(DavidisesKochbuch/cueML/basisCueML.rng) definiert die Kombination von TEI und Schema.org/Recipe und wurde von Prof. Dr.-Ing. Luttenberger mittels Roma (http://www.tei-c.org/Roma/) erstellt. cueML_v05.rng (DavidisesKochbuch/cueML/cueML_v0.5.rng) definiert die Erweiterungen unseres Vokabulares und inkludiert basisCueML.rng.