Domain-specific modeling: a Food and Drink Gazetteer

Andrey Tagarev, Laura Tolo̧si, 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 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. For general domains (such 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. For this purpose a maximally generic Food and Drink gazetteer needed to be created.

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. This might be unavoidable when working with narrowly-defined expert domains such as 'Art Nouveau' or 'Human Genes' but in this paper we describe a better approach for generating such gazetteers for broad or fuzzily-defined domains like 'Food and Drink', 'Arts', 'Sports', 'History', etc. While all examples in this paper deal with the specific FD application, the approach is easily generalizable to any domain with such an encyclopedic nature. In fact, the method presented in this paper is scalable, works with Linked Open Data and is semi-automated requiring only minimal input from domain-experts which makes it preferable for such applications.

[1] http://foodanddrinkeurope.eu/

To model encyclopedic domains we used the DBpedia project which is a great collection of general knowledge concepts extracted from the structured data of Wikipedia articles and categories. It is freely available and easily editable by anyone. The volume of information is enormous, 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 in EFD come in eleven languages.

In this paper we give a detailed description of a method for generating gazetteers of broad domains, demonstrate the application of this method to generating a Food and Drink gazetteer, and present a critical discussion of the suitability of Wikipedia for this purpose. In Section 2, we describe the EFD application that motivated the creation of our method. In Section 3 contains a step-by-step description of the method itself. Section 4 presents some insightful (from both technical and application perspective) properties of the sub-hierarchy generated from the *Food and drink* root. Section 5 describes the supervised curation of the domain which narrows its focus and is the only human input required in the algorithm. Section 6 shows the results of the data-driven enrichment analysis. Section 7 demonstrates how automation can be extended beyond generating the gazetteer and into discovering documents belonging to the domain which can then be used to further refine the gazetteer. Section 8 presents a way of extending an existing gazetteer to other languages. Section 9 presents an annotator evaluation of the resulting gazetteer. Finally Section 10 concludes the paper with comments and discussion of possible future work.

Next, we mention previous work that has 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 ([3], [4], [11]). In [10], a semi-supervised method for enriching existing ontologies with concepts from text is presented. More ambitious approaches propose unsupervised generation of ontologies ([8] [9]), using deep NLP methods. In [5], 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 FD has been approached before. Depending on the purpose of the classification, there exist models for cooking and recipes, models for ingredients and nutrients, food composition databases (EuroFIR classification [2]),

[2] http://www.eurofir.org/

models that classify additives (Codex Alimentarius GSFA [3]), pesticides (Codex Classification of Foods and Feeds [4]), traded food and beverages nomenclature (GS1 standard for Food and Beverages [5]), national-specific classification systems, etc. [12] have proposed a cooking ontology, focused on: food (ingredients), kitchen utensils, recipes, cooking actions. BBC also proposed a lightweight food ontology [6], that classifies mainly recipes, including aspects like ingredients, diets, courses, occasions.

The purpose of the EFD project is to classify food and drink objects from a cultural perspective, which is not addressed by existing models.

2 Europeana Food and Drink

The EFD Classification scheme [2] 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. 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 a large variety of cultural objects related to FD, some of which exist nowadays only in ethnographic museums. These are described in content coming from a variety of CH organizations, 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 more than half of the metadata only available in native languages.

Content is heterogeneous and varied. Examples include [2]: books on Bovine care and feeding (TEL [7]), book on tubers/roots used by New Zealand aboriginals (RLUK [8]), self-portraits involving some food (Slovak National Gallery [9]), traditional recipes for Christmas-related foods (Ontotext), colorful pasta arrangements (Horniman [10]), mortar used to mix lime with tobacco to enhance

[3] http://www.codexalimentarius.org/standards/gsfa/
[4] ftp://ftp.fao.org/codex/meetings/ccpr/ccpr38/pr38CxCl.pdf
[5] http://www.gs1.org/gdsn/gdsn-trade-item-extension-food-and-beverage/2-8
[6] http://www.bbc.co.uk/ontologies/fo
[7] http://www.theeuropeanlibrary.org/tel4/
[8] http://www.rluk.ac.uk/
[9] http://www.sng.sk/en/uvod
[10] http://www.horniman.ac.uk/

its psychogenic compounds (Horniman), food pounder cut from coral and noted for its ergonomic design (Horniman), toy 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 the above-mentioned categories are its direct or indirect subcategories.

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, interwiki links, referring to another Wikipedia or another Wikimedia project e.g. Wiktionary, Wikibooks, external links), *ii*) Informative templates, in particular Infoboxes; *iii*) Tables; *iv*) Categories; *v*) Lists, Portals, Projects.

There are several efforts to extract structured data from Wikpedia. E.g. the Wikipedia Mining software [11] [6] 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 [12], which allows semantic integration of both Europeana and classification data, and easier querying using SPARQL.

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

Table 1: Wikipedia: statistics concerning categories.

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

3.1 Wikipedia categories

Category statistics for Wikipedia are presented in Table 1. The counts are obtained from DBpedia (see [2] sec. 3.11.2). The columns have the following meaning:

- 'Wikipedia' specifies for which language the statistics are computed;
- 'art' is the number of content pages (articles);
- 'cat' is the number of category pages;
- 'art→cat' is the number of assignments of a category as parent of an article;
- 'cat per art' is the average number of category assignments per article, computed as art→cat / art;
- 'art per cat' is the average number of articles assigned per category, computed as art→cat / cat;
- 'cat→cat' is the number of assignments of a category as parent of another category;
- 'cat per cat' is the average number of parent categories per category, computed as cat→cat / cat.

As you can see, there is a great variety of categorization practices across languages. Polish uses a huge number of categories (relative to articles) and assignments. Dutch has a very small number of categories, and their application is not very discriminative ('art per cat' is very high).

Despite these differences, the categorization presents a wealth of information that our method uses for classification.

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. Following Wikipedia, our model is hierarchical and parent-child relations follow SKOS principles [7]. 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 such as *level* (i.e. shortest path from root), number of unique subcategories, etc.
3. Top-down curation: perform manual curation by experts 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 1. Choose a *root* category that defines the domain.

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 (sometimes due to bad practices 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* and 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 drastically reduces the size of the sub-hierarchy with minimal work, thus being an efficient early method for pruning.

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 collection of documents, thesauri, databases, etc. relevant to the domain, we use a general tagging algorithm to map concepts from the collection to the

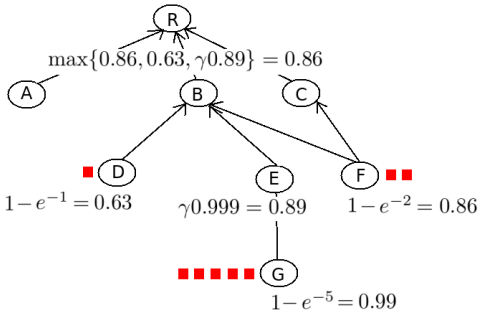


Fig. 1: Scoring categories bottom-up. Concepts mapped to categories are marked with red squares. Categories are marked with circles and named with capital letters.

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 $887,523$ categories or about 80% of all categories in the English Wikipedia (see Table 1). The categories span 26 levels below 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.

Most subcategories reachable from the selected root are *not relevant* to the domain. E.g. all the top 10 most populous categories at level 5 are irrelevant:

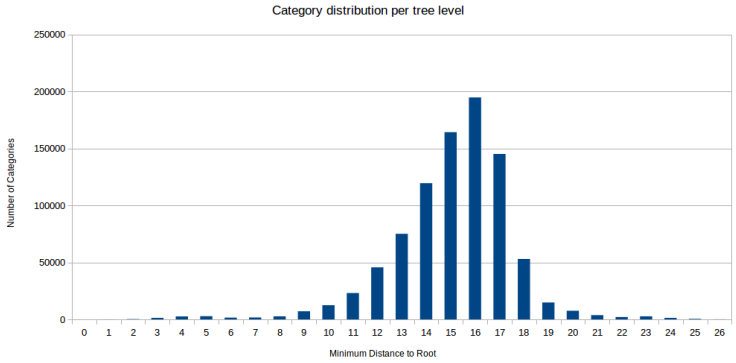


Fig. 2: Distribution of the shortest-path length from categories to the FD root category.

Oceanography, Water pollution, Physical exercise, Bodies of water, Natural materials, Country planning in the UK, etc. We discuss below reasons and examples for such a disappointing initial hierarchy.

4.1 Reasons for Irrelevant Inclusions

Semantic Drift. The main reason for irrelevance is "semantic drift": since the meaning of the Wikipedia "parent category" relation is not well-defined, the longer path one follows, the harder it becomes to see any logical connection between the two categories (ancestor and descendant). E.g. following the chain
Food and drink → *Food politics* → *Water and politics* →
Water and the environment → *Water management*,
one quickly reaches into rivers, lakes and reservoirs. Luckily it is easy to cut off major irrelevant branches early in the hierarchy.

Wrong Hierarchy. We were surprised to reach football teams. This happens along this 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: *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, similar problems still exist elsewhere.

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 some mixed subcategories that may include both relevant and irrelevant children: *Animal glandular products*: milk and its thousands of subcategories is relevant, castoreum is not; *Insect products*: honey is relevant, silk is not; *Mollusc products*: clams and oysters are relevant; pearls are not.

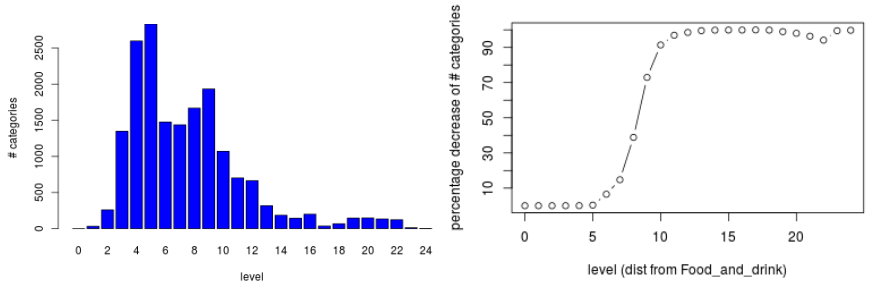
Non-human Food or Eating. Food and drink explicitly includes animal feeding, thus not all are foods for humans, e.g. *Animal feed*. The subcategory *Eating behaviors* has some appropriate children, e.g. *Diets*, *Eating disorders*, but has also some inappropriate children, e.g. *Carnivory*, *Detritivores*.

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 judgment. 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 effort from a human expert. At this step, we consider that a consensus among several experts is not needed, because only clean mismatches were removed. Examples of removed categories:
Natural materials, *Natural resources*, *Water treatment*, *Education*, *Academia*, *Academic disciplines*, *Subfields by academic discipline*, *Scientific disciplines*, *Real estate*, *Civil engineering*, *Construction*, *Water pollution*, *Property*, *Land law*, *Intelligence*, etc.
Some of these categories seem simply irrelevant, like *Civil engineering*, others could potentially lead to articles relevant to FD, like *Natural resources*. The path from FD to *Natural resources* goes through *Agriculture*, *Agroecology*, *Sustainable gardening Natural materials* (length 5). However, the category is too broad and too distant to matter, and whatever relevant articles it would link to, should be retrievable by alternative, shorter paths from the FD root. For example, *Natural resources* leads to Salt via *Minerals* and *Sodium minerals*. However, there is a shortcut from FD directly to Salt via *Foods*, *Condiments*, *Edible salt*, so there is no need to pass by *Natural resources*, which in turn adds to the hierarchy many irrelevant subcategories.

The new cardinalities per level are shown in Figure 3 a). Figure 3 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%.

The refinement of the FD hierarchy was 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



(a) (b)

Fig. 3: a)Number of categories per level after expert curation. b)Decrease of number of categories per level after expert curation.

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.

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 Objects Thesaurus, consisting of about 1500 concepts used for describing Horniman museum artefacts (700 are currently used in objects).

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 an Ontotext general-purpose concept extractor[13] that identifies Wikipedia concepts in general text. For the purpose, we concatenated all thesaurus terms 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,

[13] Customized version of http://tag.ontotext.com/

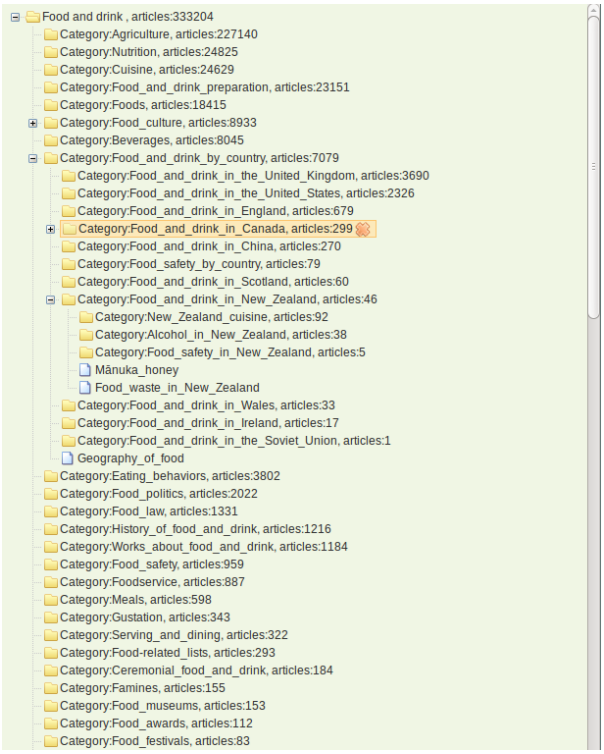


Fig. 4: Visualization interface for the FD categorization.

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

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...’

Evaluation. 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).

Table 2: The highest scoring categories w.r.t. the proposed scoring scheme.

Category	Score	Category	Score
Cooking utensils	1.00	*Crops*	0.99
Teaware	0.99	*Spices*	0.98
Serving and dining	0.99	*Agricultural machinery*	0.98
Cooking appliances	0.99	*Commercial fish*	0.98
Drinkware	0.99	*Eating utensils*	0.98
Staple foods	0.99	*Food storage containers*	0.98
Tropical agriculture	0.99	*Serving utensils*	0.98
Gardening tools	0.99	*Animal trapping*	0.98
Fishing equipment	0.99	*Food and drink*	0.95
Cooking techniques	0.99	*Recreational fishing*	0.95
Cookware and bakeware	0.99	*Breads*	0.95
Crockery	0.99	*Hunting*	0.95
Kitchenware	0.99	*Dairy products*	0.95
Spoons	0.99	*Food ingredients*	0.95
Fishing techniques and methods	0.99	*Food preparation appliances*	0.95

6.2 Scoring FD Categories w.r.t. Mapped Horniman Concepts

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 shown in table 2.

Qualitative evaluation of the scoring system: 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.

Category scoring is also useful for ranking results of a semantic search, provided that enough relevant data is collected and mapped onto the hierarchy. If a user queries a concept, the tool can return a list of Wikipedia categories relevant to the concept, ranked by relevance to the FD domain. For example, if a user searches for 'fork', the category 'Gardening tools' 0.998 will appear higher in the results than *Eating utensils* 0.982, because more concepts from the Horniman museum are mapped to *Gardening tools*.

6.3 Other CHO collections: Alinari, TopFoto and Wolverhampton

In addition to Hornimann, we performed enrichment on three more data providers' Cultural Heritage Objects. These providers were *Fratelli Alinari*, *TopFoto Partners LLP* and *Wolverhampton Arts and Museums*. The former two provided primarily visual archives with some description of the photographs' contents

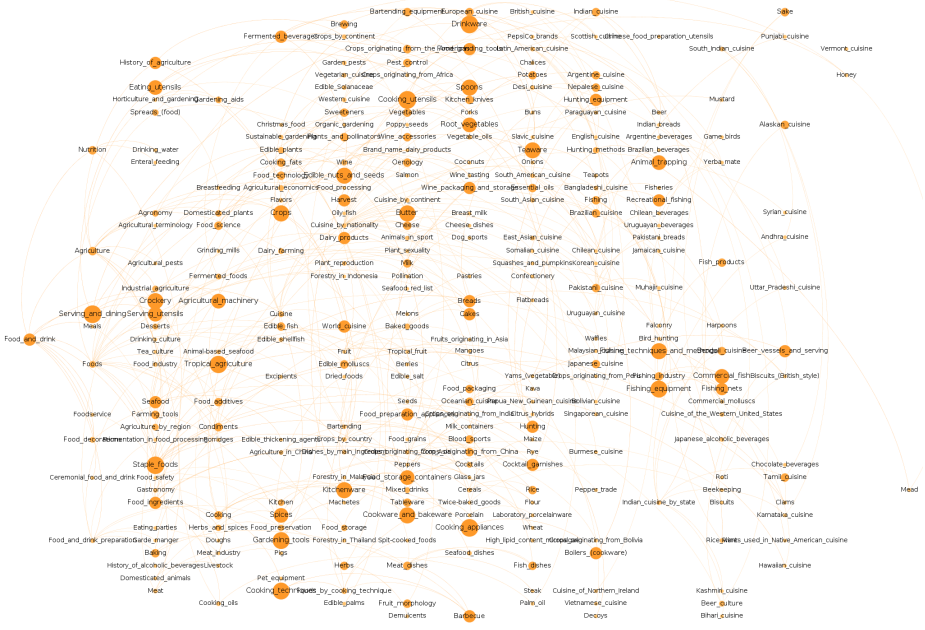


Fig. 6: Paths to *Food and drink*, activated by the bottom-up scoring scheme. Larger image available at ftp://ftp.ontotext.com/pub/EFD/Horniman3.png

while the latter provided descriptions of physical objects which often but not always came with a picture of the object.

Table 3: Enrichment comparison between providers

Provider	**CHOs**	**% of Total**	**Enriched CHOs**	**Coverage**	**Enrichments**	**per CHO**
Horniman	4351	61.3%	4351	100%	40096	9.2
Alinari	498	7.0%	187	37.6%	540	1.1
TopFoto	1814	25.5%	1780	98.1%	14052	7.7
Wolverhampton	438	6.2%	405	92.5%	4161	9.5
All	7101	100%	6723	94.7%	58849	8.3

In Table 3 we can see a summary of the results we obtained from the four different data providers. In the table the first column lists the institution's name; the second is the total number of Cultural Heritage Objects provided; the third shows what fraction of the total CHOs that represents; the fourth shows the number of CHOs that have been enriched with at least one Food and Drink concept; the fifth shows what percentage of the CHOs have been enriched; the

sixth shows the total number of concepts discovered in CHOs; the final column shows how many concepts have been discovered on average per CHO.

Unsurprisingly, our Food and Drink gazetteer performs exceptionally well on the Horniman data as we can see every object in the collection has been tagged with at least one concept and on average we discovered over nine objects per CHO. The good news is that the gazetteer has generalized very well and we can observe over 90% coverage for both *TopFoto* and *Wolverhampton* as well as nearly 95% coverage on the full data set.

The data provider we observe the worst results on is Alinari. Further exploration reveals that the reason for that is the quality of the metadata. Namely, the description field of many of their objects (which should be reserved for a description of the photo's contents and should be the most likely source of Food and Drink concepts) very often simply consists of an identical administrative message with no relation to the specific photograph or Food and Drink in general.

The conclusion is that the Food and Drink gazetteer we have produced is very well suited to our purposes so far. We intend to enrich more data provider collections in the future for further corroboration but this is solid evidence for the method's effectiveness.

7 A food and drink statistical classifier

The Food and drink hierarchy naturally provides with data for training a statistical classifier that is able to discriminate between food-and-drink and non-food-and-drink documents. Specifically, we constructed two sets of documents, *positives* and *negatives*, as follows. The *positives* are the abstracts of all Wikipedia concepts that are in the FD hierarchy, that were at least once tagged in the collection of CHOs (6.1). Thus, we only include confirmed food-and-drink-related concepts. For the *negatives*, we collected by random selection abstracts of Wikipedia concepts that are not in the FD tree. We made sure that the selection is not biased towards heavy categories (e.g. *Living people*), which would make the classification unrealistically easy (living people against food and drink concepts would be easy), by limiting the number of samples from the same category to 3. We arrived at a set of about 1200 positive samples and 4700 negatives. We trained a bag-of-words maxent classifier and achieved 89% F1 score, for a train-test split of 80% and 20% of the data, respectively. The dataset being relatively small, we consider that the performance is really good. Also, as the collections of CHOs get populated and more positive examples are collected, the performance will most likely increase.

The most relevant 50 features (stemmed) for classification are:

food, restaur, drink, cook, dish, london, fruit, brand, type, commonli, cuisin, beverag, alcohol, tea, meat, chef, sugar, plant, edibl, term, shape, bread, process, fish, wine, coffe, kitchen, flour, tradition, market, meal, crop asia, largest, varieti, metal, agricultur, consum, typic, heat, anim, tradit, prepar, flavor, grain, vari, popular, kind, britain, bottl

Table 4: Categories with many articles predicted as non FD (third column), and some articles predicted as FD (second column).

Category	FD	non FD
Olympic medalists in equestrian	57	344
New York Red Bulls players	64	259
British Darts Organisation players	51	250
Professional Darts Corporation players	43	242
American jockeys	35	236
FC Red Bull Salzburg players	34	190
Genes mutated in mice	33	183
Deaths from typhoid fever	25	141
Electro-Motive Diesel locomotives	27	139
Deaths from cholera	34	131

Clearly, the classifier succeeds to extract the most relevant words used to described FD-related concepts.

With the goal in mind of further improving the tree by removing categories not FD-relevant, we applied the classifier to the rest of the articles from the FD tree, that have not been associated with any CHO from our collection, counting above 90000. We call those *maybes*. The classifier categorizes each *maybe* article as either FD or not FD, which allows for each category in the tree to be evaluated by two counts: the number of articles predicted as FD and the number of articles predicted as not FD.

Since we cannot afford to inspect all categories (above 10000), of interest are two particular types, candidate for removal from the FD hierarchy: fist, those that have no article classified as FD, as many of them could be not FD-relevant; second, large categories with many articles that are classified as not FD (but can have FD articles also), as removing them would decrease the size of the tree by much. For the first type, we manually inspected top 300 categories with 0 FD and more than 8 not-FD articles. Out of them, 34% we judged as not FD-relevant and are candidates for removal. They contain up to 1100 articles. The second type are heavy categories, top 200 sorted by the number of not-FD-classified articles. Out of them, 25% we judged to be candidates for removal, comprising about 9600 articles.

Examples of categories marked for removal (we show these because we consider that it is interesting that they are reachable from the Food and Drink category) are shown in Table 4.

Equestrianism and related topics are reachable from Food and drink via *Agricultural occupations*, *Animal care occupations* and *Horse-related professions and professionals*. It is a large category and mostly irrelevant to the culture of food and drink.

FC Red Bull Salzburg players and *New York Red Bulls players* are linked to Food and drink by their immediate relation to the energy drink, which in turn is related to many sports events.

Various darts topics (*Professional Darts Corporation players, British Darts Organisation players*) are related to food and drink because darts are typical restaurant and pub games.

The category *Genes mutated in mice* is subcategory to *Mutated genes*, which leads to *Agriculture* via *Genetically modified organisms* and *Crops protection*. Clearly the subject turns quickly from agriculture to biology (genetics) and the relation to food and drink culture is lost.

Cholera or typhoid fever are both foodborne illnesses, hence the relation to food and drink. The specific categories *Deaths from cholera* and *Deaths from typhoid fever* are distancing away from the food and drink culture, rather belonging to medicine topics.

The divergence from food and drink to *Electro-Motive Diesel locomotives* occurs around *Agricultural machinery* and *Tractors*.

The lesson we learn is that via the category-category relations topics diverge and change subtly, such that already starting at a distance of 5-6 categories to the root, the relevance to the domain may weaken. In many cases however, the categories remain related to food and drink and get more specific. Discriminating between food-and-drink relevant and irrelevant categories at deep levels down the tree is hard, due to their large number. Hence, the classifier is a valuable tool for ranking and prioritizing the candidates for removal.

Classification of CHOs. The classifier can be used for discovering Europeana CHOs that are FD-related. For most of the CHOs that have a reasonably long description (at least a couple of sentences), we can assume that the classifier can estimate if they are related to FD with high confidence. This constitutes an automated filter for selection of objects and enriching the current collection of CHOs. Further on, via our annotation pipeline, the CHOs will be enriched with FD concepts and the evidence supporting categories in the FD tree will become more substantial. To conclude the cycle, the classifier will be updated based on the new evidence and become better. Hopefully some convergence will be reached after many such iterations, meaning that the classifier and the labels for the categories in our hierarchy will stabilize. We leave this for future work.

8 Food and drink modeling for other languages - outlook

Modeling food and drink for many other European languages is future work for Europeana Food and Drink. In this section we present preliminary analyses and approaches.

The most natural approach to modeling FD in other languages is to repeat the procedure for an arbitrary Wikipedia in language X. The *sameAs* relations between the categories in various Wikipedias allows us to use the already existing FD hierarchy in English as a reference, both for curation and for validation of new

FD trees. Ideally, there will be a one-to-one match between the trees. In practice, this is far from the truth; the category structure is different for each language. Moreover, other Wikipedias are significantly poorer (see Table 1), which means that the best outcome is a FD hierarchy in the new language which is a subtree of the English FD hierarchy (via *sameAs* relations). Our experiments show that this assumption is also not true. The following subsection shows that there is little overlap between the English reference and the FD in other languages (at least, for the three languages we tested).

8.1 Top-down category harvesting from FD root

For few European languages we extracted the tree of FD categories reachable from the root equivalent to Food and drink in those languages. In German, the root category is 'Essen und Trinken', in Bulgarian the root is 'Hrana i napitki', in Italian the root is 'Categoria:Alimentazione'. We will call these the raw FD hierarchies, for their respective languages.

In the German Wikipedia, from the FD root we reach 1252 categories, on 8 levels (see Table 5). Similarly, in the Italian Wikipedia, from the root we reach 1352 categories, on 7 levels. By comparison, in the Bulgarian Wikipedia 17849 categories are reachable from the FD root, on 24 levels. By comparison, the Bulgarian FD tree is more than 10 times larger in terms of number of categories. We do not have an a priori estimate of the 'right' size of the FD tree in a specific language, because it is influenced by the overall size of the particular Wikipedia, by the conservativeness of SKOS relations between categories, by the richness of the FD culture of the native speakers of the language, etc. However, the difference between the German and Italian on one side, and the Bulgarian raw FD hierarchies on the other side is interesting and deserves investigation.

We report some statistics about the *sameAs* relations between FD categories in different languages. We consider three types of relations, depicted in Figure 7: a) two categories that share a *sameAs* relation, in English and language X, are in the FD hierarchies of their respective languages; in the Figure they are shown in green; b) categories in one FD hierarchy, the *sameAs* of which is not in the FD hierarchy of the other language; in the figure, they are shown in red, one example for each tree; c) categories that are in one of the FD trees, but they do not have *sameAs* correspondent in the other language; in the Figure, they are shown in gray.

Figure 8 shows the distribution of categories for each level of the FD tree, for English (the curated tree), German, Bulgarian and Italian (raw hierarchies).

Figure 9 shows the distributions of the *sameAs* properties in the German, Bulgarian and Italian trees, in comparison to English. Figure 9a) shows by level, the number of Bulgarian FD categories that have *sameAs* match in the English FD tree, in green; the number of Bulgarian FD categories that have a *sameAs* pair outside the English FD hierarchy in red; and in gray, the number of Bulgarian FD categories that do not have *sameAs* pair in English. the overall picture is that the vast majority of the raw FD Bulgarian categories have a *sameAs* pair in English, outside the English FD tree. On manual inspection, we concluded that

Table 5: *sameAs* statistics for various languages. X is the language, second column shows categories that are

Language	FD in X and EN	FD only in EN	FD only in X	FD cat in X
X = de	285	603	56	1252
X = bg	239	103	11616	17849
X = it	434	331	45	1352

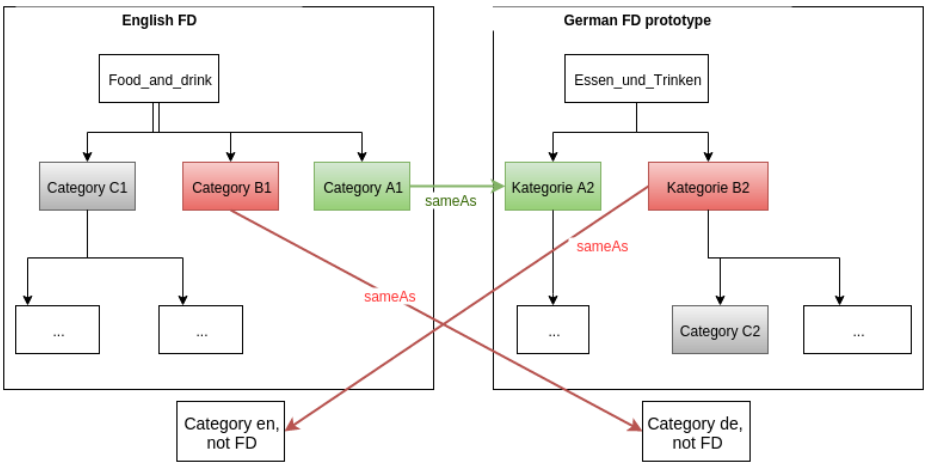


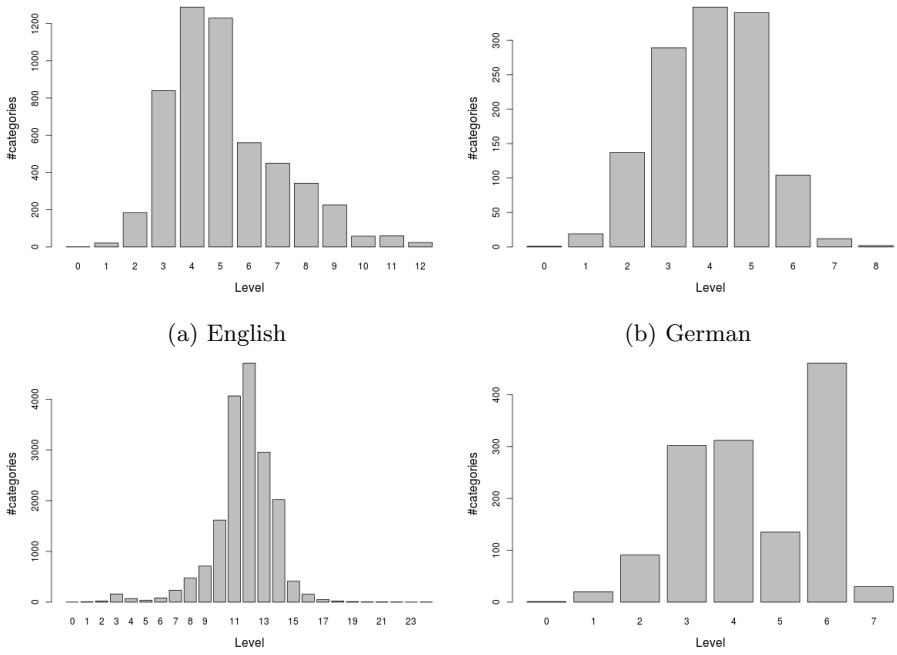
Fig. 7: *sameAs* relations between two FD trees, eg. in English and German. Categories in green are related by *sameAs* and are both in the FD trees. FD categories with *sameAs* correspondents outside the other FD tree are shown in red. Categories without *sameAs* match are depicted in gray.

many of those are related to water, hydrology, oceanography, etc., topics that we removed from the English FD hierarchy early during the project. Conversely, a large percentage of the English FD tree does not have *sameAs* pairs in Bulgarian (94%), see 9b), gray area.

The statistics for the German raw FD hierarchy are quite different from the Bulgarian. More than 16% of the German raw FD categories have *sameAs* pairs in the English FD tree (Figure 9c). Very few German raw FD categories have a *sameAs* pair outside the English tree (less than 4%). However, only 4% of the English FD categories are in the German raw FD (Figure 9d), which is less than for Bulgarian.

Our analysis shows that the Italian FD hierarchy has the highest coherence with the English FD. 32% of the Italian raw FD categories have a *sameAs* pair in the English FD, 3% have a *sameAs* pair outside the English FD and 65% do not have a *sameAs* pair in English (Figure 9e). 8% of the English FD have a

sameAs pair in the Italian raw FD, 6% have a *sameAs* pair outside the Italian FD and 86% of the English FD do not have a *sameAs* in Italian.



(a) English (b) German

(c) Bulgarian (d) Italian

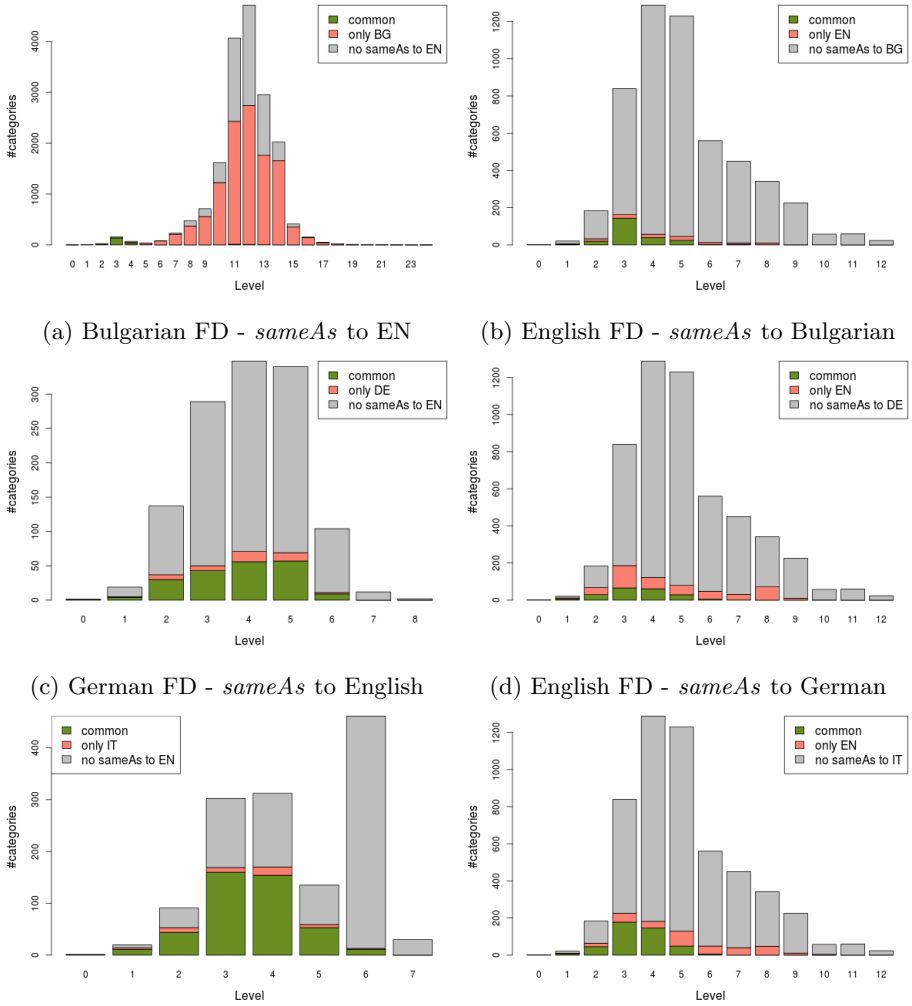
Fig. 8: Overall statistics by level of the FD tree in various languages.

8.2 Article-driven reconstruction approach

The previous Subsection showed that the *sameAs* correspondence between the FD hierarchies in English and some target language are probably not sufficient for an automatic construction of the FD in the respective language.

An alternative approach is a bottom-up reconstruction of a FD tree in language X, by following the steps:

1. start from all FD articles from the English FD, call that set A_{en}
2. getting their *sameAs* pairs in language X, call that set A_X
3. with the Wikipedia in language X represented as a graph, for which nodes are categories, leaves are articles and relations are category-category SKOS relations or article-category relations, infer (using a Steiner tree [13] for example) a minimal connected tree that contains all A_X leaves. Call this FD_X



(a) Bulgarian FD - *sameAs* to EN

(b) English FD - *sameAs* to Bulgarian

(c) German FD - *sameAs* to English

(d) English FD - *sameAs* to German

(e) Italian FD - *sameAs* to English

(f) English FD - *sameAs* to Italian

Fig. 9: Overall statistics of category *sameAs* by level of the FD tree in various languages.

4. add to FD_X all other articles contained in the categories of FD_X

The implementation of step 3. is not trivial, the Steiner tree problem being NP-complete in most of the cases. Approximations are available, for large graphs [13].

The procedure described above would ensure maximal coverage for the English FD articles and is fully automated. However, there may be FD articles in

language X that are specific to some culture and therefore not have a *sameAs* pair in English. Those would have to be discovered and added separately.

We suggest that a hybrid approach between the top-down category harvesting and bottom-up, article-driven reconstruction is optimal.

9 Evaluation

Ideally, we would compare the effectiveness of the presented approach to an already existing algorithm but there is no published work dealing with this particular task. To further complicate the situation, the iterative nature of the algorithm and intentionally nebulous definition of the domain preclude the use of many typical evaluation approaches.

9.1 Method

In the end the domain generated by the algorithm was refined sufficiently to function well with the available data but for the purposes of evaluation we settled on the alternative approach of using annotators to more objectively evaluate the quality of the resulting gazetteer.

The finalized Food and Drink domain consists of 14,368 categories and 185,020 articles. A 100 of each were randomly selected and were presented to three independent annotators that judged whether each individual article or category should be part of the Food and Drink domain. The inter-annotator agreement was 97% for categories and 98% for articles. In the cases of annotator disagreement, the choice of the two agreeing annotators was taken as correct.

9.2 Results

Level	**Count**	**Correct**
All	100	72
1	0	0
2	3	3
3	13	13
4	24	24
5	21	18
6	17	10
7	10	3
8+	12	1

Table 6: Category sample results

Level	**Count**	**Correct**
All	100	56
1	0	0
2	4	4
3	12	12
4	21	19
5	13	9
6	14	9
7	4	1
8+	32	3

Table 7: Article sample results

The results of the evaluation are summarized in two tables. Table 6 shows the evaluation of the category sample. It has three columns with *Level* denoting the

shortest distance from a given category to the root, *Count* showing how many of the sample categories were in that sample and *Correct* showing how many of those were judged by the annotators to belong to the domain. It shows an overall accuracy of 72% but more interestingly it very clearly demonstrates the effect of *semantic drift* mentioned in 4.1. Categories within five steps of the source are very accurate then meaning is quickly lost in the next few steps until categories 8 or more steps away appear to be almost entirely irrelevant to the domain.

Table 7 shows the evaluation of the articles sample. The columns have the same meaning except *Level* simply chooses an article's parent category closest to the source. Once again the *semantic drift* effect is evident but the overall accuracy is only 56%. This can be attributed to the fact that the article to category relation effectively introduces an extra step in the semantic drift.

The evaluation made it clear that the main factor in a category's accuracy is the distance from the root. This means that the top-down expert pruning discussed in Section 5, which minimizes *semantic drift* and indirectly distance from the root, really is an effective tool for refining the hierarchy. However, the evaluation also demonstrated that while we had pruned the hierarchy sufficiently for working with our data, there is still room for improvement. That can be done by either performing further pruning steps or, as suggested by the data, simply putting a maximum limit on the distance from the root.

10 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 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.

We evaluated the scoring schemes qualitatively, by showing that the categories that are 'activated' with large scores are those that describe the main topics of the Horniman thesaurus terms, namely agriculture, food serving, fishing and hunting, etc. These topics were not explicitly input to our framework, only the concrete terms like spoon, bread, cup, fishing hook, etc. A quantitative evaluation is future work, after the semantic search for FD concepts is open to

the public. Then, we plan to submit various scoring schemes with various decay parameters and compare them based on user feedback.

Despite the reasonable coverage of the domain, we identified concepts – or sets of concepts – that belong to FD, but are not found under the FD root. For example, some hunting weapons are not accessible directly from the FD root. Horniman items representing spears could not be tagged, and they should, being tools for obtaining food. We have added a number of Wikipedia parent categorizations to enlarge the FD hierarchy, eg placing "Hunting" under FD, "Livestock" under "Agriculture" (which is under FD), etc. We also split some articles and added categorizations and labels (redirects) to match specific objects that we encountered. For example:

- Created pages "Shepherd's crook" and "Tumbler (glass)" by splitting text from existing pages. Added label "Crook"
- added to "Leash" the note "Leashes are often used to tether domesticated animals left to graze alone"as justification for adding category "Livestock"

We may add "private" secondary roots to the categorization: a direct, custom connection of type `broader` to the *Food and Drink* root is a possible way to add secondary roots.

A big challenge for the EFD project is building a multilingual categorization for up to 11 languages. Our prototype is currently limited to English, but we presented some preliminary analysis and two approaches towards automatic or semi-automatic construction of the FD hierarchy in an arbitrary language. We suggested a top-down approach that builds the FD from a root category and a bottom-up approach that reconstructs relations between articles, taking advantage of the 'parallel' Wikipedias for other languages via the *sameAs* relations. We noticed that the inter-language links are not very rich for the three languages we tested: German, Bulgarian and Italian, a large percentage (above 85%) pf the English FD hierarchies having no pair in the other languages.

Acknowledgements The research presented in this paper was carried out as part of the Europeana Food and Drink project, 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. Alexiev, V. Europeana Food and Drink Classification Scheme, Europeana Food and Drink project, Deliverable D2.2, 2015. [14]

[14] http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Classification-Scheme-(D2.2).pdf

3. Cheng, CP., Lau, GT., Pan, J, Law, KH., Jones, A., Domain-Specific Ontology Mapping by Corpus-Based Semantic Similarity. 2008 NSF CMMI Engineering Research and Innovation Conference
4. Fridman Noy, N., Musen, MA., An Algorithm for Merging and Aligning Ontologies: Automation and Tool Support, Workshop on Ontology Management at the 16th National Conference on Artificial Intelligence (AAAI-99), 1999.
5. 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.
6. Medelyan, O., Milne, D., Legg, C., Witten, IH., Mining Meaning from Wikipedia, Int. J. Hum.-Comput. Stud., vol. 67(9), 2009, pp 716–754.
7. Miles, A., Bechhofer, S., SKOS Simple Knowledge Organization System Reference. W3C Recommendation. 18 August 2009.
8. Mousavi, H., Kerr, D., Iseli, M., Zaniolo, C., Harvesting Domain Specific Ontologies from Text, ICSC '14, 211–218, 2014.
9. 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.
10. 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.
11. Pinto, HS., Martins, JP., A Methodology for Ontology Integration, Proceedings of the 1st International Conference on Knowledge Capture, K-CAP '01, 2001.
12. Ribeiro, R., Batista, F., Pardal, JP., Mamede, NJ., Pinto, HS., Cooking an Ontology, Artificial Intelligence: Methodology, Systems, and Applications, Lecture Notes in Computer Science 4183, 2006, pp 213–221.
13. Gubichev, A. and Neumann, T., Fast Approximation of Steiner Trees in Large Graphs, Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012, pp 1497–1501.