Jamie LOD-iver: Enriching Historical Recipes with Linked Open Data

Melvin Wevers, *KNAW Humanities Cluster, DHLab*
Marieke van Erp, *KNAW Humanities Cluster, DHLab*
Hugo Huurdeman, *University of Amsterdam, University of Oslo*
Richard Zijdeman, *International Institute for Social History, University of Stirling*

Introduction

What we eat has been an important part of our cultural identity.[1] There are monthly recipe magazines,[2] cooking shows,[3] and in Amsterdam alone, twenty-one culinary festivals are scheduled for 2018.[4] What we eat can be distilled from recipes, which are easily accessible for contemporary research, as recipes on food websites are marked up with hRecipe or schema.org.[5] Unfortunately, this is not the case for historical recipes, often available as part of digitized archives. This means that currently we are unable to answer important questions on how food culture has changed over time. Are we using more sugar in home cooking today? To what extent has migration affected eating habits?

Historical newspapers provide a lens on customs and habits of the past. For example, recipes published in newspapers highlight what and how we ate and thought about food. The challenge here is that newspaper data is often unstructured and highly varied, digitised historical newspapers add an additional challenge, namely that of fluctuations in OCR quality. Therefore, it is difficult to locate and extract recipes from them.

This poster presents a method for extracting and enriching over 24,000 recipes from four digitized historical newspapers (*Het Parool, Trouw, De Volkskrant,* and *NRC Handelsblad*) published between 1950 and 1995 via Delpher.[6] Based on distant supervision and automatically extracted lexicons, we present our approach to identify recipes in digitised historical newspapers, to generate recipe tags, and to extract ingredient information. Moreover, we offer OCR quality indicators and their impact on the extraction process, and we enrich the recipes with links to information on the ingredients. The poster will also demonstrate some initial analyses for which these recipes can be used.

Method: Extracting and Enriching Newspaper Recipes

To extract and enrich recipes from newspapers, we devised the workflow depicted in Figure 1.

---

[1] Whatmore, Sarah, and Lorraine Thorne. "Nourishing Networks: Alternative Geographies of Food." In *Globalising Food: Agrarian Questions and Global Restructuring*, edited by David Goodman and Michael Watts, 211–24. London: Routledge, 1997; Wilson, Thomas M., ed. *Food, Drink and Identity in Europe*. European Studies 22. Amsterdam: Rodopi, 2006; Otterloo, Anneke H. van. *Eten en eetlust in Nederland, 1840-1990: een historisch-sociologische studie*. Amsterdam: Bert Bakker, 1990.
[2] https://deliciousmagazine.nl/ https://www.foodiesmagazine.nl/
[3] https://heelhollandbakt.omroepmax.nl/
[4] https://www.iamsterdam.com/en/see-and-do/whats-on/festivals/overview-culinary-festivals-and-events
[5] http://microformats.org/wiki/hrecipe; http://schema.org/Recipe
[6] Because of higher OCR quality for more recent newspapers, we selected these newspapers. For more on these papers, see: https://www.kb.nl/nieuws/2017/belangrijke-naoorlogse-kranten-digitaal-beschikbaar

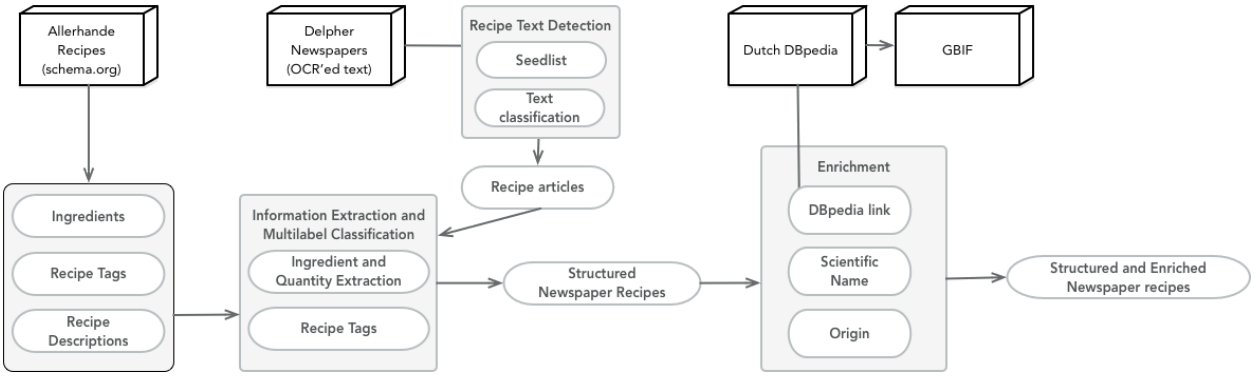

Figure 1: Newspaper recipe extraction & enrichment workflow

We first trained a text classifier on almost 7,500 recipes extracted from four national newspapers using keywords. This classifier was then used to extract over 24,000 recipes from the same national newspapers. Then, we bootstrapped 16K recipes from the *Allerhande* recipe website to obtain lists of ingredients and recipes belonging to specific to specific tags such as 'vegetarian', 'Christmas', and 'Italian'. The recipe descriptions and their associated tags were used to train a multilabel classifier, needed to assign such tags to the recipes from the newspapers. Next, we applied a rule-based tagger based on *Allerhande's* ingredients list to extract ingredients and their quantities from the recipes. The recipes were further enriched with photographs from Rijks Museum and Dutch DBpedia using SPARQL endpoints. From the latter source, we also extracted location information associated with the ingredients to enable a geographical mapping of recipes.

Analysing Recipes

Our poster shows how combining natural language processing, machine learning, and semantic web can be used to construct a rich dataset from heterogeneous newspapers for the historical analysis of food culture. By enriching the recipe information, we open up the possibility for comparative and quantitative analyses as well as rich exploration and visualization. Figure 2 visualizes a selection of the newly enriched recipes in a network diagram, organised by their ingredients. Additional text analysis such as Topic Modeling and Word Embeddings can guide qualitative studies of changing food culture. Future work could focus on extracting and enriching recipes from earlier periods and other sources than newspapers, unveiling the evolution of cuisines and the representation of cultural identities in discourse on food.

Figure 2: Recipes organised by their ingredients in a network diagram, using images from DBpedia