Domain-specific modeling: Towards a Food and Drink Gazetteer

Andrey Tagarev, Laura Toloşi, and Vladimir Alexiev

Ontotext AD, 47A Tsarigradsko Shosse, 1124 Sofia, Bulgaria
andrey.tagarev@ontotext.com

Abstract. Our goal is to build a Food and Drink (FD) gazetteer that can serve for classification of general, FD-related concepts, efficient faceted search or automated semantic enrichment. Fully supervised design of a domain-specific models *ex novo* is not scalable. Integration of several ready knowledge bases is tedious and does not ensure coverage. Completely data-driven approaches require a large amount of training data, which is not always available. In cases when the domain is not very specific (as the FD domain), re-using encyclopedic knowledge bases like Wikipedia may be a good idea. We propose here a semi-supervised approach, that uses a restricted Wikipedia as a base for the modeling, achieved by selecting a domain-relevant Wikipedia category as root for the model and all its subcategories, combined with expert and data-driven pruning of irrelevant categories.

Keywords: categorization, Wikipedia, Wikipedia categories, gazetteer, Europeana, Cultural Heritage, concept extraction

1 Introduction

Our work is motivated by the Europeana Food and Drink (EFD) project[1], which aims at categorizing food and drink-related concepts (FD), in order to digitalize, facilitate search and semantically enrich Cultural Heritage (CH) items pertaining to the 'food and drink' theme. Even though driven by the application to FD, our approach is easily generalizable to any domain that is encyclopedic in nature. For example, we can apply the approach for categorizing 'Arts', 'Sports', etc.

Modeling a domain from scratch requires interdisciplinary expertise, both in the particular domain and in knowledge-base modeling. Also, it is a tedious, time-consuming process. When the domain is very specialized, for example 'technical specifications of car building', or 'human genes', probably the process is unavoidable. However, for broader domains like 'Food and drink' (but also 'Arts', 'Sports'), we believe that using encyclopedic, LOD data is a better, more scalable approach. To model 'Food and Drink' concepts, we used Wikipedia.

Wikipedia is a great collection of general knowledge concepts. It is freely available and easily editable by anyone. The volume of information is enormous.

[1] http://foodanddrinkeurope.eu/

E.g. the English wiki has a total of 35M pages, of which 30M are auxiliary (discussions, sub-projects, categories, etc). Overall, Wikipedia has some 35M articles in over 240 languages. Multilingualism is a very important aspect that recommends the usage of Wikipedia, as CH objects at Europeana are annotated in eleven languages.

In this paper we describe the method, we show preliminary results and we present a critical discussion on the suitability of Wikipedia for the purpose of FD categorization. In Section 2, we describe the EFD application. In Section 3, we present the steps of the method. Section 4 presents some insightful (from both technical and application perspective) properties of the sub-hierarchy generated from the *Food_and_drink* root. We continue by describing the supervised curation of the domain in Section 5 and by showing the results of the data-driven enrichment analysis in Section 6. We will conclude the paper with comments and outlook in Section 7.

Next, we mention relevant articles that have addressed domain-specific modeling in the past.

1.1 Related work

Much work has been dedicated to building domain specific knowledge bases. Earliest approaches were fully supervised, domain experts defining *ex-novo* the classification model. With the development of modern NLP techniques such as concept disambiguation, concept tagging or relation extraction, semi-supervised and even unsupervised methods are emerging. For example, there are many methods for automated merging and integration of already existing ontologies ([2], [8], [9]). In [4], a semi-supervised method for enriching existing ontologies with concepts from text is presented. More ambitious approaches propose unsupervised generation of ontologies ([5] [10]), using deep NLP methods. In [7], a method is described, for generating lightweight ontologies by mapping concepts from documents to LOD data like Freebase and DBpedia and then generating a meaningful taxonomy that covers the concepts.

Classification of Food and Drink is not a new problem. There exist quite a number of resources that belong to the domain, some of them more specific than others. [3] have proposed a cooking ontology, focused on: food (ingredients), kitchen utensils, recipes, cooking actions. http://www.bbc.co.uk/ontologies/fo is a lightweight food ontology by BBC.

2 Europeana Food and Drink

The Europeana Food and Drink[2] Classification scheme (EFD classification) [6] is a multi-dimensional scheme for discovering and classifying Cultural Heritage Objects (CHO) related to Food and Drink (FD). The project makes use of innovative semantic technologies to automate the extraction of terms and co-references.

[2] http://foodanddrinkeurope.eu/

The result is a body of semantically-enriched metadata that can support a wider range of multilingual applications such as search, discovery and browse.

The FD domain is generously broad and familiar, in the sense that any human can name hundreds of concepts that should be covered by the model: ‘bread’, ‘wine’, ‘fork’, ‘restaurant’, ‘table’, ‘chicken, ‘bar’, ‘Thanksgiving dinner’, etc. In our particular application however, the model is required to cover also a large variety of cultural objects related to FD, many of which exist nowadays only in museums and are known only by a number of experts. These are described in content coming from a variety of Cultural Heritage organizations (CHO), ranging from Ministries to academic libraries and specialist museums to picture libraries. The content represents a significant number of European nations and cultures, it comprises objects illustrating FD heritage, recipes, artworks, photographs, some audio and video content and advertising relating to FD. It is heterogeneous in types and significance, but with the common thread of FD heritage and its cultural and social meaning. Metadata are available partly in English and native languages, with almost half of metadata only available in native languages.

Content is heterogeneous and varied. Examples include [6]: books on Bovine care and feeding (TEL, http://www.theeuropeanlibrary.org/tel4/), book on tubers/roots used by New Zealand aboriginals (RLUK, http://www.rluk.ac.uk/), self-portraits involving some food (Slovak National Gallery, http://www.sng.sk/en/uvod), traditional recipes for Christmas-related foods (Ontotext, www.ontotext.com), colorful pasta arrangements (Horniman, www.horniman.ac.uk/), mortar used to mix lime with tobacco to enhance its psychogenic compounds (Horniman), food pounder cut from coral, noted for its ergonomic design (Horniman), horse made from cheese (Horniman), a composition of man with roosters/geese made from bread (Horniman), poems about food and love, photos of old people having dinner, photos of packers on a wharf, photos of Parisian cafes, photos of a shepherd tending goats, photos of a vintner in his winery, medieval cook book (manuscript), commercial label/ad for consommé, etc.

2.1 Wikipedia categories related to FD

Wikipedia categories live in the namespace https://en.wikipedia.org/wiki/Category: (note the colon at the end). We discovered a number of FD categories, amongst them: *Food and drink*, *Beverages*, *Ceremonial food and drink*, *Christmas food*, *Christmas meals and feasts*, *Cooking utensils*, *Drinking culture*, *Eating parties*, *Eating utensils*, *Food and drink preparation*, *Food culture*, *Food festivals*, *Food services occupations*, *Foods*, *History of food and drink*, *Holiday foods*, *Meals*, *Works about food and drink*, *World cuisine*. Other interesting categories: *Religious food and drink*, *Food law*: topics like halal, kashrut, designation of origin, religion-based ideas, fisheries laws, agricultural laws, food and drug administration, labeling regulations, etc., *Food politics*, *Drink and drive songs*, *Food museums*. We selected https://en.wikipedia.org/wiki/Category:Food_and_drink as the root of our FD restricted model, considering that all above-mentioned categories are its direct or indirect subcategories.

Table 1: Wikipedia: statistics concerning categories.

Wikipedia	articles	cats	per art	art-cat	per art	per cat	cat-cat	per cat
English	4,774,396	1,122,598	4.25	18,731,750	3.92	16.69	2,268,299	2.02
Dutch	1,804,691	89,906	20.07	2,629,632	1.46	29.25	186,400	2.07
French	1,579,555	278,713	5.67	4,625,524	2.93	16.60	465,931	1.67
Italian	1,164,000	258,210	4.51	1,597,716	1.37	6.19	486,786	1.89
Spanish	1,148,856	396,214	2.90	4,145,977	3.61	10.46	675,380	1.7
Polish	1,082,000	2,217,382	0.49	20,149,374	18.62	9.09	4,361,474	1.97
Bulgarian	170,174	37,139	4.58	387,023	2.27	10.42	73,228	1.97
Greek	102,077	17,616	5.79	182,023	1.78	10.33	35,761	2.03

3 A method for domain-specific modeling

Wikipedia is loosely structured information. It has very elaborate editorial policies and practices, but their major goal is to create modular text that is consistent, attested (referenced to primary sources), relatively easy to manage. A huge number of templates and other MediaWiki mechanisms are used for this purpose. The structured parts of Wikipedia that can be reused by machines are: *i*) Links (wiki links, inter-language links, providing language correspondence, inter-wiki links, referring to another Wikipedia or another Wikimedia project e.g. Wiktionary, Wikibooks, external links), *ii*) Informative templates, in particular Infoboxes; *iii*) Categories; *iv*) Lists; *v*) Portals, Projects, Tables.

There are several efforts to extract structured data from Wikpedia. E.g. the Wikipedia Mining software [3] [11] allows extraction of focused or limited information. For our purpose, we prefer to use data sets that are already structured, like DBpedia. The data in RDF format is easily loaded in Ontotext GraphDB [4], which allows semantic integration of the data and easier querying.

3.1 Wikipedia categories

Category statistics for Wikipedia are presented in Table 1. 'Articles' denotes the number of content pages, 'cats' refer to the number of category pages, 'per art' is the ratio of articles to categories, 'art-cat' represents the number of assignments of a category as 'parent' of an article, 'per art' denotes the number of category assignments per article, 'per cat' stands for number of articles assigned per category, 'cat-cat' is the number of assignments of a category as 'parent' of another category, 'per cat' represents the number of parent categories per category.

[3] http://sourceforge.net/projects/wikipedia-miner
[4] http://ontotext.com/products/ontotext-graphdb/

3.2 Method overview

Our approach to domain-specific modeling is aimed at selecting a sub-hierarchy of Wikipedia, rooted at a relevant category, that covers well the domain concepts. The procedure follows the steps below:

1. Start by selecting the maximally general Wikipedia category that best describes the domain to ensure coverage. We will refer to this category as *root*.
2. Traverse Wikipedia by starting from the *root* and following `skos:broader` relations between categories to collect all *children* (i.e. sub-categories of the *root*). We also remove cycles to create a directed acyclic graph and calculate useful node metadata like *level* (i.e. shortest path from root), number of unique subcategories, etc.
3. Top-down curation: perform manual curation by an expert of the top (few hundred) categories to remove the ones irrelevant to the domain.
4. Bottom-up enrichment: map domain-related concepts to Wikipedia articles and evaluate enrichment in concepts mapped to each category. Thus, we automatically evaluate the relevance of categories, by direct evidence.

Technical details:

Step 2. Breadth-first (BF) traversal selects all categories reachable from the root. In order to obtain the domain categorization, we keep all possible edges defined by the `skos:broader` relation, but remove edges that create cycles. Cycles are logically incompatible with the SKOS system, but are not forbidden and exist in Wikipedia due to bad practice or lack of control. In order to remove cycles, we check that a potential child of the current node of the BF procedure is not also its ancestor before adding the connection. The average number of children of a category is 2.02, therefore we expect the number of categories to grow exponentially with each level until the majority of connections start being discarded for being cyclical.

Step 3. We generate a list of the few hundred most important categories (based on being close to the *root* or having many descendants) that are judged for relevance by an expert. Ones judged irrelevant are marked for removal. Removal of a category consists of a standard node-removal procedure in a directed graph, meaning that all node metadata including all incoming and outgoing edges are deleted and the node is marked as irrelevant in the repository (to be omitted in future builds). As a consequence, the sub-hierarchy may split into two or more connected components, one of which contains the root, the others being rooted at the children of the removed category. In such a case, we discard all connected components, except for the one starting at the initial *root*. The expert curation potentially drastically reduces the size of the sub-hierarchy with minimal work, thus being an efficient early method for pruning.

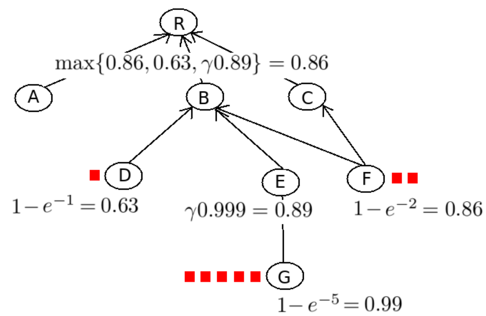


Fig. 1: Scoring categories bottom-up. Concepts mapped to categories are marked with red squares. Categories are marked with circles and named with capital letters.

Step 4. Moving away from the *root*, the number of categories of the domain hierarchy grows exponentially. Manually checking the validity of the categories w.r.t. the domain becomes infeasible. We propose a data-driven approach here: given a large collection of documents, thesauri, etc. relevant to the domain, we use a general tagging algorithm to map concepts from the collection to the hierarchy. Categories to which concepts are mapped are likely to belong to the domain, supported by evidence. For the categories to which no concepts have been mapped, we can infer their validity by using evidence mapped to children or even more distant descendants. For a leaf category (with no children) X with t concepts directly mapped to it, the score is computed as :

$$score(X) = 1 - e^{-t} \tag{1}$$

For a category Y with children $Y_1, ..., Y_n$ and t directly mapped concepts, the score is computed as:

$$score(Y) = \max\{1 - e^{-t}, \max_{i=1}^{n}\{\gamma score(Y_i)\}\}, \tag{2}$$

where $\gamma \in (0, 1)$ is a decay factor, that decreases the score of categories as they get further away from descendants with evidence (i.e. mapped concepts). Figure 1 illustrates an example, where the scores of leaf categories D, E, F are computed based on Equation 1 and the scores of categories E and B are computed using Equation 2. The scores can be used for automatically pruning categories that have a score under a certain threshold, where the threshold is level-specific.

4 Properties of the FD classification hierarchy

Following the method described in Section 3.2 we generated the *FD hierarchy*. We retrieved 887523 categories, at 26 levels under the FD root. The distribution of the number of categories by level is unimodal, peaking at the 16^{th} level, where we retrieve about $200,000$ categories (see Figure 2). The average number of subcategories of a category is 2.36. This means that the initially generated *FD hierarchy* contains about 80% of all categories in the English Wikipedia (see Table 1).

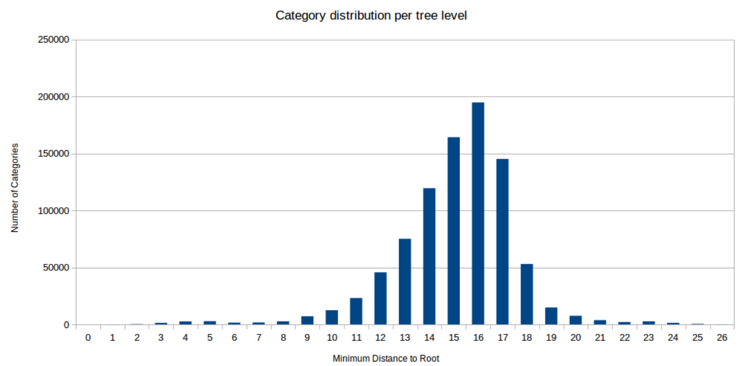


Fig. 2: Distribution of the shortest-path length from categories to the FD root category.

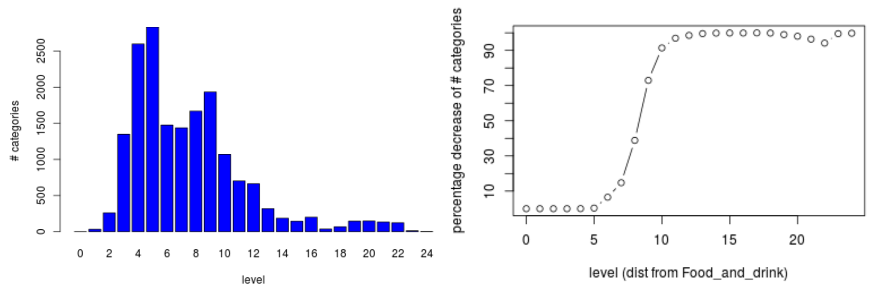
Not all subcategories reachable from the selected root are relevant to the domain, especially for such a broad domain like FD. We discuss below connections to categories not relevant to FD.

4.1 Examples of irrelevant inclusions

Due to wrong hierarchy. For instance, it is easy to reach football teams starting from FD in Wikipedia, following the chain: *Food_and_drink* → *Food_politics* → *Water_and_politics* → *Water_and_the_environment* → *Water_management* → *Water_treatment* → *Euthenics* → *Personal_life* → *Leisure* → *Sports* → *Sports_by_type* → *Team_sports* → *Football*.

The above chain contains a wrong supercategory assignment, concerning 'Euthenics'. *Euthenics* is the study of the improvement of human functioning and well-being by improvement of living conditions.*Personal life*, *Leisure* and *Sports* are correctly subcategories of *Euthenics*. But *Water treatment* should not be a supercategory of *Euthenics*. This issue was fixed on June 12, 2014 by removing *Euthenics* from *Water_treatment*. However, likely similar problems still exist elsewhere.

Due to partial inclusion. *Food_and_drink* has child *Animal_products*. Only about half of the children of *Animal_products* are relevant to the FD domain: *Animal-based_seafood*, *Dairy_products*, *Eggs_(food)*, *Fish_products*, *Meat*. Some are definitely not appropriate to FD: *Animal_dyes*, *Animal_hair_products*, *Animal_waste_products*, *Bird_products*, *Bone_products*, *Coral_islands*, *Coral_reefs*, *Hides*. Finally, there are



(a) Number of categories per level after expert curation.

(b) Decrease of number of categories per level after expert curation.

some mixed subcategories that may include both relevant and irrelevant children: *Animal_glandular_products*: milk and its thousands of subcategories is; castoreum is not; *Insect_products*: honey is, silk is not; *Mollusc_products*: clams and oysters are; pearls are not.

Due to non-human food or eating. Foods and drink explicitly includes animal feeding, thus not all are foods for humans, e.g. *Animal_feed*. The subcategory *Eating_behaviors* has some appropriate children, e.g. *Diets*, *Eating_disorders*, but has also some inappropriate children, e.g. *Carnivory*, *Detritivores*.

5 Top-down expert pruning

Supervised pruning of irrelevant categories becomes more efficient as experts are presented 'heavier' categories first; therefore we used a heuristic measure for the number of Wikipedia articles reachable from a certain category and provided them to the expert in descending order for judgement. This way, if an irrelevant category is removed, we can expect a drastic decrease of the number of nodes. The expert judged 239 of the top 250 categories in the list as irrelevant to the EFD topic. After removing them, we obtained a more focused hierarchy containing 17542 unique categories, therefore achieving a 50-fold decrease, with an hour of work from a human expert.

The new cardinalities per level are shown in Figure 5 a). Figure 5 b) reveals the levels at which the curation has the largest effect: starting with level 8, the decrease is larger than 50% and from level 11, the decrease is larger than 90%.

Further refinement of the FD hierarchy can be performed by an expert using the specially designed drill-down UI shown in Figure 4. It starts with the *root* category and displays a node's child categories ordered by our heuristic measure of weight and all articles directly linked to the node. The user can drill-down on categories to expand them in the same way and quickly mark them as irrelevant which removes them from the repository and UI.

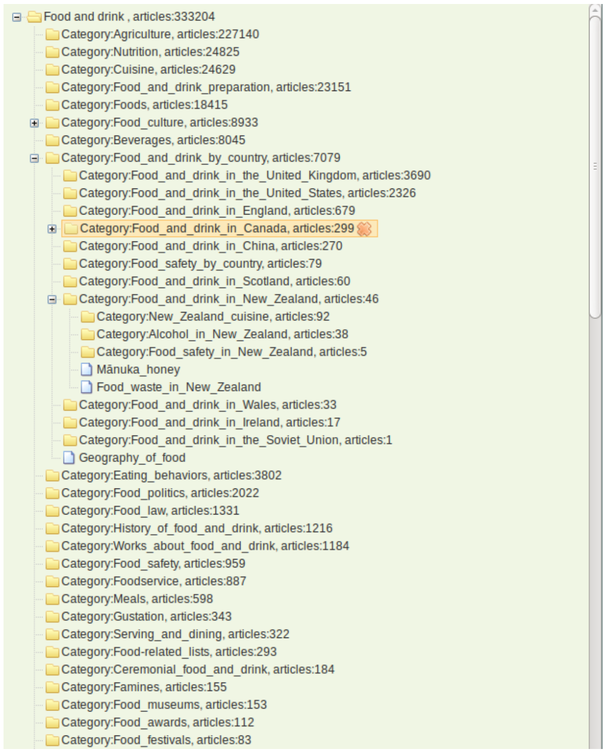


Fig. 4: Visualization interface for the FD categorization.

Fig. 5: Shark hook, an object from the Horniman museum http://www.horniman.ac.uk/collections/browse-our-collections/object/136887.

6 Bottom-up data-driven enrichment

A data-driven approach for estimating category relevance was described at Step 3 of our method (see Section 3.2). To demonstrate the approach, we considered the Horniman thesaurus, consisting of about 700 concepts used for describing Horniman museum artefacts.

The Horniman thesaurus is a shallow hierarchy consisting of four levels. At the second level, the classification is most informative: *agriculture and forestry*, *domestication of animals*, *food processing and storage*, *food service*, *hunting, fishing and trapping*, *narcotics and intoxicants: drinking*. For example, the object *shark hook* (Figure 5) belongs to the following path: *tools and equipment: general*, *hunting, fishing and trapping*, *fish hooks*, *shark hooks*.

6.1 Mapping the Horniman thesaurus to Wikipedia articles

We use Ontotext general-purpose concept extractor[5], that identifies Wikipedia concepts in general text. For the purpose, we concatenated all thesaurus terms

[5] http://tag.ontotext.com/

into several pseudo-documents, grouped by the second level category. The concept extractor relies on the context of each candidate for disambiguation, in the sense that the word 'mate' form the thesaurus entry 'mate teapot' would be mapped to http://en.wikipedia.org/wiki/Mate_(beverage), in the context of other terms regarding drinking, and not to other senses, listed in the disambiguation page http://en.wikipedia.org/wiki/Mate. In order to create context, we delimited the thesaurus terms in the pseudo-documents by comma (','). Eg., the pseudo-document for 'hunting, fishing and trapping' starts with:

'hunt and fishing trap, fishing net, spring trap, mantrap, mole trap, spear, fish spear, eel spear, elephant spear, spike wheel trap, spindle, snare trap, marmot snare, bird snare, sinker, net sinker, sheath, hunting knife sheath, shellfish rake, clam digger, sample, arrow poison, reel, quiver, poison, no-return trap, fish trap, nose clip, net, hunting net, hand net, fishing net, dip net, pig net, pigeon net, scoop net, line, fish line, lure, fly, cuttlefish lure, knife, hunting knife, keep, rat trap, fishing rod, float, line float, net float, fishing float, fish hook, ice-hole hook, halibut hook, gorge, pike hook, salmon hook, shark hook...'

The concept extractor returned 337 unique Wikipedia concepts, with an estimated Precision 0.91 of and estimated Recall of 0.7. For example, *shellfish rakes*: correctly identifies https://en.wikipedia.org/wiki/Shellfish, but incorrectly returns the redirect https://en.wikipedia.org/wiki/Train for rake, instead of https://en.wikipedia.org/wiki/Rake_(tool).

6.2 Scoring FD categories w.r.t. Horniman concepts mapped

Of all 337 concepts, 219 are in the FD hierarchy. Using our scoring scheme, we 'activated' 451 categories on the path to the FD root. The highest scoring are:

Cooking_utensils (1.00), *Teaware* (0.99), *Serving_and_dining* (0.99), *Cooking_appliances* (0.99), *Drinkware* (0.99), *Staple_foods* (0.99), *Tropical_agriculture* (0.99), *Gardening_tools* (0.99), *Fishing_equipment* (0.99), *Cooking_techniques* (0.99), *Cookware_and_bakeware* (0.99), *Crockery* 0.99), *Kitchenware* (0.99), *Spoons* (0.99), *Fishing_techniques_and_methods* (0.99), *Crops* (0.99), *Spices* (0.98) *Agricultural_machinery* (0.98), *Commercial_fish* (0.98), *Eating_utensils* (0.98), *Food_storage_containers* (0.98), *Serving_utensils* (0.98), *Animal_trapping* (0.98), *Food_and_drink* (0.95), *Recreational_fishing* (0.95), *Breads* (0.95), *Hunting* (0.95), *Dairy_products* (0.95), *Food_ingredients* (0.95), ...

Note that we retrieve Wikipedia categories concerning the broad topics of the Horniman thesaurus that were not explicitly input to our method: agriculture, domestic animals, food processing and storage, hunting and fishing, drinking. Figure 6 shows all the categories up to the FD root that get 'activated' by the bottom-up scoring, meaning that they get a positive score.

7 Comments and future work

We presented ongoing work on developing a FD categorization, with the purpose of classifying Cultural Heritage items from Europeana. To this end, we introduced a lightweight, SKOS categorization that borrows Wikipedia categories

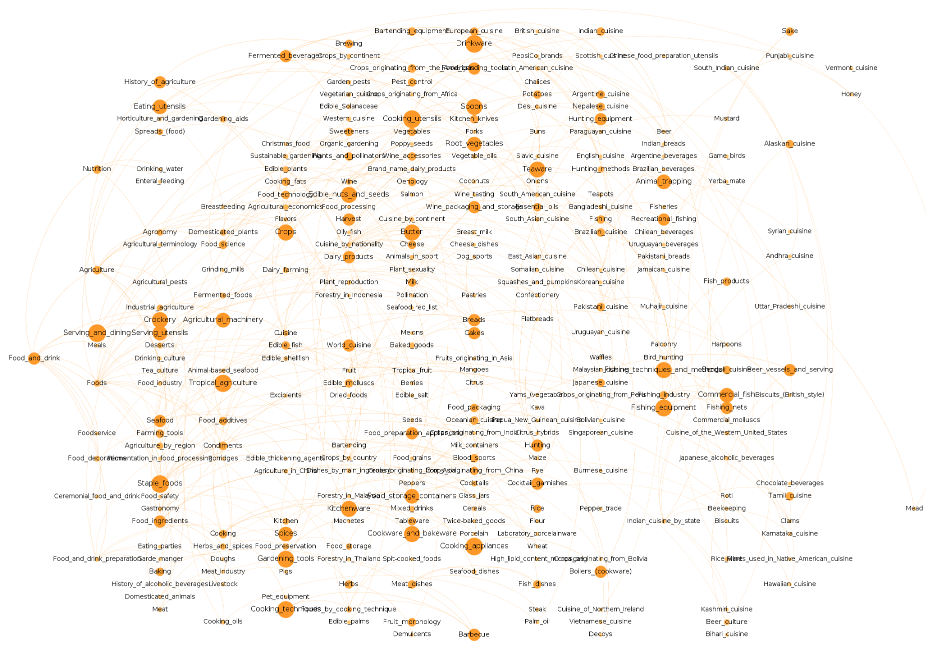


Fig. 6: Paths to *Food_and_drink*, activated by the bottom-up scoring scheme.

related to FD. Our preliminary results show that Wikipedia categories are rich enough to provide a good initial coverage of the domain. In fact, we showed that there are a large number of irrelevant categories that need to be removed by supervised curation. We developed an interactive visualization tool that allows experts to remove irrelevant categories and update the knowledge base.

We also presented a bottom-up, data-driven method for scoring categories with respect to concepts identified in Cultural Heritage collections, such as Horniman museum artefacts. We showed that by using this scoring scheme, a subhierarchy of FD is supported by evidence and thus confirmed to belong to the domain. This of course does not mean that the remaining categories are not food-and-drink relevant. Clearly, as more resources (e.g. recipes, books, see Section 2) are being processed and mapped to our classification scheme, more evidence will be gathered, for more accurate estimation of relevance of categories.

Despite the reasonable coverage of the domain (a qualitative remark, based on our observations! A quantitative estimate is hard to obtain), we identified concepts – or sets of concepts – that belong to FD, but are not found in our classification scheme. For example, some hunting weapons are not accessible directly by traversing Wikipedia from *Food_and_drink*. Some Horniman items representing spears could not be tagged, and they should, being tools for obtaining food. We plan to add secondary roots to the categorization, in this case 'Hunting' would be a good candidate. A direct, custom connection of type `broader` to the *Food_and_Drink* root is a possible way to add secondary roots.

A big challenge to the EFD project is building a multilingual categorization, for up to 11 languages. Our prototype is currently limited to English, but we believe extension is not hard, as we will take advantage of the 'parallel' Wikipedias for other languages. A possible approach for language X is to use all Wikipedia articles currently mapped to the English FD, get their correspondents in language X and start building the hierarchies bottom-up, to the corresponding FD root in language X. Thus, we ensure that all concepts from the English categorization would be covered by the categorization in language X. Of course, we would keep in mind that language-specific concepts may not be covered and thus may need to be added, either by some data-driven procedure, or supervised.

Acknowledgements The research presented in this paper was carried out as part of the Europeana Food and Drink project (http://foodanddrinkeurope.eu/) of the Europeana Fundation, co-funded by the European Commission within the ICT Policy Support Programme (CIP-ICT-PSP-2013-7) under Grant Agreement no. 621023.

References

1. Agirre, E., Barrena, A., De Lacalle, OL., Soroa, A., Fern, S., Stevenson, M., Matching Cultural Heritage items to Wikipedia. 2012
2. Cheng, CP., Lau, GT., Pan, J, Law, KH. Jones, A., Domain-Specific Ontology Mapping by Corpus-Based Semantic Similarity.
3. Ribeiro, R. and Batista, F. and Pardal, JP. and Mamede, NJ. and Pinto, HS., Cooking an Ontology, Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Computer Science 4183, 2006, pp 213–221, .
4. Parekh, V., Gwo, J., Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies, In Proceedings of the International Conference of Information and Knowledge Engineering, 2004.
5. Mousavi, H., Kerr, D., Iseli, M., Zaniolo, C., Harvesting Domain Specific Ontologies from Text, ICSC '14, 211–218, 2014.
6. Alexiev Europeana Food and Drink Classification Scheme, http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Classification-Scheme-(D2.2).pdf Europeana Food and Drink project, Deliverable D2.2, 2015.
7. Medelyan, O., Manion, S., Broekstra, J., Divoli, A., Huang, AL., Witten, IH., Constructing a Focused Taxonomy from a Document Collection. The Semantic Web: Semantics and Big Data, Lecture Notes in Computer Science 7882, 2013, 367–381.
8. Fridman Noy, N., Musen, MA., An Algorithm for Merging and Aligning Ontologies: Automation and Tool Support, In Proc. of the Workshop on Ontology Management at the 16th National Conference on Artificial Intelligence (AAAI-99), 1999.
9. Pinto, HS., Martins, JP., A Methodology for Ontology Integration, Proceedings of the 1st International Conference on Knowledge Capture, K-CAP '01, 2001.
10. Mousavi,H., Kerr, D., Iseli, M., Zaniolo, C., OntoHarvester: An unsupervised ontology generator from free text, CSD Technical Report #130003, University of California, Los Angeles 2013.
11. Medelyan, O., Milne, D., Legg, C., Witten, IH., Mining Meaning from Wikipedia, Int. J. Hum.-Comput. Stud., vol. 67(9), 2009, pp 716–754.