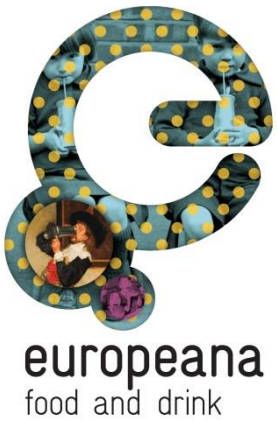


Grant Agreement 621023

Europeana Food and Drink

D2.2 Classification Scheme

Deliverable number	D2.2
Dissemination level	PU
Delivery date	February 2015
Status	Final
Author(s)	Vladimir Alexiev (ONTO)

This project is funded by the European Commission under the
ICT Policy Support Programme part of the
Competitiveness and Innovation Framework Programme.

Revision History

Revision	Date	Author	Organisation	Description
V0.1	15/01/2015	Vladimir Alexiev	ONTO	Initial Draft
V0.2	26/01/2015	Elena Lagoudi	PS	First review
V0.3	27/02/2015	Diantha Osseweijer	CAG	Second review
V0.4	20/02/2015	Vladimir Alexiev	ONTO	Final Version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Contents

1 Introduction

The Europeana Food and Drink Classification scheme (EFD classification) is a multi-dimensional scheme for discovering and classifying Cultural Heritage Objects (CHO) related to Food and Drink (FD). To support the broadest possible range of re-use models, we are building upon existing terminologies to develop and apply a multilingual taxonomy for FD collections which will be used to tag, discover and aggregate relevant material by theme.

The project will also use innovative semantic technologies to automate the extraction of terms and co-references. The result will be a body of semantically-enriched metadata that can support a wider range of multi-lingual applications such as search, discovery and browse.

1.1 Background

The core concept of the EFD Best Practice Network is to kick-start the creative and commercial re-use of digital content relating to food and drink from the culture sector to drive new commercial applications, relationships and partnerships.

FD serves the dual purpose of providing a powerful thematic focus to inspire creative re-use of digital cultural content while offering sufficient breadth to support a wide range of applications and approaches.

EFD WP2 will identify, describe, enhance, license and upload a body of high-quality digital assets and their associated metadata, to support the delivery of commercial applications and public engagement activity.

These objectives will be facilitated by the work under Task 2.2:

- Devise, assemble and develop an EFD classification scheme to support classification and resource discovery
- Develop approaches to enrich semi-automatically the assets and their associated metadata using this classification scheme

These objectives will be facilitated by the work under Task 2.3:

- Map local metadata structures to the requirements of the Europeana Data Model
- Ingest the primary digital assets into the Europeana Cloud infrastructure
- Ingest the descriptive metadata and identifiers into Europeana
- Apply the relevant licensing conditions drawn from the Europeana Content Re-use Framework

These objectives will be facilitated by the work under Task 2.4:

- Ingest new content contributions to Europeana Cloud infrastructure
- Ingest user-generated content alongside existing EFD assets
- Apply appropriate licensing and metadata schemata to new content

2.1 Europeana Food and Drink Classification Scheme

1.2 Overview of Content

The initial content that the consortium content partners had committed to, was described in a table included in the DoW and analysed in detail in D2.1, Inventory of EFD Content.

Here are some key findings about the EFD content:

- The content comes from a variety of Cultural Heritage organizations, ranging from Ministries to academic libraries and specialist museums to picture libraries. See sec. 1.9 for specific examples.
- The content represents a significant number of European nations and cultures
- The content comprises of objects illustrating FD heritage, recipes, artworks, photographs, some audio and video content and advertising relating to FD
- The content is heterogeneous in types and significance, but with the common thread of FD heritage and its cultural and social meaning
- The content is mainly documented in Dublin Core (also see sec. 3.1)
- The content metadata are available partly in English and native languages, with almost half of metadata only available in native languages

The content will be ingested into Europeana via MINT (NTUA) and was to be hosted in Europeana Cloud. Europeana Cloud as a project is only initializing now (February 2015), so the initial contact to proceed this only happened recently. This obviously solved the issue of the “homeless” Europeana Food and Drink content, as it has now been decided that the content will be ingested into Europeana Cloud, while its metadata will be ingested in Europeana via MINT. EFD content will be a pilot body of content to test the feasibility of Europeana Cloud, which is a pure infrastructural solution allowing to develop interfaces using the API it exposes. WP2 team is liaising with Pavel Kats, from Europeana Cloud, to outline the work flow..

As demonstrated by various projects within the Europeana ecosystem, the key issue with creative and commercial re-use seems to be findability and exploitability of assets.

Users of digital cultural content seek and retrieve it from aggregators and repositories where they can browse, search and find the desired content amongst a mosaic of diverse content publishers. The interfaces for content discovery often offer poor user experience and lack the functionality of intuitively presenting diverse cultural material, as metadata expressivity and overall quality varies.

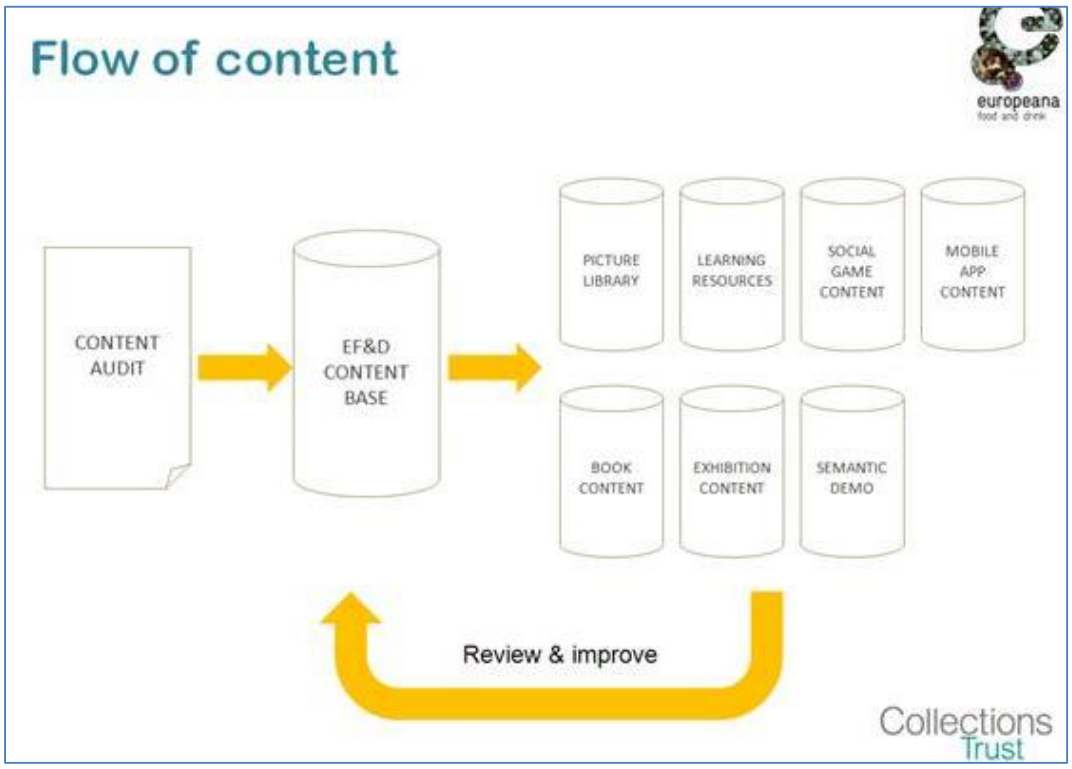
There is a need for more detailed, domain-specific curation, presentation and publication of digital heritage content.

1.3 Previous Planning

The content (digital assets) was to be ingested into the Europeana Cloud infrastructure and the metadata into Europeana. However, the Europeana Cloud is not yet ready, so for the past few months the Consortium has investigated various

solutions for hosting this content and researched their functionalities and specifications.

It was agreed that, following the content audit that produced D2.1, the content's inventory, the content base was to be built from which all the 7 products would draw content from, as described in the schematic below.



The Consortium agreed that, for the project's content to be findable and exploitable, a more commercial focused application has to be used. It was decided, since the Content Base needs to provide the underlying infrastructure from which the other products will draw their content from, it has to be merged with the Picture Library, so that the feasibility of this entrepreneurial model that the Consortium is developing for GLAMS can be piloted.

The requirements and functionalities of this Content Base/Picture Library are summarized as follows:

- Storing high-quality images & media
- Storing associated metadata
- Managing rights associated with content & metadata
- Easy workflows for ingestion, mapping, data management and export
- Export to LIDO/EDM compatible format
- Uses existing technology/infrastructure
- Scalable to meet future demand

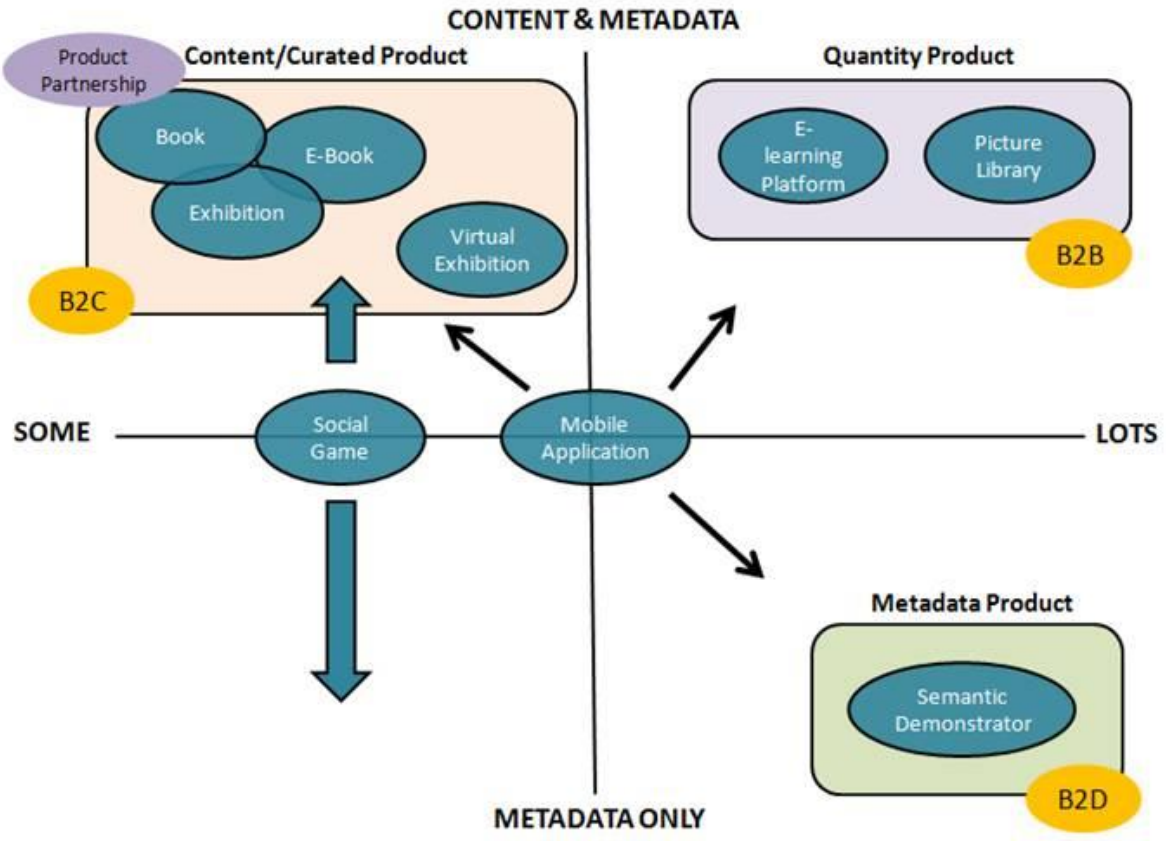
1.4 Role of the Deliverable in the Project

WP2 as a whole will support the objective of identifying, describing, enhancing, licensing and uploading a body of high-quality digital assets and their associated metadata, to support the delivery of commercial applications and public engagement activity.

The Content Base developed throughout the project will be available for cultural institutions, creative industries, professional users and third party developers in order to easily search for the cultural resources that meet their retrieval criteria so as to use and re-use them for the development of applications, products and services.

The result will be a body of high quality digital assets and semantically-enriched metadata that can support a wider range of (multi-lingual) natural language applications such as search, discovery and browsing.

The Europeana Food and Drink Content Base will feed content into the applications and products that the Consortium will develop. Based on the quantity of content and metadata needed, the applications were split into the following diagram:



1.5 EFD Classification Purposes

Expanding on the previous section, the EFD classification is intended to serve a variety of purposes:

- Classification of content provided by the EFD project partners
- Enable semi-automatic enrichment of content assets and their associated metadata
- Enable discovery of content already existing in Europeana that is also related the topic of interest (see sec.5.9)
- Form the basis for a future Food and Drink Channel on Europeana, a goal that the EC has posited
- Provide the core semantic information underpinning the Semantic Demonstrator application
- Enable semantic search and faceting

The EFD Classification scheme, will be an invaluable tool for search, browsing and discovery of relevant FD heritage content. While it is a classification (a way to select from a set of values), it will also inform structural decisions, leading towards the development of a FD heritage profile for EDM.

1.6 Semantic Search

To explain the point about semantic search and faceting: consider the examples in sec.3.15. Each of the captured concepts has additional useful information:

- It is part of a meaningful hierarchy, e.g.:
  - Chicken is Meat
  - Pestle is <grinding and milling equipment>[1], together with grinders, mortars, grindstones, manos, and (according to Getty) sausage stuffers
  - Austral Islands is in French Polynesia
  - Christmas Bread is a kind of Christmas food
  - Bread Loaf is a type of bread
  - Bread Loaf is associated with Bulgarian Cuisine
- A lot of these have additional information, e.g.
  - Bulgarian Cuisine is made/used in Bulgaria
  - Austral Islands has coordinates S 23° 0' 0", W 150° 0' 0"
  - Bread is mostly made of grains

This enables powerful searches by concept, higher-order concept, geospatial (within rectangle or near a place), etc. It also enables faceting by object type, food or ingredient type, location etc.

1.7 Query Expansion

Europeana offers a simple query translation service. E.g. a query for "beer" in 6 languages[2] returns a translatedQuery that can be used for multilingual search:

- beer OR "Beer" OR "Öl" OR "Cerveza" OR "Birra" OR "Bier" OR "Bière"

[1] http://www.getty.edu/vow/AATHierarchy?find=&logic=AND&note=&subjectid=300024716
[2] http://europeana.eu/api/v2/translateQuery.json?languageCodes=en,de,fr,it,es,sv&wskey=api2demo&term=beer

The technically more complex semantic enrichment and semantic search is needed because a mere word translation:

- does not accommodate hyponyms, e.g. Lager or Pilsner
- does not prevent ambiguities. E.g. Recipe can be related to food or to Medicine; fork could mean cutlery, an agricultural instrument (pitch fork) or a musical device (tuning fork)

[Olensky 2012] describes a classical failure of naïve query translation in Europeana (using the older enrichment tool AnnoCultor): If you search for "poison" in the collections provided by Swiss institutions, you may find photographs from India and Indian movie covers. The reason is that objects were enriched with the term "poison" and its multilingual equivalents. In Latvian "poison" means Inde, which is the same keyword the French-speaking domain expert gave to the objects to describe their content: India.

Multilingual expansion that does not disambiguate will decrease precision so much that any gains from increased recall will not be perceived as positive.

One requirement for a well-constructed taxonomy is to enable Query Expansion, i.e. finding objects associated with a lower-level concept (e.g. Lager) when the user searches for a higher level concept (e.g. Beer). For this to work well:

- The association should be to a semantic item, as free of ambiguities as possible
- The thesaurus should distinguish between different kinds of broader relations (e.g. generic, partitive and instantial) and have proper composability between them (see [Alexiev 2014a]).

The Getty thesauri (sec.3.2) are one of the few that support proper composability and expansion. Wikipedia Categories (sec.□) are on the opposite side of the spectrum, since they offer no guaranteed of semantic consistency whatsoever.

1.8 Complexity of Classification

Many Europeana-related projects have developed classification schemes to cover their respective domains, e.g.:

- ECLAP developed an ontology (called "vocabulary") related to performing arts
- PartagePlus developed a thesaurus related to Art Nouveau

Because these professional domains are relatively closed, these schemes were developed by the respective project partners, are relatively small, and often could be sourced or correlated from a few existing datasets

- The ECLAP vocabulary[3] has about 10 classes and 90 properties

[3] http://www.eclap.eu/schema/eclap/

- The PartagePlus thesaurus was gathered by the partners and inspired by Getty AAT. The majority of concepts (97%) were successfully matched to AAT [Kailus 2014, slide 7]

However, we believe that the EFD classification faces a unique problem:

- The activities of obtaining, producing, consuming and enjoying FD hold central importance for humanity, therefore FD has a prominent role in human culture
- Many objects from various collections, and many topics of discussion, have (or may conceivably be judged to have) a relation to FD.
- The variety of content associated with FD is **staggering** (see next section)
- Therefore we cannot hope that the FD classification can be limited to a small number of controlled entries, can be assembled by the EFD project partners alone, or can be considered as a closed, final, frozen artefact

1.9 Variety of Content

The variety of CHOs collected by EFD partners is staggering (see the EFD content survey):

- books on Bovine care & feeding (TEL)
- book on tubers/roots used by New Zealand aboriginals (RLUK)
- self-portraits involving some food (Slovak National Gallery)
- traditional recipes for Christmas-related foods (Ontotext)
- colourful pasta arrangements (Horniman)
- 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 & 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 consomme
- etc etc

The sample classifications in sec. 3.15 show that we need a wide set of classification elements.

The variety of CHOs related to FD that already exist in Europeana is even bigger, see sec. 5.9.

Our thesis that we need an open-ended classification is further demonstrated by the variety of examples in the dataset investigations (sec.0).

1.10 Multilingual Requirements

EFD covers content in 11 languages. In addition to English content, the content in the following languages has been translated into English:

- Hungarian
- Lithuanian
- Romanian
- Dutch
- French

Content in the following languages has not been translated into English yet:

- Greek
- Italian
- Polish
- Bulgarian
- Spanish

Additional Europeana content (see sec. 6) will increase the number of languages to over 20. Therefore strong preference is given to data sources that have good coverage across a significant number of the required languages. This is important both for:

- semantic enrichment: be able to recognize indexing terms in text in various languages
- semantic concept search: allow the user to query in his own language, and find objects referring to the concept, not a particular language term. The dangers of naïve multilingual expansion have been described in sec. 1.7

1.11 Approach

We believe that it is not productive to aim for a **closed** or **final** EFD classification. See in particular sec.3.8. Rather, it should be an **open-ended** artefact (dataset) that:

- is continuously refined through human-computer interaction
- descends the timeframe of the EFD project and is hopefully adopted by Europeana for the upcoming FD Channel

So we adopt an iterative approach, using machine learning and human-computer interaction:

- Whenever a concept/category is used to tag a CHO, we mark it as appropriate to the domain.
- We also trace backward toward the root (category "Food and drink", see sec. 3.8.3) and mark all categories along the way as appropriate.
- Then we leverage semantic enrichment to find other CHOs that mention the same concept/category

- We show these candidate CHOs to the user and ask for feedback, i.e. to point out some positive and some negative examples
- We learn from the negative examples, e.g. if many of the rejected CHO have the word "fragment" or "shard" (see sec. 6.1), we put it on a blacklist.
- The user can cut out branches from the category hierarchy as inappropriate (see sec. 3.8.4). This is done in a crowd-sourced fashion.

In this way we achieve a positive feedback loop:

- Confirmed concepts/categories are used to discover more CHOs relevant to FD
- Confirmed CHOs are used to augment the category hierarchy by marking the directly applied and parent categories as appropriate
- Confirmed CHOs are used to suggest new terms for the classification
- Disconfirmed CHOs are used to learn terms for the black list

We call this **dual semantic enrichment**, since both:

- appropriate objects are discovered and enriched with confirmed categories, and
- the set of confirmed categories is augmented

This is similar to the Wikipedia BiTaxonomy approach to building article-and-category hierarchies (see sec. 3.18), but that is built automatically in an unsupervised fashion.

In our case we need a human-computer interaction loop, since often the appropriateness judgements are subjective. Implementing an attractive interactive application for this is a major task of the Semantic Demonstrator.

2 Classification Facets

The EFD classification has to cover the following dimensions (facets). The first few (places, cultures, people) are generic and there are well-known sources to cover them, so we deal with them in his section, referring to sec. 3 Datasets as needed

For the last few (concepts, events, festivities) we have to look for FD-specific datasets (or subsets thereof), so we discuss them in detail in sec. 3 Datasets

2.1 Places

The spatial relation or coverage of a CHO is very important, as it provides a useful semantic facet that enables two important searches:

- Geo-spatial, e.g. by coordinate bounding box or radius (nearby). Ontotext GraphDB has a special geo-spatial index[4] to enable such queries.
- By place hierarchy: administrative and/or physical. Administrative places are defined by political boundaries (present or historic) and include countries,

[4] https://confluence.ontotext.com/display/GraphDB6/GraphDB-SE+Geo-spatial+Extensions

populated places, etc. Physical places are defined by physical features and include continents, mountains, rivers, etc.

The EFD metadata schema should be able to capture the role that a particular place has towards the CHO, e.g.:

- Made
- Used
- Found

Datasets:

- Wikipedia (sec. 3.7.1) / DBpedia / Wikidata have 818k places
- Getty TGN (sec. 3.2) has 1.2M places, but a strong North American bias (800k). It has a rich place type system (deeper than GeoNames)
- GeoNames[5] has the widest coverage with 9.6M places, ranging from continents to notable streets, squares, hotels. The problem is disambiguation (see below).

GeoNames has 680 Feature Codes organized in 9 Feature Classes:

Cl	Feature Class	Count
A	Administrative Boundary Features (country, state, region,...)	303,814
H	Hydrographic Features (stream, lake, ,,,)	1,742,469
L	Area Features (parks, area, ,,,)	309,264
P	Populated Place Features (city, village,,,,)	3,256,345
R	Road / Railroad Features (road, railroad)	36,658
S	Spot Features (spot, building, farm)	1,812,606
T	Hypsographic Features (mountain, hill, rock,,,,)	1,081,794
U	Undersea Features (undersea)	13,955
V	Vegetation Features (forest, heath,,,,)	30,784
	undefined	5,293
	Total	**8,592,982**

Disambiguation of place names is a big problem, because people take place names with them when they emigrate, create new territories/states, or discover a new place.

- Search for Guadalajara and you'll find[6] 106 places, then a long list of places in "Spain> Castille-La Mancha> Guadalajara".
- Even if you limit to Mexico[7] you'll find 96 places, of which 17 populated places! The problem is how to filter out the small pueblos (population 5-10). To filter out the small pueblos you can include the feature Population in Machine Learning

But there are many other ambiguity problems to be resolved, e.g.

[5] http://www.geonames.org/
[6] http://www.geonames.org/search.html?q=guadalajara&startRow=100
[7] http://www.geonames.org/search.html?q=guadalajara&country=MX

- How to determine place type (feature code)
- How to use parent places to identify the correct place. E.g. the old Europeana enrichment tool (AnnoCultor) did not take them into account. Instead it had a rule: it used all countries, but sub-country places only in Europe. As a result, a CHO with explicit text "Guadalajara, Mexico" would be enriched with the entities Mexico and… Guadalajara, Spain.
- How to extract such useful features from the context (surrounding text)

2.2 Cultures

Getty AAT (sec. 3.2) has well-developed styles, periods, and cultures hierarchies.[8] Browse that URL to explore them. Using the ID from the URL, we can find the number with this query: 5491

```
select (count(*) as ?c) {?x gvp:broaderExtended aat:300264088}
```

Wikipedia probably also covers all these cultures, but the hierarchy in AAT makes it much easier to do query expansion. For example, a CHO classified as Culture = Austral Islands (sec. 5.1) will also be found[9] when querying for Polynesian or Oceanic.

So far we've encountered only one culture present in Wikipedia but not in AAT: Nyangatom (see sec. 5.6)

2.3 People and Organisations

Some FD-related CHOs are related to specific agents (people, organisations, conventions etc). The EFD metadata schema should be able to express the kind of relation, e.g.

- Creator
- Production role (painter, engraver, etc)
- Subject

Important agent datasets include:

- VIAF, [10] an aggregation of the Authority Files of 20 national libraries and 15 other contributors (including US LoC, DE GND, FR BnF and SUDOC, Getty ULAN etc)[11]
- Getty ULAN (sec. 3.2). It has detailed person info (names, dates, events, associative relations) and excellent provenance info.

[8] http://www.getty.edu/vow/AATHierarchy?find=&logic=&note=&subjectid=300264088
[9] http://www.getty.edu/vow/AATHierarchy?find=&logic=note=&subjectid=300021975
[10] http://viaf.org
[11] https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control#Resolve_against_VIAF_links

2.1 Europeana Food and Drink Classification Scheme

A recent EuropeanaCreative report [Alexiev 2015] researched in detail the available name datasets, and concluded the following.

- The best datasets to use for name enrichment are VIAF and Wikipedia/Wikidata (sec. 3.12)
- There are few name forms in common between the "library-tradition" datasets (dominated by VIAF) and the "LOD-tradition datasets" (dominated by Wikidata). VIAF has more name variations and permutations, Wikidata has more translations
- VIAF is much bigger: 35M persons/orgs. Wikidata has 2.7M persons and maybe 1M orgs
- Only 0.5M of Wikidata persons/orgs are coreferenced to VIAF, with maybe another 0.5M coreferenced to other datasets, either VIAF-constituent (eg GND) or non-constituent (e.g. RKDartists). A lot can be gained by leveraging coreferencing across VIAF and Wikidata. Wikidata has great tools for crowd-sourced coreferencing
- The best approach to use these datasets is to load them to a local repository, and resolve the coreferences, in order to ensure levels of service

A number of people were indicated as referenced in the contributed content during the Content Inventory process. We look them up by way of example:

- Henry Loveridge: **none**
- Frederick Hardy: VIAF 6 possible matches, WD 1 match
- George Washington: http://viaf.org/viaf/31432428, https://www.wikidata.org/wiki/Q23
- Matilde Balzan (Maltese TV Presenter): **none**, not even on https://mt.wikipedia.org
- Apicius. http://viaf.org/viaf/191439325, https://www.wikidata.org/wiki/Q114982 or https://www.wikidata.org/wiki/Q117841
  - VIAF also lists related people, e.g. André, Jacques (1910-1994) and works e.g. "Alimentation et la cuisine à rome"
  - WD describes him as " cook who found a way of packing fresh oysters in the second century CE" and lists him first
  - WD has 4 entries, of which the 2 listed above are about the same person (Marcus Gavius Apicius). VIAF has 1 entry for this person. This is an example that more coreferencing is needed, in this case to leverage from VIAF to WD.
- Nikosthenes https://viaf.org/viaf/96538320/, https://www.wikidata.org/wiki/Q275165
- Pieter de Pannemaeker (19th century Ghent artist): **none**

We can see that some FD-related people are not famous enough to have an entry yet. There are two possible approaches to deal with this:

- Create an entry in Wikidata. This is really easy (little data is required) and there are no restrictive notability criteria.
- Create an entry in the shared thesaurus (see sec. 2.6)

2.4 Concepts

This is the most complicated part of the classification, since it includes

- Object types
  - Objects that appear in any museum or gallery: photograph, still life, painting
  - FD-specific and Agricultural objects: containers, vessels, dishes, cutlery, field-work instruments (e.g. ploughs), objects made of food (e.g. see sec. 5.11), etc
- Agricultural and FD-related tools used as depicted subject
- Foods and Drinks: extensive lists of foods, varieties, brands, styles
  - Used as material
  - As depicted subject
- Food ingredients and materials. Allows search by ingredient (e.g. maize) or class of ingredient (e.g. grains)
- Activities related to FD and agriculture, e.g. eating, formal dining, bar-hopping, fasting, feasting, degustation, shepherding, ploughing, cooking, boiling, simmering
  - As depicted subject
  - As qualifier of the object
  - As qualifier of the food (e.g. cooking technique)
- Person types related to FD, e.g. chefs, brewers, vintners, barista…

2.5 Foods, Drinks and Ingredients

We believe that the Foods, Drinks and Ingredients concepts should form the core of the EFD classification, but their acquisition is also the hardest. We acquire them from Wikipedia categories, lists, portals and projects (sec 3.7 and following).

- Food classification, e.g. National/ Local/ Traditional
- Location (spatial coverage). This can be acquired from appropriate categories using little manual work. E.g. a food with category "X Cuisine" can be associated with place "X" (Bulgarian Cuisine → Bulgaria), but there is no perfect alignment between cuisines and places
- Period (time coverage). There is little reliable information to that effect. Spatio-temporal coverage would be even more interesting (e.g. Mamaliga was first used in Romania and was later "imported" to Bulgaria under the name Kacamak), but there is even less information. The topic of mobility and employment of recipes across cultures is complex and requires research by a food historian.
- Main ingredient used in a recipe/dish.

2.6 Events

- Food and Drink related events/festivals, e.g. beer fests, food fairs, Tomatina, etc
- Religious holidays, e.g. Easter, Christmas, Pentecostal fast. Their relation to food, e.g. Christmas foods, Easter foods, food during fasting
- Festivities, customs, traditions, holidays related to Food
- Historic and remembrance days, e.g. World War I or Independence Day celebrations

- Human life events, e.g. weddings, christenings, funerals
- Agricultural work cycles and associated festivals, e.g. crop harvest, Autumn olive gathering

While the above are event types or periodic events, specific events are also interesting:

- Historic events, e.g. Boston Tea party
- Broader historic periods, e.g. the Stuarts, the Industrial revolution, the opium wars and the interwar period

We acquire these from Wikipedia lists (3.9) and categories (3.8)

2.7 Themes

EFD consortium partners have expressed an interest in defining a small thesaurus of common topics of Interest (themes). These themes can provide an important feature for the applications. Themes under consideration may include the following (the list and hierarchy is by no means final). Please note that these are fixed values, not lists of values (e.g. lists of fests/events are covered in the previous section):

- Cultural and Traditions
  - customs and traditions
  - heritage foods and recipes
  - cultures and food
- Industrial and Industrial/craft
  - agriculture
  - traditional food production and
- Time-based themes (see previous section), e.g.
  - Daily life
  - Traditional holidays, remembrances, feasts
- Socio-cultural phenomena
  - famine
  - immigration
  - emigration
  - economic crisis
  - war-time food and advice
  - nostalgia
- Social use of food and drink
  - food fests
  - wine and beer fests
  - drinking culture
  - healthy eating

If the themes are a small number (10-100), we'll create them in a shared spreadsheet and then convert to SKOS.

If the themes (or other shared classifications) will be bigger (hundreds or thousands of entries), a more serious collaboration environment needs to be deployed. An appropriate tool is VocBench[12], an open source SKOS+SKOSXL editor that works directly over semantic repository (Ontotext GraphDB). Although technically more complex, this approach is more scalable, and has the benefit that it can incorporate already existing SKOS thesauri that need to be edited, translated or customized.

3 Datasets

This section (representing the bulk of this report) is dedicated to an investigation of existing datasets that we can use for the EFD classification. We have a strong preference for thesauri available as Linked Open Data (LOD) in RDF, since this allows semantic integration by loading them to the semantic repository adopted by EFD (Ontotext GraphDB)[13].

If a dataset that we want to reuse is not available as LOD but in some other structured format, we convert it to RDF. In some cases we provide links to downloadable CSV or other files, or to web pages that best present the information about a particular item.

3.1 Thesauri Used by Content Partners

A survey was performed[14] asking the content partners what kinds of thesauri they use. The results are reproduced below.

- CAG: Dutch version of the AAT (Am-MovE), and a specialized 'CAG-thesaurus' for specific agricultural objects
- IAPH: own thesaurus[15] that is available on the web[16]
- Wolverhampton: tried to use the British Museum object thesaurus
- RMCA: Getty AAT
- Horniman: use a thesaurus but is not clear which one
- VUFC: use a thesaurus but is not clear which one
- The other partners do not use particular thesauri

The EFD classification takes into account the ones that are used. They will be integrated in the semantic store and used during classification. However, each partner needs to take care to coreference their thesaurus to some of the global LOD datasets described below.

[12] http://aims.fao.org/vest-registry/tools/vocbench-2
[13] http://graphdb.ontotext.com/
[14] https://basecamp.com/2069212/projects/7016737/messages/31036412
[15] http://www.iaph.es/web/canales/conoce-el-patrimonio/tesauro-pha/
[16] http://www.iaph.es/tesauro/init.htm

3.2 Getty Thesauri

The Getty Art and Architecture Thesaurus (AAT) and the Thesaurus of Geographic Names (TGN, another Getty thesaurus) were published as LOD in 2014[17]. The Unified List of Artist Names (ULAN) will be published in Apr 2015. A SPARQL endpoint is provided[18] and the representation is well documented ([Alexiev 2014b]), including sample queries.[19]

AAT is a very important thesaurus in CH. It's been in development for over 25 years and is famed for its good organisation/structure and adherence to standards.

- The International Terminology Working Group (ITWG) headed by Getty coordinates translation efforts, thus AAT is fully translated to Dutch, Spanish and Chinese. Other translations (notably German) are in progress
- AAT includes almost no pre-coordinated terms, which allows more flexibility in classification and avoids a combinatorial explosion. This is unlike e.g. Library of Congress Subject Headings (LCSH) that has subjects such as "16th century Italian love poetry"
- AAT includes intermediate levels (hierarchies and guide terms) used to organize the thesaurus. It goes 11 levels deep
- AAT distinguishes between kinds of broader relation (broader generic vs broader partitive), which allows proper composition of the broader relation (see [Alexiev 2014a])
- AAT has 43016 concepts, 307421 labels and 45091 "broader" relations. Thus every concept has an average of 1.048 parents: even though there are multiple parents (1 preferred and potentially several non-preferred), AAT is close to a mono-hierarchy. The number of relations was obtained with this query:

```
select (count(*) as ?c) {?x gvp:broader ?y. ?x skos:inScheme aat:}
```

AAT is organized by kind of concepts in the following major groupings (facets):[20]

- Concepts
- Physical Attributes
- Styles and Periods
- Agents
- Activities
- Materials
- Objects
- Brands

AAT is a multi-hierarchy: in addition to a preferred parent, many concepts have secondary parents. For example, this query lists the concepts with most ancestors:

[17] http://vocab.getty.edu
[18] http://vocab.getty.edu/sparql
[19] http://getty.ontotext.com/doc/#Sample_Queries
[20] http://www.getty.edu/vow/AATHierarchy?find=&logic=AND&note=&subjectid=300000000

```
select ?x (count(*) as ?parents) {
  ?x skos:inScheme aat: ; gvp:broaderExtended ?y
} group by ?x order by desc(?parents)
```

Protomai capitals[21] has 46 ancestors and is nested 11 levels deep[22]:

Protomai capitals < capitals (column components) < <capitals and capital components> < column components < <columns and column components> < <supporting and resisting elements> < structural elements < <structural elements and components for structural elements> < architectural elements < <components by specific context> < components (objects) < Components (Hierarchy Name) < Objects Facet

All Getty thesauri include names in various languages, including historic names and period of use, and strong scholarly attestation (sources and contributors for every item and every name). They have not only a well-organised hierarchy, but also rich associative relations, e.g. including "technique used-for product" (ULAN), "ally of" (TGN), "person-student" (ULAN).

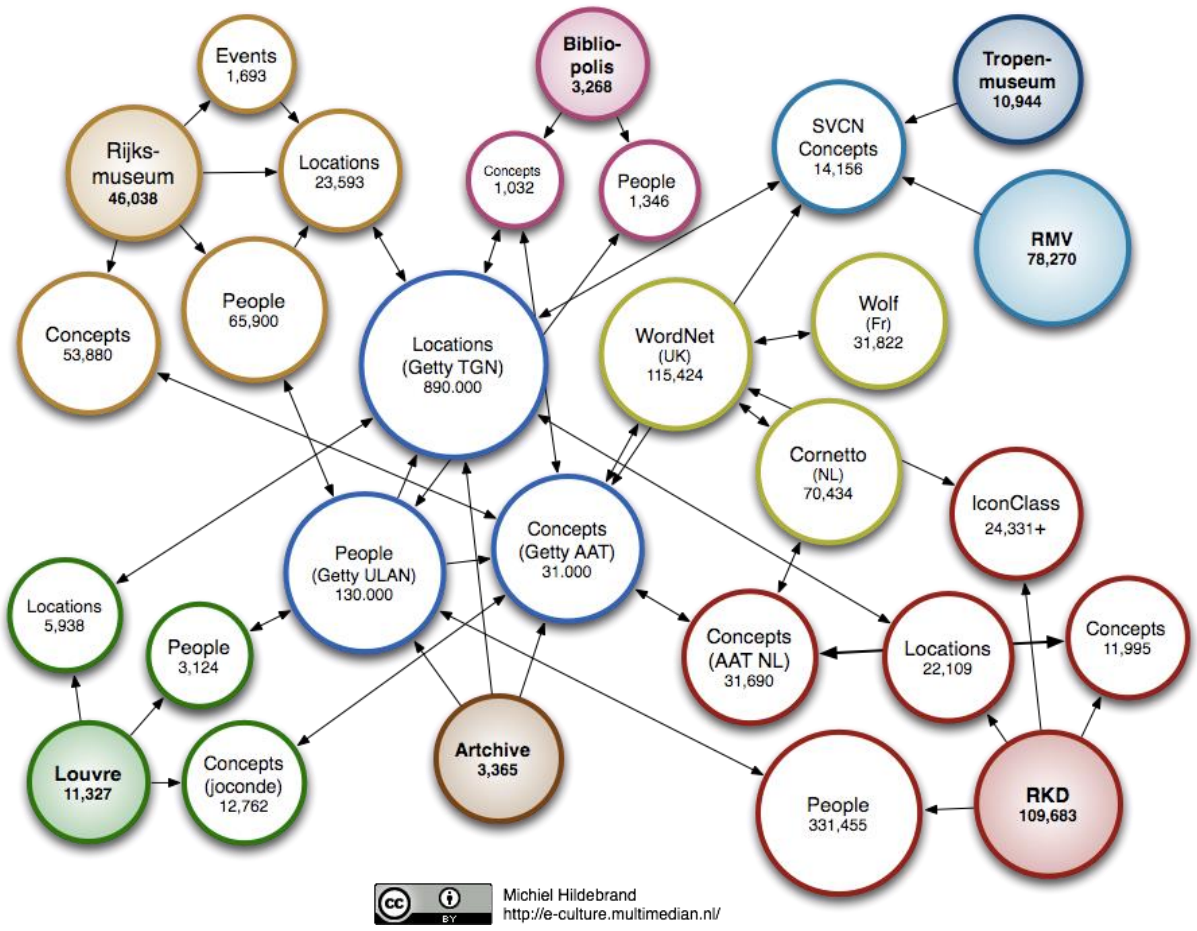
TGN has 1.2M places, with a large concentration in North America (800k). In comparison, DBpedia has 850k places and GeoNames has 8M. TGN places have a rich system of types (deeper than the Feature Codes of GeoNames), latitude/ longitude, and a few have altitude.

ULAN has 195k persons, 40k organisations, and several hundred "unknown person of given nationality". ULAN has person types, biographies, life events including type, dates and place.

CONA describes works of art so it has very rich structure, but is still in formation (10-20k objects).

The Getty thesauri are quite well coreferenced due to their central role in the Cultural Heritage LOD cloud:

[21] http://vocab.getty.edu/aat/300001695
[22] http://www.getty.edu/vow/AATHierarchy?find=&logic=AND&note=&page=1&subjectid=300001695



People have been connecting to them long before their publication as LOD. In addition to various thesauri, they are fairly well coreferenced in Wikidata. See [Alexiev 2015 sec 5] for a description of Wikidata coreferencing facilities.

EFD Applicability

AAT has about 1k concepts relating to FD, e.g.

- porringers:[23] Small, shallow, flat-bottomed bowls or basins, used for soup, stews, or similar dishes.
- porringer spoons:[24] designed to serve soup, stews, or similar dishes from a porringer
- biggins:[25] coffeepots having separate containers, often in the form of a muslin bag
- drawknives:[26] Two-handled curved knives utilized by pulling them toward the user
- backing drawknives:[27] Drawknives with a slight downward thrust of the handles for cutting outside curves

[23] http://vocab.getty.edu/aat/300042964
[24] http://vocab.getty.edu/aat/300043165
[25] http://vocab.getty.edu/aat/300215510
[26] http://vocab.getty.edu/aat/300023702
[27] http://vocab.getty.edu/aat/300023709

2.1 Europeana Food and Drink Classification Scheme

There is a whole hierarchy of Culinary Equipment[28] with 866 items:

```
select (count(*) as ?c) {?x gvp:broaderExtended aat:300199765}
```

One can get a list of these items with a query like this:

```
select * {?x skos:broaderExtended aat:300199765;
  gvp:prefLabelGVP/xl:literalForm ?l}
```

These AAT hierarchies are relevant to FD. They are largely orthogonal, so we can sum the items.

AAT id	Hierarchy	Items	Note
300199765	Culinary Equipment	866	
300024716	<grinding and milling equipment>	15	Total 21, about 15 are FD
300254496	Food	85	
300004332	<eating and drinking spaces>	19	
300203437	closures (container components)	8	
300006335	food processing plants	10	
300120719	beverage processing plants	6	
300069097	feasts	4	
	Total	**1013**	

As you can see from the last few, we're "fishing" concepts since AAT does not have a hierarchy or collection dedicated to FD. Since AAT is dedicated to visual arts, it's missing basic FD concepts such as chef, vintner, aioli, etc. AAT can be used in the EFD classification in various ways:

- The FD hierarchies described above
- The whole Objects Facet, since any object (e.g. painting, poem, etc) may have a **subject** relating to FD
- The AAT organization into facets is well-established and should be used as a model for any EFD dedicated thesaurus

TGN is a place thesaurus, so it is generally applicable to FD. ULAN is somewhat applicable, but it doesn't have "food" people (chefs, vintners etc) unless they also had some relation to art.

The Getty thesauri are based on SKOS, SKOS-XL and a number of auxiliary ontologies (see the documentation). EFD is in a unique position to utilize them fully because of ONTO's involvement in their publication, and understanding of the semantic representation.

[28] http://www.getty.edu/vow/AATHierarchy?find=&logic=AND&note=&subjectid=300199765

2.1 Europeana Food and Drink Classification Scheme

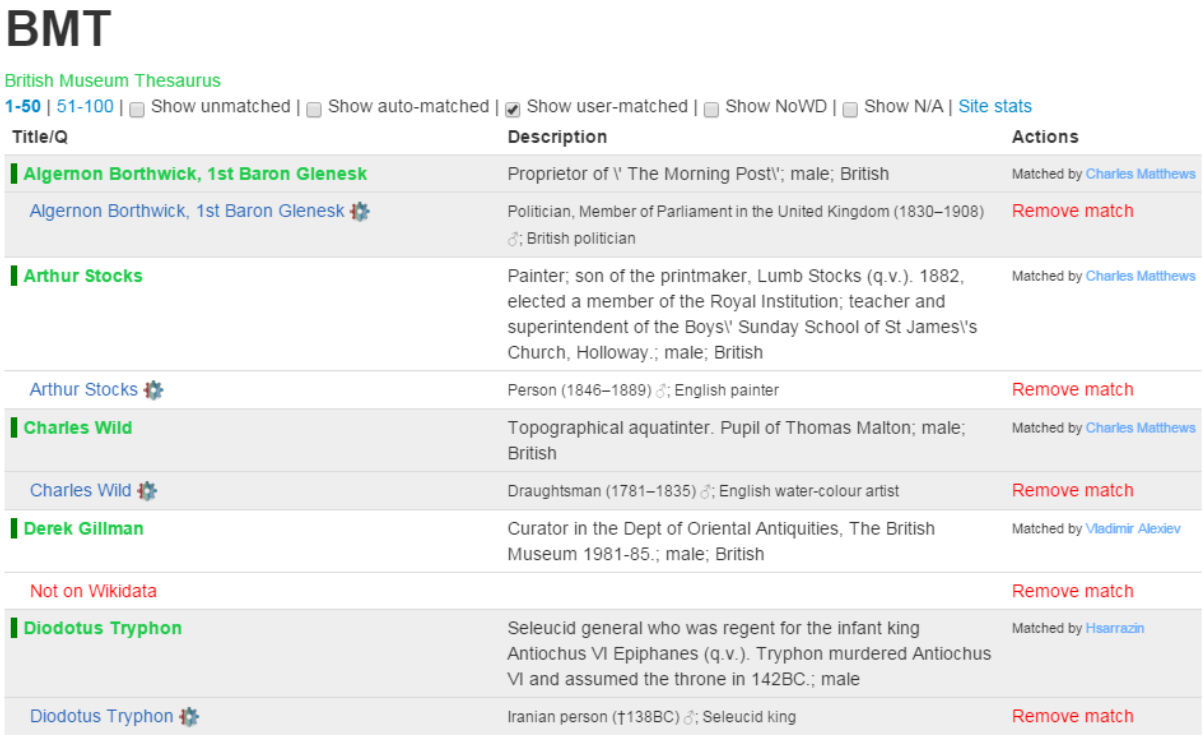
3.3 British Museum Thesauri

The British Museum uses a large number of thesauri[29] (this page also lists Yale Center for British Art thesauri). They include person-institution, place, place type, dimension, acquisition association, production association, inscription type, material, technique, object type, etc.

- They are available as LOD and SPARQL query endpoint (published by ONTO as part of the ResearchSpace project) at http://collection.britishmuseum.org/sparql
- The Portable Antiquities Scheme has extracted the thesauri as CSV[30].

The BM thesauri are not coreferenced to any other thesaurus or dataset. ONTO initiated the process of coreferencing to Wikidata, starting with the BM person-institution thesaurus:

- A corresponding property was added to Wikidata[31]
- An enriched CSV dataset was produced[32]
- The dataset was loaded to the Mix-n-Match tool[33] by Wikimedia Germany (listed as "BMT")
- Coreferencing has started, e.g. see below



BMT

British Museum Thesaurus

1-50 | 51-100 | Show unmatched | Show auto-matched | Show user-matched | Show NoWD | Show N/A | Site stats

Title/Q	Description	Actions
Algernon Borthwick, 1st Baron Glenesk	Proprietor of \' The Morning Post\'; male; British	Matched by Charles Matthews
Algernon Borthwick, 1st Baron Glenesk	Politician, Member of Parliament in the United Kingdom (1830–1908) ♂; British politician	Remove match
Arthur Stocks	Painter; son of the printmaker, Lumb Stocks (q.v.). 1882, elected a member of the Royal Institution; teacher and superintendent of the Boys\' Sunday School of St James\'s Church, Holloway.; male; British	Matched by Charles Matthews
Arthur Stocks	Person (1846–1889) ♂; English painter	Remove match
Charles Wild	Topographical aquatinter. Pupil of Thomas Malton; male; British	Matched by Charles Matthews
Charles Wild	Draughtsman (1781–1835) ♂; English water-colour artist	Remove match
Derek Gillman	Curator in the Dept of Oriental Antiquities, The British Museum 1981-85.; male; British	Matched by Vladimir Alexiev
Not on Wikidata		Remove match
Diodotus Tryphon	Seleucid general who was regent for the infant king Antiochus VI Epiphanes (q.v.). Tryphon murdered Antiochus VI and assumed the throne in 142BC.; male	Matched by Hsarrazin
Diodotus Tryphon	Iranian person (†138BC) ♂; Seleucid king	Remove match

[29] https://confluence.ontotext.com/display/ResearchSpace/Meta-Thesaurus+and+FR+Names#Meta-ThesaurusandFRNames-Metathesaurustable
[30] https://github.com/findsorguk/bmThesauri
[31] https://www.wikidata.org/wiki/Property_talk:P1711
[32] https://github.com/VladimirAlexiev/bmThesauri/raw/master/bmPerson-institution-better.tsv.gz
[33] https://tools.wmflabs.org/mix-n-match/

EFD Applicability

The BM Object Type thesaurus has 5845 entries. There is no hierarchy and about 500 are related to FD, e.g.

URL	Concept
http://collection.britishmuseum.org/id/thesauri/x5112	agricultural equipment
http://collection.britishmuseum.org/id/thesauri/x5113	agricultural tool/implement
http://collection.britishmuseum.org/id/thesauri/x5413	beer-bottle
http://collection.britishmuseum.org/id/thesauri/x5414	beer-container
http://collection.britishmuseum.org/id/thesauri/x72562	beer-dipper
http://collection.britishmuseum.org/id/thesauri/x5415	beer-glass
http://collection.britishmuseum.org/id/thesauri/x5416	beer-jug
http://collection.britishmuseum.org/id/thesauri/x5417	beer-mug
http://collection.britishmuseum.org/id/thesauri/x5418	beer-pot
http://collection.britishmuseum.org/id/thesauri/x5419	beer-skimmer
http://collection.britishmuseum.org/id/thesauri/x5420	beer-strainer
http://collection.britishmuseum.org/id/thesauri/x5421	beer-trough
http://collection.britishmuseum.org/id/thesauri/x5422	beer-vessel

3.4 US Department of Agriculture Standard Reference

The USDA SR is a massive dataset describing nutritional values of some 10k foods. It covers not only the 3 major nutrients, but about 375 micro-nutrients as well. ONTO is familiar with this dataset, since it was used in the development of Edamam[34], a semantic recipe search engine.

EFD Applicability

USDA SR is not appropriate for EFD, because its focus is on nutrition not culture, and the collected CHOs rarely identify the specificity of food identification needed by SR (e.g. it describes over 100 kinds of steaks).

3.5 FAO AGROVOC

AGROVOC is a controlled vocabulary developed by the UN Food and Agriculture Organization (FAO).[35] It has been in development for 24 years and covers food, nutrition, agriculture, fisheries, forestry, environment, biology, etc.

AGROVOC has 32k concepts in 21 languages (Arabic, Chinese, Czech, English, French, German, Hindi, Hungarian, Italian, Japanese, Korean, Lao, Persian, Polish, Portuguese, Russian, Slovak, Spanish, Thai, Turkish and Ukrainian).

[34] http://edamam.com
[35] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

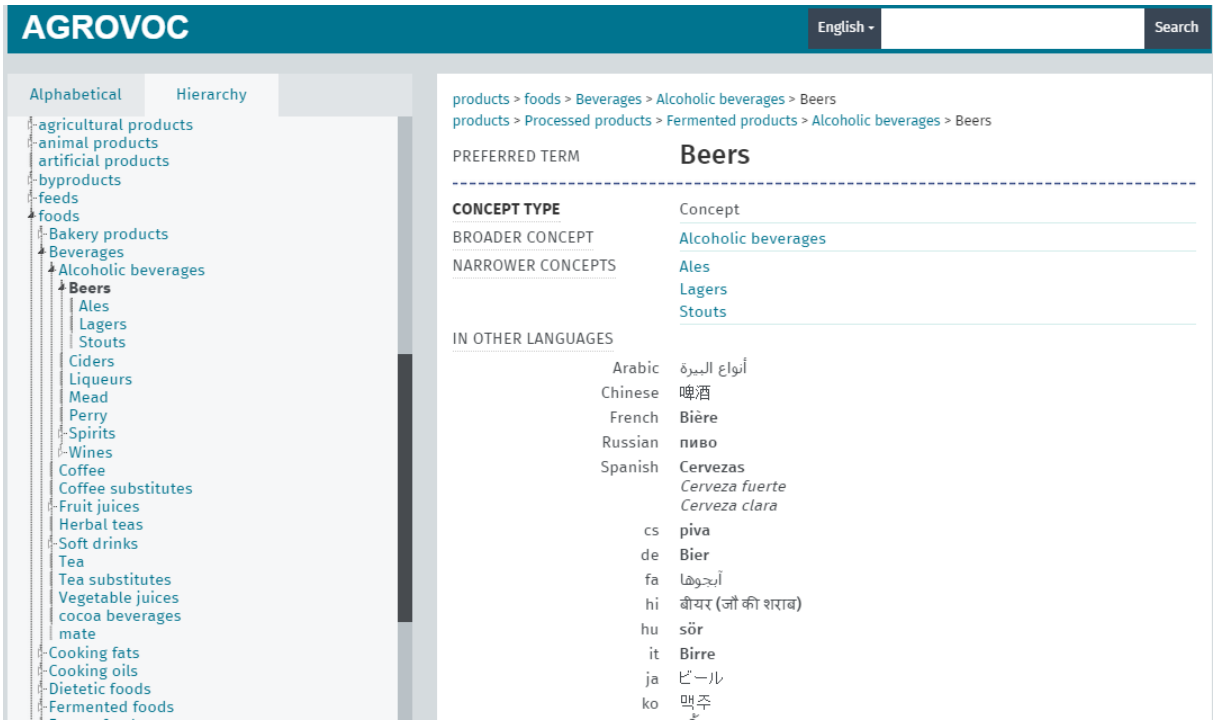
2.1 Europeana Food and Drink Classification Scheme

AGROVOC is aligned (coreferenced) with 13 other multilingual SKOS[36] related to agriculture, including the US National Agricultural Library Thesaurus (NALT).

AGROVOC is managed and edited with VocBench[37], an open source SKOS editor that works directly with a semantic repository (Ontotext GraphDB). AGROVOC is available for searching/ browsing on the web,[38] and also as LOD[39] based on SKOS and SKOS-XL.

EFD Applicability

Parts of AGROVOC are applicable to EFD. For example, here is the information for "beer"[40]. In contrast, AAT doesn't have an entry for "beer".



However, AGROVOC contains many items of specialized technical or scientific interest. Conversely, AGROVOC does not contain enough items relating to:

- Brands or varieties. AGROVOC has only 3 varieties, whereas Wikipedia lists some 50 Beer styles[41], another 25 Types (including Boza,[42] which I do not believe is a beer: it is sold to kids in confectionery shops in Bulgaria). It also has a great variety of brands, e.g. see German beers.[43]

---

[36] http://aims.fao.org/aos/agrovoc/void.ttl
[37] http://aims.fao.org/vest-registry/tools/vocbench-2
[38] http://aims.fao.org/standards/agrovoc/functionalities/search
[39] https://aims-fao.atlassian.net/wiki/display/AGV/Releases
[40] http://aims.fao.org/skosmos/agrovoc/en/page/c_864
[41] https://en.wikipedia.org/wiki/List_of_beer_styles
[42] https://en.wikipedia.org/wiki/Boza
[43] https://en.wikipedia.org/wiki/Category:German_beer_styles

- Traditions, festivities, rituals, festivals, even silly things like drinking games[44]

If we may so summarize: for the purposes of EFD, AGROVOC puts too much emphasis on the scientific and production side of food, and not enough on the consumption, fun and appreciation side. For this reason we have not investigated AGROVOC and will turn to it only if we need a concept we cannot find elsewhere.

3.6 EC EUROVOC

The EUROVOC Thesaurus is the official EC thesaurus for dealing with all kinds of topics that are important to government business:

- It is a well-established and widely used thesaurus, in particular in EC and legal publishing (e.g. the EC Publications Office, EuroLex, etc)
- It is multidisciplinary and includes an “Agri-Foodstuffs” hierarchy (which is however poorer than AGROVOC)
- It is multilingual, including terms in 23 languages of the EU plus Serbian.
- It is published in SKOS schema as an LOD dataset.

EUROVOC can also be loaded in VocBench over Ontotext GraphDB. However, we have estimated its relevance to EFD to be much lower than AGROVOC, and therefore have not investigated it further.

3.7 Wikipedia

With the previous section and the examples in sec.3.15 we hope to have set the stage for the idea that specialized thesauri like AAT (visual arts) and AGROVOC (agriculture) will cover only partly a topic as wide as FD. It would be possible to cover the domain using combinations of terms from several thesauri. LOD makes such data integration relatively easy, at least as it comes to data schemas. But still it's much easier to use data from a fewer number of sources, because then the problems of schema alignment, update schedule and other integration issues become simpler.

So we turn to a source of encyclopedic nature. Wikipedia has the goal to become "the Sum of all Knowledge", freely available and easily editable by anyone. It is gradually working towards this source. The volume of information is enormous. Some numbers for the EFD languages as of Feb 2015:

Wikipedia	URL	articles
English	http://en.wikipedia.org	4,774,396
Dutch	http://nl.wikipedia.org	1,804,691
French	http://fr.wikipedia.org	1,579,555
Italian	http://it.wikipedia.org	1,164,000
Spanish	http://es.wikipedia.org	1,148,856
Polish	http://pl.wikipedia.org	1,082,000

[44] https://en.wikipedia.org/wiki/Beer_pong

Bulgarian	http://bg.wikipedia.org	170,174
Hungarian	http://hu.wikipedia.org	272,323
Romanian	http://ro.wikipedia.org	256,022
Lithuanian	http://lt.wikipedia.org	168,823
Greek	http://el.wikipedia.org	102,077
Total		**12,522,917**

- Statistics are obtained[45] from[46] several[47] sources, and in some cases estimated.
- Articles counts only content pages (articles). E.g. enwiki 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

There is a great overlap between Wikipedia language editions through the so-called Inter-language links (lower left of every article). For a popular thing, there are many articles that talk about it in the different editions. For example, there are 730 persons described in over 20 of the 28 language editions integrated in DBpedia2014[48]. enwiki often serves as a central hub, since it's the largest.

This overlap is a major strength, since it enables even "lesser" language partners (e.g. Greek, Bulgarian) to index content using another language that they understand (most often enwiki).

3.7.1 Language Overlap and Total Things

It is uncertain how many thing the above 12.5M articles describe. Since a similar thing may have widely differing articles in two languages, reflecting local traditions, knowledge and experience. E.g.

- https://en.wikipedia.org/wiki/Mămăligă is the same as https://bg.wikipedia.org/wiki/Мамалига, and
- https://en.wikipedia.org/wiki/Kačamak is the same as https://bg.wikipedia.org/wiki/Качамак
  - But are the two the same or different?
- How different is https://en.wikipedia.org/wiki/Polenta from those two, really?

We estimate the degree of language overlap from DBpedia[49] and its cross-language statistics[50], and also obtain useful breakdowns by type of thing (Person, Place, Organization, Work…)

[45] https://en.wikipedia.org/wiki/Wikipedia:Statistics
[46] http://stats.wikimedia.org/EN/Sitemap.htm
[47] http://stats.wikimedia.org/EN/TablesArticlesTotal.htm
[48] http://wiki.dbpedia.org/Datasets2014/CrossLanguageOverlapStatistics?v=hj1
[49] http://wiki.dbpedia.org/Datasets2014/DatasetStatistics
[50] http://wiki.dbpedia.org/Datasets2014/CrossLanguageOverlapStatistics?v=hj1

2.1 Europeana Food and Drink Classification Scheme

	things	tot.art	ovl	EFD thing	EFD.art	en	nl	fr	it	es	pl	hu	bg	el
Person	1,471,254	2,571,237	1.75	1,182,684	2,066,917	1,445,104	54,879	134,749	189,448	99,147	96,135	22,444	19,047	5,964
Athlete	305,848	648,646	2.12	243,936	517,343	268,773	26,113	65,782	67,932	31,527	45,890	5,919	3,625	1,782
Actor	38,881	95,609	2.46	23,242	57,153	6,501	8,117	14,019	508	13,831	10,106	2,519	1,552	0
Artist	164,501	334,625	2.03	111,496	226,803	96,282	16,656	32,562	15,621	34,898	20,180	6,633	2,404	1,567
Musical Artist	67,142	136,621	2.03	48,685	99,065	45,089	5,959	11,138	15,113	14,594	6,924	0	48	200
Politician	61,636	91,206	1.48	52,994	78,418	40,343	1,805	11,461	4,893	7,460	10,639	1,025	0	792
Place	818,539	2,362,619	2.89	577,498	1,666,883	735,062	202,393	148,586	177,524	156,377	211,084	19,992	10,865	5,000
Populated Place	518,727	1,850,430	3.57	362,162	1,291,923	478,351	183,335	118,716	160,582	133,947	191,208	16,331	5,321	4,132
Building	77,852	94,795	1.22	73,546	89,552	68,582	2,373	6,926	3,888	4,455	2,549	513	236	30
Airport	13,958	28,393	2.03	11,192	22,767	13,649	1,050	1,499	1,069	1,921	3,392	187	0	0
Bridge	4,239	6,078	1.43	3,392	4,864	3,543	305	0	216	444	249	107	0	0
River	29,254	58,977	2.02	18,170	36,632	26,295	1,257	3,957	2,099	3	1,859	730	378	54
Organisation	266,551	428,833	1.61	210,681	338,948	241,286	12,234	27,542	15,554	15,955	15,288	4,829	3,924	2,336
Company	64,308	105,412	1.64	48,800	79,991	58,400	2,490	8,180	5,337	1,077	3,054	831	430	192
Educ.Institution	51,261	65,714	1.28	44,308	56,801	49,172	775	2,943	918	1,709	845	158	175	106
Band	33,183	63,179	1.90	23,324	44,407	30,572	2,057	5,177	0	0	4,054	949	1,348	250
Sports Team	35,287	84,204	2.39	26,404	63,008	28,357	5,751	5,844	7,900	5,585	5,048	2,322	1,032	1,169
Work	462,124	913,895	1.98	355,517	703,070	411,295	30,126	61,212	78,975	50,374	37,363	12,449	5,204	16,072
Musical Work	185,644	357,081	1.92	150,470	289,425	180,308	8,774	22,065	31,309	21,379	17,406	5,374	1,792	1,018
Album	134,199	249,303	1.86	109,975	204,302	123,374	4,786	14,426	30,252	12,897	12,406	3,952	1,364	845
Single	47,230	88,386	1.87	36,444	68,202	45,433	2,982	5,621	0	7,296	5,000	1,276	428	166
Film	92,490	246,316	2.66	63,253	168,452	87,282	10,239	15,669	24,156	12,140	12,555	3,188	2,095	1,128
Book	64,965	83,033	1.28	46,799	59,815	31,029	953	3,549	6,083	2,217	1,687	608	556	13,133
Software	33,154	81,174	2.45	24,663	60,384	31,401	3,878	8,980	7,145	6,284	1,187	1,105	99	305
Television Show	31,843	55,378	1.74	25,874	44,998	29,466	2,077	4,373	570	3,544	3,071	1,152	559	186
Event	74,104	130,250	1.76	60,320	106,023	45,377	5,552	23,123	18,118	5,050	4,064	3,627	753	359
Celestial Body	35,489	116,881	3.29	24,358	80,222	32,864	1,666	0	16,974	2,541	13,312	10,581	138	2,146
Species	278,793	667,278	2.39	207,621	496,930	252,166	129,539	0	22,810	64,950	23,474	94	3,897	0
Disease	6,628	17,064	2.57	5,125	13,194	6,078	1,088	0	1,868	2,397	1,448	0	29	286
TOTAL	**3,413,482**	**7,208,057**	**2.11**	**2,591,435**	**5,472,187**	**3,169,232**	**437,477**	**395,212**	**521,271**	**396,791**	**402,168**	**74,016**	**43,857**	**32,163**
Food	6,606													

- Total articles for the selected types are 7.2M. Total scores of items are 3.4M, thus the degree of overlap **ovl** is 2.11 (each thing is described in that many articles)
- The articles for the selected types in the EFD languages are 5.4M (lt and ro are missing from this statistic). We use the same **ovl** factor, which is imprecise but is our best guess. (With more languages, the popular things will be described in more articles, but also more unique articles will be introduced for local things).
- So we come to an estimate of 2.6M things for the selected types in the EFD languages.

The results are also available in more readable form[51] (tab EFD). But perhaps not all of these types of things are relevant to FD

- Celestial Bodies: certainly remove all, unless we want to discuss the possibility that the Moon is made of Green Cheese
- Species: there are a lot of edible species (e.g. a whole category of articles on Maize and maize products), but the majority of taxonomic classification is not related to FD. Remove 80%
- Events: Plenty of events are related to food (festivities, religious holidays…)
- Places: all are potentially relevant
- Works, Organizations, Persons: remove 70-80% since not that many FD-related works, organizations and people are notable.

[51] https://docs.google.com/spreadsheets/d/1O5Y6q3tOCGoxQeI5Ydo8mukcIN3VNvynoX0wzQknD34/edit

In conclusion, there are a million or so items potentially usable the in EFD Classification.

3.7.2 Total Pages Related to EFD

The table in sec 3.6 totalled 12.5M articles in the EFD languages, while the one in 3.7.2 has 5.4M. Discounting 400k for the two missing languages (lt, ro), that still leaves a shortage of 6.7M articles.

The reason is that the things counted in the DBpedia type statistics above are Named Entities (NE): articles that are mapped to a class (see sec. 3.11). (Another reason is that the table in 3.7.2 counts only the major NE classes, but leaves a long-tail of other types out)

Conceptual articles that are not NE are just as important for the EFD Classification. If you compare the NEs (e.g. Places, Persons) to TGN and ULAN, you can compare the conceptual articles to AAT. They provide core concepts for classification.

Taking the total 12.5M articles in the EFD languages and applying the overlap ratio 2.11, we estimate 5.9M **things or concepts** of any type in all EFD Wikipedia editions.

It's very hard to venture a guess as to how many of these articles are related to FD, because this is a subjective question. But our experimental investigations suggest that at least 10-20% or 600k to 1.2M Wikipedia things/concepts are relevant to FD.

3.7.3 Wikipedia Structure and Access

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, but it is still **text**. The structured parts of Wikipedia that can be reused by machines are:

- Links: wiki links, inter-language links (providing language correspondence/ overlap), inter-wiki links (referring to another wikipedia or another Wikimedia project e.g. Wiktionary, Wikibooks), external links
- Informative templates, in particular Infoboxes
- Categories (see next)
- Lists (see later)
- Portals, Projects, Tables

There are several efforts to extract structured data from Wikpedia. E.g. the Wikipedia Mining software allows extraction of focused/limited information. But we prefer to use data sets that are already structured: DBpedia (3.11) or Wikidata (3.12). We get the data in RDF and put it in Ontotext GraphDB, which allows semantic integration of the data and easier querying.

In the following 3 sections we talk of Wikipedia and illustrate with Wikipedia links, but we actually access the data from these structured sources.

3.8 Wikipedia Categories

To sift through 12.5M articles to find those relevant to FD we've adopted an iterative feedback-driven approach. This cannot be done by a single entity alone, so we adopt an iterative feedback-driven approach.

Wikipedia Categories can provide some help. They serve to organise and interlink Wikipedia:

- Each article may have a number of categories, usually placed at the bottom. We'll say that the category is a "parent" of the article, and that may be thought informally as "instance of" but very often is not
- Categories are managed as pages (just like articles), so they may also have "parent" categories. This establishes a poly-hierarchy that may be thought informally as "sub-class" but very often is not

Categories are first and foremost a navigation device: they allow the user to find more content on a topic of interest. As such, they do not follow specific distinctions between instance, subclass, part of, or other semantic relations.

Consider e.g. a category "books of X" where X is an author. In addition to being applied to each book of X, it is typically applied on the article for X himself (which does not mean he is considered to be a book). One could try to use NLP techniques and heuristics such as "if the category name includes the article name, make a specific relation rather than **instance of**".

3.8.1 Category Counts

Category statistics are obtained from DBpedia (see sec. 3.11.2).

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	1.70
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
Subtotal	**11,825,749**	**4,417,778**	**2.68**	**52,449,019**	**4.44**	**11.87**	**8,553,259**	**1.94**

Hungarian	272,323							
Romanian	256,022							
Lithuanian	168,823							
Total	**12,522,917**							

The columns are as follows:

- **articles**: number of content pages
- **cats**: number of category pages
- **art/cat**: ratio of articles to categories
- **art<cat**: number of assignments of a category as "parent" of an article
- **per art**: category assignments per article
- **per cat**: articles assigned per category
- **cat<cat**: number of assignments of a category as "parent" of another
- **per cat**: number of parent categories per category

The categorisation is extensive.

- For every 2.66 articles, there is a category. In plwiki, there are **twice** as many cats as arts, and in nlwiki the number of cats is quite poor
- Every article is assigned 4.44 cats on average. In plwiki that's 18.6 cats!
- Every category is assigned 11.87 articles on average. In nlwiki that's 29.25, so nlwiki categories have lesser discriminative power
- Every category has 1.94 parents on average, and this number is quite consistent across wikipedias. Thus the categories form a strong poly-hierarchy: in a mono-hierarchy this number is 1 and in AAT it's 1.05. 2 parents may not seem like much, but in fact it's quite extreme: a cat at level 10 (and there are deeper levels) has up to 1024 ancestors on average (some ancestor paths may converge, so the number may be smaller).

Just like articles, categories may also have inter-language links. We have not researched the degree of language overlap, but we can assume it's the same as for articles or a bit lower, so we assume 2. So for the EFD languages we have 2.2M distinct categories that categorize 2.6 distinct articles.

This number is confirmed by a recent (Dec 2014) count of Wikidata classes (across all Wikipedias) done by Ontotext:

- 2,409,399 items are instance of Q4167836 "Wikimedia category page"

3.8.2 FD Categories

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

We find the following especially curious/interesting

- Metaphors referring to food and drink
- Religious food and drink
- Food law: topics like halal, kashrut, designation of origin, religion-based ideas, fisheries laws, agricultural laws, food and drug administration, labelling regulations, etc
- Food politics
- Drink and drive songs
- Food museums

https://en.wikipedia.org/wiki/Category:Food_museums leads to over 500 museums, including food, farm, mill, agriculture, beverage, beer, tea, chocolate, and even salt museums! https://en.wikipedia.org/wiki/Category:Salt_museums alone has 15 museums!!!

3.8.3 FD Category Acquisition

https://en.wikipedia.org/wiki/Category:Food_and_drink is the root of all FD categories. (Note: Nutrition includes 27 categories besides Food and Drink, such as: Alcohol and health, Cooking, Diets, Food and drink preparation, Obesity, Prebiotics, Probiotics. We may decide to use some of those branches as well.)

The Wikipedia API[52] allows you to download categories.[53] It returns a number of formats, but none is RDF:

- xml: most readable, 1 line per result
- json: js data struct
- php=yaml: php data struct
- txt: indented text showing data struct
- dbg: indented text showing data struct
- wddx: xml similar to SPARQL results

The most important shortcoming is that the API returns only the immediate categories. Probably because of the large branching factor (2 parents on average), Wikipedia developers have consistently declined to implement "Flattening"[54] i.e. the expansion of subcategories, which would be similar to transitive inference.

So instead, we used a SPARQL repository, loaded several language instances of DBpedia (see sec. 3.11) and wrote some SPARQL queries to fetch sub-categories.

We downloaded 887660 FD-related out of 4799920 categories, or 18.5%. This is some concrete indication about the potential percentage of FD-related content in Wikipedia.

We did not transitive inference enabled in the repository, so the query took several hours. These queries won't run on the public instance of dbpedia because of time and memory limits: Virtuoso 42000 Error TN...: Exceeded 1000000000 bytes in transitive temp memory

3.8.4 Category Problems

We explored this large number of FD categories and found various problems:

Loops

Food_and_drink is connected to itself. This means that there are loops, in this case through Nutrition. We deal with them by removing redundant edges, going back towards the root "Food and drink" or across categories the same distance from the root. We call this "Category Combing".

Spillage Due to Wrong Hierarchy

By "spillage" we mean the inclusion of many irrelevant categories.

---

[52] https://www.mediawiki.org/wiki/API:Categorymembers
[53] http://en.wikipedia.org/w/api.php?action=query&list=categorymembers&format=xml&cmtitle=Category%3AFood%20and%20drink&cmprop=title%7Ctype&cmtype=page%7Csubcat&cmlimit=100
[54] https://phabricator.wikimedia.org/T3497

E.g. we noticed various football teams, and investigated this spillage. It turned out that's due to the following 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

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 sub-cats of Euthenics. But water treatment should not be a super-cat of euthenics.

You can see at DBpedia[55] the Wikipedia page revision[56] (27 October 2013) that caused the incorrect information:

```
category:Euthenics skos:broader category:Water_treatment
```

This was fixed on 12 June 2014 with comment "removed Category:Water treatment - Euthenics is not water treatment". You can verify at the up-to-the-minute DBpedia site [57] that Euthenics has only 2 parent categories: category:Social_sciences, category:Quality_of_life. Unfortunately many other similar cases are still present.

Spillage 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:

[55] http://dbpedia.org/page/Category:Euthenics
[56] http://en.wikipedia.org/wiki/Category:Euthenics?oldid=578958104
[57] http://live.dbpedia.org/page/Category:Euthenics

- Animal_dyes
- Animal_hair_products
- Animal_waste_products
- Bird_products
- Bone_products
- Coral_islands
- Coral_reefs
- Hides

Finally, there are some mixed cats that may include both relevant and irrelevant children:

- Animal_glandular_products: milk and its thousands of subcats is; castoreum is not
- Animal_fat_products
- Insect_products: honey is, silk is not
- Mollusc_products: clams and oysters are; pearls are not
- Whale_products: meat is; baleens are not (though may be)

A lot of the mixed cats don't have a second parent to indicate them as relevant.

We deal with spillage through blacklists (manual cleaning of branches) and interactive human input to cut out specific branches.

Non-Human Food/Eating

Foods and drink explicitly includes animal feeding, thus

- Not all are foods for humans, e.g. Animal_feed

The subcat Eating_behaviors:

- has some appropriate children, e.g. Diets, Eating_disorders
- has some inappropriate children, e.g. Carnivory, Detritivores

Service Categories

There are various categories that have only managerial functions and are not meaningful in terms of content. These include:

- Cuisine templates
- Drink templates
- Food and drink portals
- Food and drink templates
- * Stubs

We deal with them through black lists

3.8.5 Category Enrichment

Categories can be enriched with extra links. E.g.

- The article "Cozunak" has categories "Bulgarian cuisine" and "Christmas foods"
- If we relate "Bulgarian cuisine" to place "Bulgaria", that will connect "Cozunak" to Bulgaria for geo-searching
- If we relate "Christmas foods" to event "Christmas", that will connect "Cozunak" to Christmas for searching by event and religious festivity

This is in essence semantic enrichment and can use similar NLP techniques that we use to enrich CHOs. These links may already exist (e.g. if "Bulgarian cuisine" is one of the categories applied to "Bulgaria") but can be made more explicitly by using NLP.

3.9 Wikipedia Lists

Wikipedia lists have a similar purpose to categories, but are created and maintained differently. Each has its role[58]. Lists are article pages, including an explicit link of other articles. Because they are manually curated:

- Lists usually represent a concentration of strong content on a particular topic
- However, the content is not exhaustive

A recent (Dec 2014) Wikidata count (across all Wikipedias) shows:

- 186,119 items with type Q13406463 "Wikimedia list article"

To get an overview of this large number there a several resolutions:

- Many lists can be found by navigating from this point: https://en.wikipedia.org/wiki/List_of_lists_of_lists
- There are categories dedicated to lists, through which one can find lists on a topic of interest. E.g. https://en.wikipedia.org/wiki/Category:Food-related_lists

Unlike Wikipedia categories that are extracted in structured form in DBpedia, lists are not extracted. There are two sorts of lists (made with bullets or with line breaks), and the editorial guidelines prefer the former. It would be possible to create a relatively simple extractor for lists (or hopefully we can find an existing one). It should handle:

- multi-column lists,
- lists interspersed with headings,
- spatial references listed as second links in the same bullet (e.g. California, Bretagne)

[58] https://en.wikipedia.org/wiki/Wikipedia:Categories,_lists,_and_navigation_templates

3.9.1 FD Lists

We have explored several FD-related lists and are convinced that we should implement the list extractor mentioned above. Not only these provide excellent topical lists (e.g. Christmas foods), but we can use the harvested pages to augment the score of their categories.

First we start from **Categories** of FD-related Lists, e.g.

- Food-related lists
- Lists of foods
- Lists of beverages
- Lists about Oktoberfest
- Alcohol-related lists
- Lists of restaurants
- Bibliographies of food
- Dessert-related lists
- Lists of restaurants
- Lists of food television series episodes

Then we leverage these categories to harvest a number of lists (below is a small sampling):

- https://en.wikipedia.org/wiki/List_of_Christmas_dishes
- https://en.wikipedia.org/wiki/List_of_culinary_fruits
- https://en.wikipedia.org/wiki/List_of_culinary_herbs_and_spices
- https://en.wikipedia.org/wiki/List_of_culinary_vegetables
- https://en.wikipedia.org/wiki/List_of_edible_seeds
- https://en.wikipedia.org/wiki/List_of_festivals
- https://en.wikipedia.org/wiki/List_of_foods
- https://en.wikipedia.org/wiki/List_of_meat_animals
- https://en.wikipedia.org/wiki/List_of_micronutrients
- https://en.wikipedia.org/wiki/List_of_sandwiches
- https://en.wikipedia.org/wiki/List_of_types_of_seafood
- https://en.wikipedia.org/wiki/Lists_of_beverages
- https://en.wikipedia.org/wiki/Lists_of_festivals
- https://en.wikipedia.org/wiki/Lists_of_prepared_foods
- https://en.wikipedia.org/wiki/Lists_of_restaurants
- https://en.wikipedia.org/wiki/Wine_festival#Festivals

3.10 Wikipedia Portals and Projects

Wikipedia Portals provide an overview of a domain, and a large number of links to articles in that domain. FD-related portals include:

- https://en.wikipedia.org/wiki/Portal:Food
- https://en.wikipedia.org/wiki/Portal:Drink

- https://en.wikipedia.org/wiki/Portal:Beer
- https://en.wikipedia.org/wiki/Portal:Wine
- https://en.wikipedia.org/wiki/Portal:Agriculture_and_Agronomy
- https://en.wikipedia.org/wiki/Portal:Health_and_fitness

The Food portal includes several lists of dynamically included pages that are updated often. These include: article, picture, person, recipe, ingredient, quote.

Wikimedia Projects are concentrated efforts to accomplish some tasks and to share information. While Portals are consumer-oriented, Projects are contributor-oriented. We have found them a good source of information on FD. We have explored the following projects:

- https://en.wikipedia.org/wiki/Portal:Food/WikiProjects
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Food_and_drink
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Breakfast
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Beer
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Wine
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Food_and_drink/Beverages_Task_Force
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Food_and_drink/Herbs_and_Spices_task_force
- https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Food_and_drink/Desserts_task_force
- https://www.wikidata.org/wiki/Wikidata:WikiProject_Food

Furthermore, if EFD decides to enrich the coverage of a particular kind of data, the best way to approach it is to join one of these projects and contribute to Wikipedia or Wikidata.

Unlike the other Wikipedia information provided in the previous sections, portals and projects have little in the way of machine-readable information. They have an appropriate type in Wikidata, and that's it.

Wikipedia Navigation Templates are nicely presented link lists like this one at the bottom of the article https://en.wikipedia.org/wiki/Steak_pie:

v · t · e	British pies [hide]
Sweet	Apple pie · Bakewell tart · Banoffee pie · Bedfordshire clanger · Black bun · Custard tart · Manchester tart · Mince pie · Rhubarb pie · Treacle tart
Savoury	Bacon and egg pie · Bedfordshire clanger · Bridie · Butter pie · Chicken and mushroom pie · Corned beef pie · Cornish pasty · Cottage pie · Cumberland pie · Curry pie · Devizes pie · Fish pie · Game pie · Homity pie · Killie pie · Meat and potato pie · Melton Mowbray pork pie · Pork pie · Scotch pie · Shepherd's Pie · Squab pie · Stargazy pie · **Steak pie** · Steak and kidney pie · Steak and oyster pie · Woolton pie
Manufacturers	Clark's Pies · Dickinson & Morris · Fray Bentos · Ginsters · Higgidy · Holland's Pies · Mr Kipling · Peter's · Poole's · Pork Farms · Pukka Pies · Shire Foods · Square Pie · Wall's · Wrights Pies

We don't yet know whether these links are easily available in structured form.

3.11 DBPedia

DBpedia is a knowledge base extracted from Wikipedia that has been in development for 7 years. Since DBpedia provides stable URLs for all kinds of entities (URLs come from Wikipedia), it is the centre of the LOD cloud and is linked from 207

other LOD datasets[59]. DBpedia relies on an elaborate Extraction Framework that is described in [Lehmann 2013], together with detailed statistics.

A number of general extractors obtain all kinds of useful information from the page:

- The structured info described in 3.7.3
- Abstract (text before the first heading)
- Numbers, dates, times (a highly non-trivial problem given the variety of languages and editorial practices)
- Every national-language property from referenced templates
- Etc etc

In the next step the extractor looks for Infobox templates mapped to a class.[60] The mapping is performed in a crowd-sourced fashion using a wiki and templates such as: TemplateMapping, PropertyMapping, ConditonalMapping, etc. The DBpedia ontology is also maintained there. Many of the Wikipedia articles don't have such a template, or are not yet mapped. E.g.

- Aristotle is a NE (Infobox **philosopher**) but Mathematics is not (no Infobox for **science**)
- Strawberry and Chocolate (Infobox **film**) is NE (a creative work), but Michelin Guide is not (there is Infobox **book** but it has not been used here)
- Chocolate is a NE (Infobox **prepared food**)
- Taco Bell is a NE (Infobox **company**) but Restaurant is not
- Stella Artois is a NE (Infobox **beverage**) but Drinking Culture or Beer Styles is not
- Nyangatom_people (the Culture used in sec. 5.6) is not. Not sure whether there is an infobox for culture, but it has not been used here

The accurate extraction of types and values and the quality of the DBpedia ontology depends on the mapping editors. It is an open effort, so anyone can join and add the mappings he is interested in. Ontotext has been very active in this effort since Dec 2014, in particular regarding mapping consistency and additions for BG DBpedia. We attended the DBpedia meeting in Dublin (Feb 2015)[61] and will be part of the DBPedia Ontology Committee. See for example:

- DBpedia Ontology and Mapping Problems[62]
- Adding a DBpedia Mapping[63]
- bg.dbpedia.org launched[64]

[59] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/
[60] http://mappings.dbpedia.org
[61] http://wiki.dbpedia.org/meetings/Dublin2015
[62] http://vladimiralexiev.github.io/pres/20150209-dbpedia/dbpedia-problems-long.html
[63] http://vladimiralexiev.github.io/pres/20150209-dbpedia/add-mapping-long.html
[64] http://vladimiralexiev.github.io/pres/20150209-dbpedia/bg-dbpedia-launched.html

3.11.1 Food Class in DBpedia

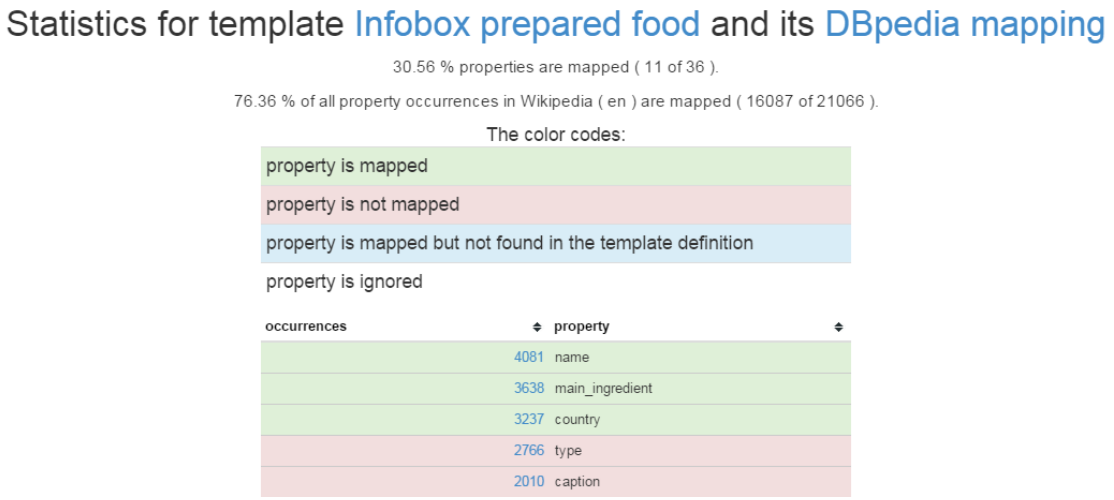
DBpedia includes a class Food, which currently has 6,606 items on EN DBpedia (http://dbpedia.org/sparql)

```
select * {?x a dbpedia-owl:Food}
```

It includes:

- Foods, e.g. Brown_bread, Amandine_(dessert)
- Food ingredients, e.g. Brine_(food), Cranberry_sauce
- Brands, e.g. Hangar_1_Vodka
- Dishes/recipes, e.g. Hasty_pudding

They come from templates like[65] Infobox prepared food, Food-product-stub, Food-stub, Meat-stub, Korean cuisine, etc. Exploring the first template, we can see which properties are mapped and which are not yet mapped[66]



The class hierarchy[67] regarding Food is still quite modest:

- Food (edit)
  - Beverage (edit)
    - Beer (edit)
    - Vodka (edit)
    - Wine (edit)
      - ControlledDesignationOfOriginWine (edit)
  - Cheese (edit)

[65] http://mappings.dbpedia.org/index.php?title=Special%3ASearch&search=food&go=Go
[66] http://mappings.dbpedia.org/server/templatestatistics/en/?template=Infobox_prepared_food
[67] http://mappings.dbpedia.org/server/ontology/classes/#Food

Many foods in Wikipedia don't have infoboxes (e.g. Aioli). All these are reasons why the number of items in class Food is quite small (we estimate that Wikipedias have at least 100k foods).

As explained in the previous section, we can easily edit the ontology and mappings to improve them. But even without that, DBpedia offers easy access to some very important info: the categories.

3.11.2 Counting Categories

Categories and the links between them are represented as in the following query patterns. We can count categories and assignments between articles and categories

```
select (count(*) as ?c) {?x a skos:Concept}     # cat
select (count(*) as ?c) {?x dcterms:subject ?y} # art<cat assignment
select (count(*) as ?c) {?x skos:broader ?y}    # cat<cat assignment
select (count(*) as ?c) {?x skos:subject ?y}    # art<cat topical assignment
```

The topical article of a category is the main article describing it, e.g.:

```
dbcat:Programming_languages skos:subject dbr:Programming_language
```

Note: this should use foaf:focus instead of skos:subject, see this bug[68].

The statistics are shown in sec. 3.8. We ran the queries at http://dbpedia.org, http://nl.dbpedia.org, http://fr.dbpedia.org, http://it.dbpedia.org, http://es.dbpedia.org, http://pl.dbpedia.org, and http://bg.dbpedia.org/sparql[69]. The other dbpedias (hu, ro, lt) are not yet live, so we don't have their numbers. It is possible to:

- Use SPARQL Federated queries from a single point. But not all dbpedia servers support federation, and not all in the same way.
- Load all datasets in one repository, and run just one query.

Loading all 11 dbpedias in Ontotext GraphDB is what we plan to do for EFD. See next section.

3.11.3 owl:sameAs and Smushing

A crucial feature of Wikipedia pages are the inter-language links, which establish that the same thing is described with several pages (see sec. 3.7.1). This applies to both articles and categories. The category networks in two Wikipedias are not the same, but the inter-language links establish points of correspondence between the networks.

These links are represented in DBpedia as owl:sameAs. The semantics of owl:sameAs is one of "smushing": the two nodes are merged, and all their statements

[68] https://github.com/dbpedia/extraction-framework/issues/301
[69] Established Jan 2015 by Ontotext

(incoming and outgoing) are applied to that merged node. In fact many nodes can be declared sameAs, in which case the smushing applies to the whole cluster (e.g. sec. 3.7 mentions 700 famous people that are described in more than 20 Wikipedias). Ontotext GraphDB allows efficient handling of sameAs[70] so it is feasible to load the DBpedia datasets in all 11 EFD languages and enforce the sameAs.

Al DBpedia2014 downloads are provided on a download server[71] and documented at the dbpedia wiki[72]. Each language has about 160 files, and it takes some careful analysis to decide which of them should be loaded. A few of the interesting files are not documented, but should still be downloaded.

This will enable cross-leveraging of article data and category networks. Most importantly, it will join up the labels from all DBpedias, enabling strong multilingual access. This is especially important for EFD content that won't be translated to English, because it will allow semantic enrichment in the native language to connect to items in the merged DBpedia.

However, it will also exacerbate data quality problems, since a mistake in one DBpedia will be proliferated to the merged dataset.

3.12 Wikidata

Wikidata is a crowd-sourced knowledge base similar to FreeBase (to be retired/merged to Wikidata) or the Google Knowledge Graph. Its purpose is to manage Wikipedia data centrally and provide it to Wikipedia templates on demand. It has been seeded from Wikipedia: there is a Wikidata item for every article and category page, each item has multilingual labels and aliases, some class assignment ("instance of"), and some have descriptions. Most people (humans) have life dates, places, occupations. Some areas are very well developed (e.g. taxons).

The Wikidata data model is flexible and allows any number of references and qualifiers to be added to a statement (called claim in Wikidata). Excellent editorial facilities are provided, e.g. autocomplete searches through all names of an item or property, and the most popular ones are shown first.

All inter-language links are now managed in Wikidata, replacing a quadratic number of textual links. Some wikipedias are starting to externalize more of their data to Wikidata (e.g. itwiki, bgwiki).

Wikidata makes RDF dumps regularly. These are described in [Erxleben 2014] and include:

- a simple variant (statements without qualifiers)
- a full variant (uses non-standard reification patterns to express the qualifiers)

---

[70] https://confluence.ontotext.com/display/GraphDB6/GraphDB-SE+Reasoner#GraphDB-SEReasoner-sameAsOptimisation
[71] http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/
[72] http://wiki.dbpedia.org/Downloads2014

The main benefits of Wikidata are:

- Data consistency, due to careful import processes and human data editing
- Good labels that are easily available (don't have to load 11 data sets)
- Good direct types (but the class hierarchy is a mess)
- Strong coreferencing to other datasets (e.g. VIAF), see sec. 2.3

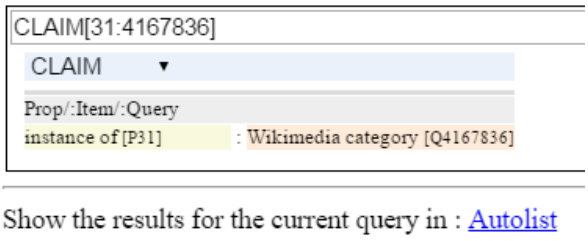
The main disadvantage of Wikidata is the still-limited data scope. For example:

- Categories are marked appropriately, but article<category assignments are not available
- Companies are marked appropriately, but web homepages are not available

3.12.1 Counting Categories

Wikipedia categories are available in Wikidata. Eg https://www.wikidata.org/wiki/Q8949729 is "Category:Christmas food". It includes multilingual labels, but does **not** include parent cats, nor articles using the cat.

Wikidata provides a weird, wonderful and powerful query language called WDQ. Let's count categories. First go to the WDQ query editor[73] to "click a query"



CLAIM[31:14827288] means "instance of: Wikimedia category".

Now you can either:

- Click the Autolist[74] link that returns a page of items and the total count. You can also intersect items of interest with particular categories. If you're logged in, you can even do batch updates.
- Make an API[75] call, adding an appropriate parameter to count without returning items[76]

[73] http://wdq.wmflabs.org/wdq/
[74] http://tools.wmflabs.org/autolist/autolist1.html?q=CLAIM%5B31%3A14827288%5D
[75] https://wdq.wmflabs.org/api_documentation.html
[76] https://wdq.wmflabs.org/api?q=CLAIM[31:14827288]&noitems=1

3.12.2 Wikidata FD Hierarchies

Wikidata includes some number of classes related to FD, and items of these types

- Visual Food Hierarchy[77]: 2050 items as of 22-Jan-2015
- Visual Drinks Hierarchy[78]: 184 items as of 22-Jan-2015

This is even less than the DBpedia food class. Nevertheless, the information is very well collated. It's best visualised with the Reasonator application, e.g. for Cheese[79].

- "Classification" is the class hierarchy in Wikidata (that's not related to Wikipedia categories)
- "From related items" shows inverse relations, e.g.
- Antoine Roussel is a fromager, i.e. has worked with cheese
- There are 259 subclasses of Cheese
- On the right is correlation to external Authority Files and links to wikipedias and related Wikimedia projects

[77] http://tools.wmflabs.org/wikidata-todo/tree.html?q=2095&rp=279
[78] http://tools.wmflabs.org/wikidata-todo/tree.html?q=40050&rp=279
[79] https://tools.wmflabs.org/reasonator/?&q=10943

2.1 Europeana Food and Drink Classification Scheme

Reasonator

Random item | English | Find | Other

cheese (Q10943)

奶酪 | 芝士 | Zachary Caruso | The Cheese | 起士 | 起司

generic term for a diverse group of milk-based food products

subclass of	dairy product food produced from the milk of mammals
instance of	dairy product food produced from the milk of mammals
topic's main category	Category:Cheese Wikimedia category page

Classification

Name	Description
Entity	something that exists
Object	mental construct defined by the perception of something as a unitary physical or abstract entity
Abstract object	classifications that denote whether a term describes an object with a physical referent or one with no physical referents
Identifier	codification of something that exists by a subject
Information	sequence of symbols that can be interpreted as a message by a subject
Work	distinct intellectual or artistic creation
Product	anything that can be offered to a market that might satisfy a want or need
Food	any substance consumed to provide nutritional support for the body
Food product	Lebensmittel, die vorwiegend der Ernährung des Menschen dienen
Dairy product	food produced from the milk of mammals
Cheese	generic term for a diverse group of milk-based food products

The above section uses data from WikiData Query, which can be ~10min out of date. You can switch to the slower live version

From related items

field of work	Henri Boursault crémier au Perreux sur Marne connu comme l'inventeur du Boursault
	Antoine Roussel fromager
depicts	Man with a glass of wine painting
	Nature morte (lièvre, canard, bouteilles, pain et fromage) painting by Jean-Baptiste Oudry
	Still life with cheese, artichoke and cherries painting by Clara Peeters
material used	Käsebrot Brot mit einem Belag aus Käse
	cheeseburger hamburger topped with cheese
subclass of	259 items. Show items
is a list of	list of French cheeses Wikimedia-Liste country : France [AL]
part of	crust outer layer of some cheeses
has part	Käsespätzle
	Cholera Gemüsekuchen mit Lauch, Kartoffeln, Käse und Äpfeln aus dem Kanton Wallis in der Schweiz

Related media

Commons category : Cheese
Commons gallery : Cheese

External sources	
BNCF	3086
Commons category	Cheese
Commons gallery	Cheese
GND	4029176-5
Freebase	/m/01nkt
NDL	00573131

Wikimedia projects

Current language Wikipedias	
en	Cheese
Big Wikipedias	
de	Käse
es	Queso
fi	Juusto
fr	Fromage
hu	Sajt
it	Formaggio
ja	チーズ
nl	Kaas
pl	Ser
pt	Queijo
ru	Сыр
sv	Ost
Wikimedia Commons	
commons	Cheese
Wikiquote	
ca	Formatge
cs	Sýr
en	Cheese
es	Queso
it	Formaggio
nn	Ost
pl	Ser
sk	Syr
sv	Ost
Other Wikipedias	
af	Kaas
als	Käse
an	Queso
ang	Cīese
ar	جبن
arc	ܓܒܢܐ
arz	جبنه
ast	Quesu
ay	Kisu
az	Pendir
ba	Сыр
bar	Kaas
bat_smg	Sūris
be	Сыр
be_x_old	Сыр
bg	Сирене
bn	চিজ
bo	རྫས་
br	Keuz (boued)
bs	Sir
bxr	Бисалар
ca	Formatge
ce	Нехча
chr	ᎤᏁᏉᎢ ᏭᏫᏁ
ckb	پەنیر
cs	Sýr
cv	Чăкăт
cy	Caws
da	Ost
el	Τυρί
eml	Furmâj
eo	Fromaĝo
et	Juust
eu	Gazta
fa	پنیر
fiu_vro	Juust
fur	Formadi
fy	Tsiis
ga	Cáis
gan	奶酪

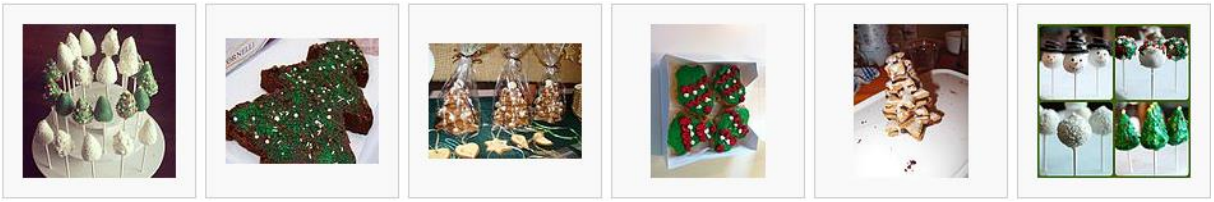
3.13 Wikimedia Categories

Wikimedia Commons is a multimedia library shared between all Wikipedias. It's organized similar to Wikipedia, with individual pages and categories.

E.g. the page https://commons.wikimedia.org/wiki/Cheese has a number of sections on serving cheese, types of cheese, etc (for a painter this would be his paintings, organized by genre, year etc)

The category https://commons.wikimedia.org/wiki/Category:Cheese lists correspondences to Wikipedia categories, and 40 sub-cats. Notable sub-cats include Pronunciation of cheese (9 audio files), Wiki loves cheese (181 pictures of cheese).

Wikimedia has its own set of categories, not necessarily parallel with the Wikipedia categories. E.g. https://commons.wikimedia.org/wiki/Category:Christmas_tree-shaped_food exists in Wikimedia but not in Wikipedia.



The two hierarchies are aligned to a large extent.

Wikimedia is extracted in DBpedia 2014 together with all the Wikipedias.

3.14 Wiktionary

Wiktionary is a collaboratively developed multilingual dictionary. Actually it's a set of dictionaries. E.g. the English Wiktionary includes words from all kinds of languages, but the definitions and explanations are in English. It has the same basic structure as Wikipedia.

Unsurprisingly, it has words and categories related to FD, e.g.

- https://en.wiktionary.org/wiki/Category:Foods
- https://en.wiktionary.org/wiki/Category:Beverages

As you can see, the category structure is bifurcated and breaks down into:

- Language-independent sub-categories (e.g. beverages> alcoholic, coffee, milk, tea).

- Language-dependent categories that are semantically on the same level (e.g. beverages> bg:Beverages, en:Beverages, etc)

The language-dependent categories further break-down into sub-categories, e.g.:

- https://en.wiktionary.org/wiki/Category:en:Beverages
- https://en.wiktionary.org/wiki/Category:en:Alcoholic_beverages

Correspondence between words in the same dictionary (i.e. translations) and to other dictionaries is provided, e.g. for airan#English[80] (redlinks are entries not yet written)

[80] https://en.wiktionary.org/wiki/airan#English

2.1 Europeana Food and Drink Classification Scheme

Etymology [edit]

From Turkish *ayran*.[1]

Pronunciation [edit]

[eboʃ] This entry needs pronunciation information. If you are familiar with enPR or the IPA then please add some!

Noun [edit]

airan (*plural* **airans**)

1. A Turkish and Altaic drink out of yoghurt, water and salt.

Translations [edit]

±a Turkish and Altaic yoghurt drink

Select targeted languages

- Arabic: عيران (ʿayrān), شنينة (šinīna)
- Armenian: այրան (ayran)
- Azeri: ayran (az)
- Bashkir: айран (ayran)
- Georgian: აირანი (airani)
- Greek: αριάνι (ariáni), αϊράνι (aïráni)
- Kazakh: айран (ayran), шалап (şalap)
- Kurdish:
  - Kurmanji: dew (ku), çeqilmast (ku)
  - Sorani: دۆ (do)
- Kyrgyz: чалап (çalap)
- Macedonian: ајран (ajran)
- Persian: دوغ (duğ)
- Polish: ajran
- Portuguese: ayran *m*
- Russian: айра́н (ru) *m* (ajrán)
- Spanish: ayran *m*
- Tajik: айрон (ayron), дӯғ (düğ)
- Tatar: әйрән (tt) (äyrän)
- Turkish: ayran (tr)
- Uzbek: ayron (uz), çalop

Wiktionary may be even a better source for the EFD classification than Wikipedia, since it has an entry for every word, whereas Wikipedia does not have separate entries for words representing "trivial" concepts.

However, the coverage of Wiktionary is still low compared to Wikipedia. E.g:

- Wiktionary does not have an entry for Kacamak as food. https://en.wiktionary.org/wiki/kaçamak is "a loophole" (Turkish)
- The word Качамак (in Cyrillic) does not have any entry, not even in https://bg.wiktionary.org
- In contrast, https://en.wikipedia.org/wiki/Kačamak and 6 more wikipedias (bg, es, fr, sr, tr, uk) describe the food

We have not yet analysed the structure of Wiktionary categories.

The structure of Wiktionary pages is totally different from Wikipedia pages, and is more regular. Wiktionary is available as RDF using the Lemon ontologies. They are part of the upcoming **Linguistic Linked Data** cloud, together with NIF, OLIA, etc. See a brief introduction.[81]

[81] http://vladimiralexiev.github.io/Multisensor/20141008-Linguistic-LD

3.15 Wordnet Domains in Yago2

WordNet[82] is a major linguistic resource that has stimulated the development of similar resources in many other languages, their inter-connection into a grid, and integrations with Wikipedia such as BabelNet. It deals with Synsets (synonym sets) and word forms that express their meanings.

WordNet includes a field called lexical file info that groups senses into topics or "domains"

- Eg for **Aioli** and other sauces it's <noun.food>
- But for **cooking** it's <verb.creation> or <verb.change>, i.e. not a semantic domain

WordNet Domains[83] is a novel dataset that provides 200 hierarchical[84] domains. It is integrated in Yago2[85] and available for download[86] as RDF (file yagoWordnetDomains.ttl).

- There are 68862 synsets (the full WordNet is about 300k)
- 97% of them have a domain (67k). The average is 1.3 domains per synset (total 87k)
- There are 167 unique domains, the biggest one is "factotum" (21%)

Domain assignment is not perfect, but the precision is quite high. E.g. this is a mistake:

```
<wordnet_food_bank_113368900>    <hasWordnetDomain>  <wordnetDomain_money>
```

Proper assignment depends on the sense. E.g. one can be puzzled by this:

```
<wordnet_mother_115106674> <hasWordnetDomain>  <wordnetDomain_gastronomy>
```

But the sense referred to is **mother**: "a stringy slimy substance consisting of yeast cells and bacteria; forms during fermentation and is added to cider or wine to produce vinegar". This is an example that the domain is useful for Word-Sense Disambiguation, i.e. during semantic enrichment.

According to the original site, the domains relevant to FD are "alimentation" and "gastronomy". However, we discovered a wider selection of domains (listed in decreasing order of FD relevance):

- food, gastronomy, agriculture, animal_husbandry, fishing, hunting
- home

---

[82] http://wordnetweb.princeton.edu
[83] http://wndomains.fbk.eu/
[84] http://wndomains.fbk.eu/hierarchy.html
[85] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
[86] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/

- animals, plants, biology, commerce, economy

By searching for keywords " food, gastronomy, agriculture, animal_husbandry, fishing, hunting", we discovered 43 more FD-relevant synsets scattered around 20 other domains, e.g.

```
<wordnet_fishing_eagle_101615703> <hasWordnetDomain> <wordnetDomain_animals>
<wordnet_hunting_dog_102087122> <hasWordnetDomain> <wordnetDomain_animals>
<wordnet_food_fish_102512938> <hasWordnetDomain> <wordnetDomain_animals>
<wordnet_fishing_boat_103351262> <hasWordnetDomain> <wordnetDomain_nautical>
<wordnet_food_processor_103378174> <hasWordnetDomain> <wordnetDomain_home>
<wordnet_food_hamper_103378342> <hasWordnetDomain> <wordnetDomain_factotum>
<wordnet_hunting_knife_103549473> <hasWordnetDomain> <wordnetDomain_factotum>
<wordnet_food_105811214> <hasWordnetDomain> <wordnetDomain_factotum>
<wordnet_fishing_license_106550381> <hasWordnetDomain> <wordnetDomain_law>
<wordnet_hunting_license_106550552> <hasWordnetDomain> <wordnetDomain_law>
<wordnet_food_107555863> <hasWordnetDomain> <wordnetDomain_chemistry>
<wordnet_gastronomy_107572206> <hasWordnetDomain> <wordnetDomain_tourism>
<wordnet_fresh_food_107572712> <hasWordnetDomain> <wordnetDomain_chemistry>
<wordnet_convenience_food_107593549> <hasWordnetDomain> <wordnetDomain_chemistry>
<wordnet_devil's_food_107631109> <hasWordnetDomain> <wordnetDomain_chemistry>
```

Overall, we produced 2 files with the following number of concepts:

- 2711 yagoWordnetFood.ttl
- 3178 yagoWordnetFoodPlus.ttl

The latter has more concepts at the expense of lower precision. E.g. out of the examples below, we consider the first 3 relevant but the last 2 not relevant:

```
<wordnet_topiary_104454908>      <hasWordnetDomain>  <wordnetDomain_agriculture> .
<wordnet_trammel_net_104469345>  <hasWordnetDomain>  <wordnetDomain_fishing> .
<wordnet_vineyard_104536038>     <hasWordnetDomain>  <wordnetDomain_agriculture> .
<wordnet_pink_disease_114281678> <hasWordnetDomain>  <wordnetDomain_agriculture> .
<wordnet_potato_wart_114281851>  <hasWordnetDomain>  <wordnetDomain_agriculture> .
```

The final result includes the following domains and sample synsets:

- foods (Aioli), techniques (cooking, boiling), utensils (bread knife, plate, pot), taste (pungency, spiciness), other (kosher)

An search for drink, beer, wine, alcohol, whisky, vodka produces 119 concepts.

- Beer & wine varieties are in both food & chemistry, so are included above

This investigation can be extended. E.g. these items are relevant:

- Beer mugs: domain "factotum"
- Beer cans: domain "furniture" (this is a mistake)
- Beer gardens: domain "buildings"

3.16 Types Over Wikipedia

As already mentioned, classifying Wikipedia articles and/or making sense of the Wikipedia category hierarchy is a hard task. There are a number of approaches that attempt classification, i.e. deriving types for Wikipedia articles. Some of them use the categories, others use NLP approaches, e.g. by parsing the definition (first sentence) of each article, usually in some combination

- AirPedia[87] [Aprosio 2013]
- DBtax[88] (Fossatti)[89] *
  - Just released with the it.dbpedia.org dataset
- Heuristic Types[90] [Paulheim 2013]
  - Integrated in the DBpedia 2014 dataset
- LinkedHypernyms[91] [Kliegr 2014]
- UMBEL[92] [Giasson 2015] *
  - Integrated in DBpedia for several years
- WiBi[93] [Flati 2014] *
- Yago2 [de Melo 2010], [Biega 2013], [Mahdisoltani 2015]
  - Integrated in DBpedia for several years. Includes very specific classes (e.g. for Angela Merkel: FemaleHeadsOfGovernment, GermanPeopleOfPolishDescent, GermanWomenInPolitics, MinistersForChildren,YoungPeopleAndFamilies) derived from Wordnet. But not quite developed for the FD domain

We evaluate the FD relevance of the ones marked with * in the following sections

3.17 DBtax

DBtax is a brand new dataset. The FD part[94] covers:

- 5792 foods and drinks in one class (dbtax:Food)
- Discussing with the authors the possibility of splitting Drinks from Foods
- Includes some food manufacturers, eg: Yupi_(confectioner), an Indonesian gummy jelly manufacturer
- Includes food brands, eg: Trolli, a German gummie brand; Treet, a canned meat product similar to Spam; Treets or Knabbelknikkers, a confectionery brand
- Includes food micro-ingredients like Lactobacillus_kimchii, a major ingredient in Korean kimchi
- Includes botanical names often used as foods, eg
  - Craterellus_cornucopioides, Laccocephalum_mylittae and Lactarius_fragilis: cooking Mushrooms

---

[87] http://www.airpedia.org/download/dbpedia-entity-types-in-31-languages/
[88] http://it.dbpedia.org/downloads/dbtax/
[89] http://it.dbpedia.org/2015/02/dbpedia-italiana-release-3-4-wikidata-e-dbtax/?lang=en
[90] http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/en/instance_types_heuristic_en.nt.bz2
[91] http://ner.vse.cz/datasets/linkedhypernyms/ , http://boa.lmcloud.vse.cz/LHD/en.LHDv1.draft.nt.gz
[92] https://github.com/structureddynamics/UMBEL
[93] http://wibitaxonomy.org/download.jsp
[94] http://it.dbpedia.org/downloads/dbtax/A-Box/Food.ttl

- Triticeae: a botanical tribe that includes major crop genera: wheat, barley, and rye.

Precision:

- An evaluation of 100 items couldn't find any that are not related to Food.
- The inclusion of Triticeae is perhaps wrong, since it itself is not a food, but the crops classified under it are food ingredients

Recall:

- We have not performed extensive recall evaluation, but judging from the number of items the recall is under 25%
- Includes eg Aioli (mayonnaise sauce), Air_chocolate, Zest_(ingredient), Kyopolou (vegetable paste), Cozonac (sweet bread), Kyselo (mushroom soup), etc etc etc
- Omits eg Shallot (an onion), Chipotle (dried chilli), Lactobacillus_delbrueckii_subsp._bulgaricus (a major ingredient in Bulgarian yogurt), etc, etc

3.18 WiBi (Wikipedia BiTaxonomy)

WiBi (Wikipedia Bitaxonomy) is an approach to the automatic creation of two hierarchies for Wikipedia:

- Page hierarchy, based on the definition (first sentence) of each page
- Category hierarchy

The two hierarchies reinforce each other and are built in an iterative fashion until a fix-point is reached. The WiBi downloads[95] include important comparisons to DBpedia, MENTA/Yago and WikiNet (see directory "datasets"), showing that WiBi is best-in-class. WiBi contains the following distribution files:

file	statements	description
categorytaxonomy	594,933	Category Taxonomy: a tree made by selecting one parent per category
lemmataxonomy	4,270,248	Page Lemmas: the main word characterizing a page (e.g. "food", "beer", "brand")
pagetaxonomy	3,859,733	Page Taxonomy: after disambiguation of the lemmas, using a semantic step and BiTaxonomy Algorithm

We have not performed a comprehensive evaluation of the utility of WiBi for EFD, especially the refined category taxonomy. But below we provide an experimental/anecdotal evaluation.

[95] http://wibitaxonomy.org/wibi-ver1.0.tar.gz

WiBi Category Taxonomy

- WiBi selects one parent category for each category, e.g. "People in food and agriculture occupations < People by occupation". The other parent "Food and drink" is suppressed
- This makes a proper taxonomic tree: all pages classified under "People in food and agriculture occupations" can be legitimately classified under "People by occupation"
- However, it loses the link to our topic of interest (FD)
- Furthermore, WiBi loses some legitimate sub-categories. For example, in WiBi the only sub-categories of "People in food and agriculture occupations"[96] are
  - Gastronomy occupation
  - Maltsters
  - Winemakers
  - Wintners
- But it has 30 subcategories, of which only "Food and drink biography stubs" is not an appropriate subclass of people. Eg:
  - Beekeepers
  - Baristas
  - Food engineers (!?!)
  - Food scientists
  - Food writers
  - Wine merchants

WiBi leaves only the following sub-categories of the root Food and drink

WiBi Categories	**Missing Categories**
Beverages	Agriculture
Ceremonial food and drink	Animal products
Food and drink appreciation	Aphrodisiacs
Food and drink by country	Beverages
Food and drink media (1)	Cuisine
Food and drink preparation	Eating behaviors
Food and drink terminology	Famines
Food and the environment	Food-related lists
Food culture	Food allergies
Food decorations	Food and drink literary awards
Food safety	Food and drink portals (2)
Foods	Food and drink preparation
Foods named after people	Food and drink templates (2)
Foodservice	Food awards
History of food and drink	Food festivals
Meals	Food law
Nutrition	Food museums

[96] https://en.wikipedia.org/wiki/Category:People_in_food_and_agriculture_occupations

2.1 Europeana Food and Drink Classification Scheme

WiBi Categories	Missing Categories
	Food politics
	Food riots
	Gustation
	Hunger
	Metaphors referring to food and drink
	Organic food
	People in food and agriculture occupations
	Serving and dining
	Works about food and drink

Notes:

1. Only in WiBi
2. Not useable as topical caetgory
3. WiBi is generated in Jun1 2014. We don't have information which Wikipedia dump it used
4. The right column uses data extracted Aug 2014.

WiBi Page Lemmas

A "lemma" or "hypernym lemma" is the "main word" characterizing a page. We've investigated lemmas related to "beer".

page	description	WiBi	?
Beer	an alcoholic beverage	beverage	OK
Draught beer	beer served from a cask or keg rather than from a bottle or can	beer	OK
Ice beer	marketing term for pale lager beer brands which have undergone some degree of fractional freezing	term	NOK
Fix (beer)	a brand of Greek lager beer by the FIX brewery	brand	maybe
Victoria Bitter	a lager produced by Carlton & United Breweries	beer	OK

We see some inconsistency in WiBi:

- Draught beer and Ice beer both refer to processes/techniques ("Draught" is a method of serving and "Ice" is a method of production). The former's lemma is "beer" (OK) but the latter is "term" (NOK since it's too generic)
- Fix and Victoria Bitter are both brands of beer. But the former's lemma is "brand" while the latter is "beer"

WiBi Page Taxonomy

page	description	WiBi	?
Beer and food matching	Now redirects to https://en.wikipedia.org/wiki/Beer#Beer_and_society, referring to " beer sommelier, who informs restaurant patrons about beers and food pairings"	Beer	NOK
Delta Corporation	a beer and soft drink company of Zimbabwe	Beer	NOK
Konig Brauerei	Brewery situated in Duisburg	Beer	NOK
Beer style	categorize beers by factors such as colour, flavour, strength, ingredients, production method, recipe, history, or origin	Beer	NOK

- These mistakes are due to the linguistic (NLP) approach used by WiBi to extract a page hierarchy from the first sentence of each page (its description or definition)
- E.g. the description of "Konig Brauerei" is "The **König Brewery** is situated in the Beeck area of Duisburg; amongst other beers, it brews…".
- Since the page name includes the term Brewery, the Wikipedia editors have not explicitly said "is a brewery".
- WiBi uses dependency parsing to find the copula "is", then looks for a main noun phrase to the right and finds "other **beers**"

Comparing Lemma vs Page Taxonomy Coverage

- WiBi.lemmataxonomy.ver1.0.txt has 168 items with lemma "beer"
- WiBi.pagetaxonomy.ver1.0.txt has 139 items classified as "Beer"
- As you see, the page taxonomy has thrown out many "beer" lemma candidates, most of them legitimate, eg:
  - Crown Lager
  - Crown Pilsner

WiBi Conclusion

Our conclusion is that WiBi is not directly usable for EFD due to:

- the removal of **topical** category links towards the topic of interest (WiBi leaves only taxonomical links)
- a large number of omissions
- a smaller number of incorrect classifications

However, we'd like to continue this investigation, in order to leverage some of the WiBi knowledge for cutting off branches of the category network that are inappropriate for our topic of interest

3.19 UMBEL

UMBEL is a selection of about 10% of OpenCyc, to be used as a mapping layer by other ontologies. UMBEL uses some specific terminology:

- UMBEL Reference Concepts (RC) are the main UMBEL items (there are about 35k)
- Super-types are top-level groupings in UMBEL (about 30)
- UMBEL uses supertype-specific relations to express the mappings (e.g. umbel:relatesToFoodDrink links to a RC in the FoodDrink super-type)

We looked at release 1.01, Sep 2014[97]. The UMBEL ontology and reference structure[98] was updated Sep 2014. We also used the DBPedia→UMBEL mapping files,[99] which are from Feb 2011 (4 years old). We provide count by unique subjects, unique objects, and triples (statements). The total number of mapping triples is 5422366. The object is always a RC, denoted by the prefix rc:

file	triples	subj	nsubj	prop	nobj	coverage
cycPages	16034	dbr:	15697	umbel:correspondsTo	16032	46% of rc
dbpediaOntologyPages	896423	dbr:	659527	rdf:type	184	15% of dbr
wikipediaCategories	2987	dbr:Category:	2984	umbel:correspondsTo	2789	10% of cat, 10% of rc
wikipediaCategories-ListsIndividuals	4395002	dbr:	1808782	umbel:relatesTo*	1668	50% of dbr
wikipediaCategories-SVIndividuals	111920	dbr:	102956	umbel:relatesTo*	2484	1.3% of dbr

Below we provide a sampling and an explanation of each file

3.19.1 cycPages

Individuals to the concepts they (nearly) represent

```
dbr:Copper_extraction umbel:correspondsTo rc:CopperOre .
dbr:Downhill umbel:correspondsTo rc:DownhillSkiing .
dbr:N umbel:correspondsTo rc:LatinCapitalLetterN .
dbr: umbel:correspondsTo rc:Guerrero_StateMexico .
dbr:Raduga_KSR-2 umbel:correspondsTo rc:CruiseMissile_AS5Kelt .
dbr:Marker_pen umbel:correspondsTo rc:Marker_WritingImplement .
dbr:Grizzly_bear umbel:correspondsTo rc:GrizzlyBear .
```

This establishes a relatively straightforward correspondence from Wikipedia pages (DBpedia individuals) to UMBEL RC. It covers half of RC. Disambiguation is performed (e.g. https://en.wikipedia.org/wiki/Downhill means the skiing discipline). There are some errors (e.g. the middle line above states that the namespace http://dbpedia.org/resource/ correspondsTo rc:Guerrero_StateMexico, which is false).

3.19.2 dbpediaOntologyPages

Dbpedia type assignment to RC

[97] http://www.mkbergman.com/1795/umbel-version-1-10-released/
[98] https://github.com/structureddynamics/UMBEL/tree/master/Reference%20Structure
[99] https://github.com/structureddynamics/UMBEL/tree/master/Named%20Entities/Wikipedia

```
dbr:Jim_Covert a rc:Athlete .
dbr:Revision3 a rc:Organization .
dbr:Jabiru_Aircraft a rc:Organization .
dbr:Yoho_National_Park a rc:Location_Underspecified .
dbr:Trigoniaceae a rc:FloweringPlant .
dbr:KXKS_(AM) a rc:RadioStation_Organization .
dbr:Eastview,_Tennessee a rc:Town .
```

Establishes a type assignment for 15% of DBpedia individuals (all of them are Named Entities, not abstract concepts). Uses 184 RC as classes, which is only 0.5%. For comparison, DBpedia has about 650 classes. The type assignment is very accurate. In a random sample of 100 assignments, we found only these errors:

```
dbr:Swiss_American a rc:CommunityOrganization . # generic group, not an org
dbr:Sheffield_F.C. a rc:Organization . # more specifically rc:Club_Organization
```

It even makes a distinction between Automobile and Engine (the latter page has unfortunately been deleted).

```
dbr:Mitsubishi_Lancer__Sixth_generation a rc:Automobile .
dbr:Toyota_Celica__Third_generation__2.0_L_I4_21R a rc:AutoEngine .
```

We performed a frequency analysis. The top types (out of 184) are as follows (we've highlighted some lines that may be relevant to FD). There are some weird cut-offs for 20k and 10k (such round numbers are not likely to occur naturally) that we've reported as a bug[100]

count	type	Note
31031	rc:Athlete	
20000	rc:Club_Organization	
20000	rc:PopulatedPlace	
20000	rc:Village	
17128	rc:Building	
16775	rc:Business	
11895	rc:Artist	
11779	rc:Event	
11621	rc:Actor	
11513	rc:Place	
10901	rc:Politician	
10586	rc:MusicalComposition	
10541	rc:PersonWithOccupation	
10164	rc:BiologicalLivingObject	
10000	rc:Album_IBO	
10000	*rc:Animal*	
10000	rc:AutoEngine	
10000	rc:Band_MusicGroup	
10000	rc:BaseballPlayer	
10000	*rc:Bird*	

[100] https://github.com/structureddynamics/UMBEL/issues/2

2.1 Europeana Food and Drink Classification Scheme

count	type	Note
10000	rc:BodyOfWater	
10000	rc:Book_CW	
10000	*rc:Fish*	
10000	rc:Planet	No mistake: includes asteroids
8677	*rc:Mollusk*	
3851	rc:AnimalBodyPart	
2505	rc:Country	Includes historic (e.g. Etruscan civilization, Khasars) and a few mistakes (e.g. Organisation of African Unity, History of Venezuela)
488	*rc:Drink*	
241	*rc:Grape*	

3.19.3 wikipediaCategories

Categories to RC. This is a manual mapping of very high quality, unfortunately very small (3k)

```
dbr:Category:Shanxi umbel:correspondsTo rc:Shanxi_ProvinceChina .
dbr:Category:Cerebellum umbel:correspondsTo rc:Cerebellum .
dbr:Category:Caste umbel:correspondsTo rc:CasteSystemBeliefs .
dbr:Category:Poncho umbel:correspondsTo rc:Poncho .
dbr:Category:Scavengers umbel:correspondsTo rc:Scavenger .
dbr:Category:Anorectics umbel:correspondsTo rc:Anorectic .
dbr:Category:Roses umbel:correspondsTo rc:RoseBush .
```

3.19.4 wikipediaCategoriesListsIndividuals

Assigns 1.8M DBpedia individuals (50%) to some RC, each individual mapping to 2.44 RC on average.

```
dbr:SODA_Off-Road_Racing umbel:relatesToActivity rc:Game .
dbr:Toshiko_Sato umbel:relatesToPersonType rc:Spy .
dbr:Humphrey_IV_of_Toron umbel:relatesToEvent rc:BirthEvent .
dbr:Estefan_Ciuculescu umbel:relatesToEvent rc:BirthEvent .
dbr:Russ_Scarritt umbel:relatesToEvent rc:BirthEvent .
dbr:Julius_Gitahi umbel:relatesToEvent rc:BirthEvent .
dbr:Kate_Seelye umbel:relatesToPersonType rc:Graduate .
```

This is the biggest part of the mapping. It uses only 5% of RC, so this can be considered a sort of categorization. The 31 relations umbel:relatesTo* reflect the specific Super-Type (UMBEL domain), so this also provides a mapping to super-types.

The specificity is low. Considering the sampling above, only 2-3 of 7 statements are useful:

- umbel:relatesToEvent rc:BirthEvent is useless: everyone has been born

- rc:Graduate is not very useful: most people have graduated from some school at some point in time. (The article has this sentence "Seelye graduated from Amherst College")

Many DBpedia individuals are mapped to several RC (2.44 on average). So the mapping also includes this statement about Kate_Seelye, which is accurate and useful:

```
dbr:Kate_Seelye umbel:relatesToPersonType rc:Journalist .
```

3.19.5 wikipediaCategoriesSVIndividuals

Assigns DBpedia individuals to RC using a different method (Semantic Vectors). Much smaller than the previous file (only 1.3% of individuals are mapped). The specificity is very high: all statements below are useful.

```
dbr:Circus_Joseph_Ashton umbel:relatesToOrganizationType rc:Circus .
dbr:Fairey_Ultra-light_Helicopter umbel:relatesToProductType rc:TransportHelicopter
dbr:Vigna_umbellata umbel:relatesToPlant rc:VignaOWahuensis .
dbr:Walls_of_Ston umbel:relatesToFacility rc:Fortification .
dbr:Fishscale_cocaine umbel:relatesToDrug rc:Cocaine .
dbr:Fudge umbel:relatesToOrganizationType rc:Confectionery .
dbr:Headache_attributed_to_a_substance_or_its_withdrawal umbel:relatesToDisease
rc:Headache .
```

The documentation implies this is a subset of the previous file, but that is not so. E.g. this statement is only found here. This means that we can safely use both files together.

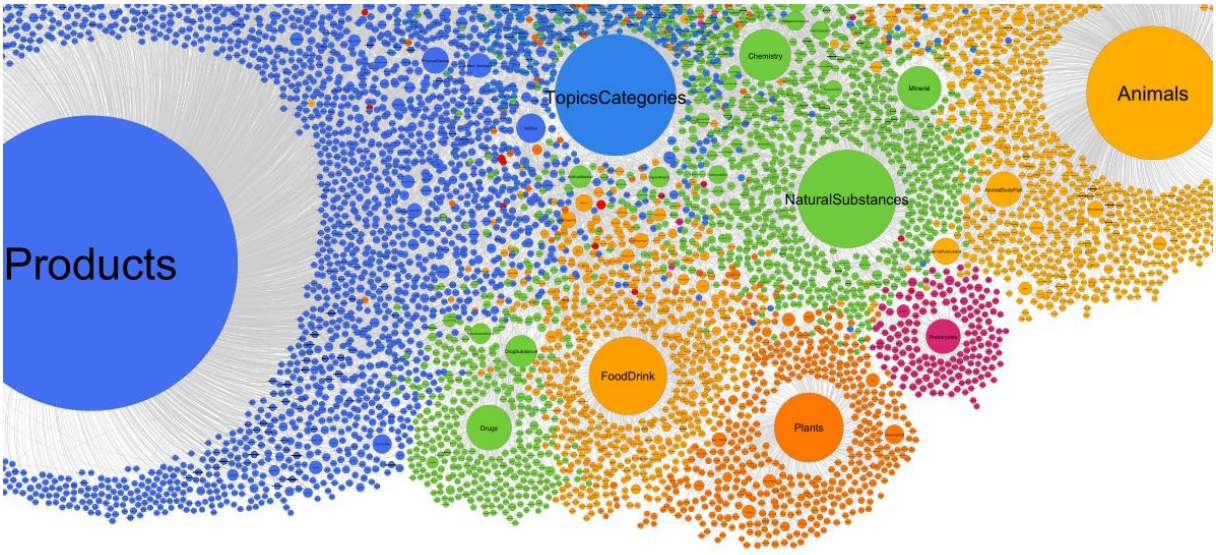
```
dbr:Abatis umbl:relatesToFacility rc:Fortification .
```

3.19.6 UMBEL FD SuperType

UMBEL has a SuperType "FoodDrink" that is directly related to our topic. It is sizable (the yellow constellation in the bottom-middle of this graph)[101]

[101] http://www.umbel.org/resources/graph/

2.1 Europeana Food and Drink Classification Scheme



We used the following approach to extract FD-related statements and DBpedia individuals:

- Extract all statements from the 5 mapping files (see the preceding sections) that mention "food" or "drink". That's 23,840 statements.
- Count the statements per RC. There are 416 RC (1.1% of all RC). The first 10% (42 RC) have 85% (20,330) of the statements, followed by a long tail with few statements.
- We assign a judgement to each RC whether it's relevant to FD: 1=yes, 0=maybe, -1=no. E.g. we've assigned 0 to all creative works (Books, Movies, TV series) because they may have FD as subject or not.

196 RC with 22124 statements are relevant to FD: we show the top and the bottom ones. We save this as **food-RC-from-statements.txt**

8035	rc:Cuisine
1940	rc:Drink
1869	rc:Food
954	rc:Cheese
843	rc:Wine
741	rc:DessertFood
599	rc:ProteinStuff
447	rc:Bread
416	rc:Vegetable_Food
274	rc:Tea_Beverage
251	rc:Whisky
245	rc:Soup
236	rc:Sauce
227	rc:Condiment
207	rc:SoftDrink
192	rc:Pie
1	rc:Cabbage_Foodstuff

1	rc:Brownie_Food
1	rc:BrownRice_Foodstuff
1	rc:Broccoli_Foodstuff
1	rc:Bisque_Soup
1	rc:Bean_Foodstuff
1	rc:Bakery
1	rc:BabyFood
1	rc:AmericanStyleFastFoodCuisine
1	rc:Alcohol_Compound

70 RC with 1134 statements are "maybe", e.g.:

139	rc:Movie_CW
134	rc:Organization
108	rc:Album_CW
96	rc:Cytokine
61	rc:Song_CW
57	rc:Game
48	rc:Book_CW
47	rc:MediaSeriesProduct
39	rc:Cellulose
38	rc:CommercialOrganization
28	rc:CooperativeOrganization
26	rc:Product
24	rc:Novel_CW
24	rc:NewZealand
22	rc:JournalSeries
21	rc:MagazineSeries
16	rc:SocialAidOrganization

149 RC with 582 statements are not relevant, e.g.:

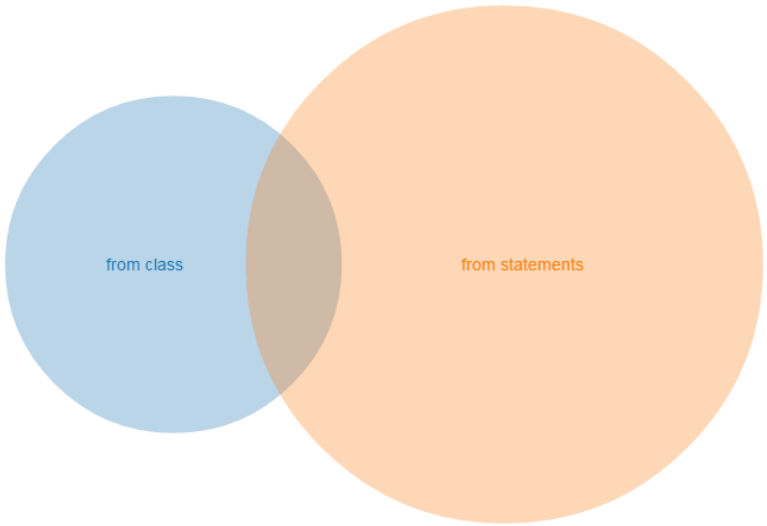
54	rc:Law
53	rc:JustinianCode
33	rc:BirthEvent
32	rc:Science
31	rc:Fodder
30	rc:MusicPerformanceOrganization
26	rc:Politics
23	rc:Building
22	rc:Ministry
12	rc:TradeUnion
10	rc:Architecture

So the number of relevant statements is 23840 - 1134/2 - 582 = 22691

umbel_reference_concepts.n3 has a mapping RC→super-type There are 83 RC in the FD super-type, we save them as **food-RC-from-class.txt**. Some examples:

```
rc:FoodGroup rdfs:subClassOf umbel:FoodDrink .
rc:BlandTaste rdfs:subClassOf umbel:FoodDrink .
rc:SaltingFood rdfs:subClassOf umbel:FoodDrink .
rc:ThingConformingToKosherDietaryLaws rdfs:subClassOf umbel:FoodDrink .
rc:Bottled_alcoholic_beverage rdfs:subClassOf umbel:FoodDrink .
rc:HalalCuisine rdfs:subClassOf umbel:FoodDrink .
rc:FoodIngredientOnly rdfs:subClassOf umbel:FoodDrink .
rc:Animal_HumanFoodSource rdfs:subClassOf umbel:FoodDrink .
```

When we join both RC files to **food-RC-joined.txt**, we obtain 263 RC (diagram made with venn.js)[102]



Now we use this set of RC to search in the 5 mapping files again.

- From wikipediaCategories.ttl we extract **food-cats.txt**, 121 DBpedia categories definitely related to FD, e.g.:

```
dbr:Category:Brandy umbel:correspondsTo rc:Brandy_Liquor
dbr:Category:Butter umbel:correspondsTo rc:Butter
dbr:Category:Cakes umbel:correspondsTo rc:Cake
dbr:Category:Carbohydrates umbel:correspondsTo rc:CarbohydrateStuff
dbr:Category:Roe umbel:correspondsTo rc:Caviar
dbr:Category:Cellulose umbel:correspondsTo rc:Cellulose
dbr:Category:Champagne umbel:correspondsTo rc:Champagne
dbr:Category:Cheese umbel:correspondsTo rc:Cheese
```

- From wikipediaCategoriesListsIndividuals and wikipediaCategoriesSVIndividuals. we extract **food-from-RC.txt**

These are 44237 statements about 37811 DBpedia individuals related to FD. Of them:

- 16324 statements in the FoodDrink domain (using relation relatesToFoodDrink)
- 27913 statements in other domains.

[102] https://github.com/benfred/venn.js/

Some examples from other domains, with counts:

```
    13 umbel:relatesToPersonType rc:Cheesemaker
   930 umbel:relatesToPersonType rc:Chef
    40 umbel:relatesToPersonType rc:Vegan
    16 umbel:relatesToPhenomenon rc:Fermenting
    81 umbel:relatesToPlant rc:Almond
    47 umbel:relatesToPlant rc:CornPlant
   226 umbel:relatesToPlant rc:Crop
   567 umbel:relatesToPlant rc:Cultivar
   366 umbel:relatesToPlant rc:Fruit
    30 umbel:relatesToPlant rc:PeanutPlant
 12568 umbel:relatesToPlant rc:Plant
    57 umbel:relatesToPlant rc:TeaPlant
    45 umbel:relatesToPlant rc:WheatPlant
    85 umbel:relatesToProductType rc:DrinkingVessel
  1004 umbel:relatesToProductType rc:EatingVessel
   118 umbel:relatesToProductType rc:FoodUtensil
   231 umbel:relatesToProductType rc:HouseholdAppliance
   136 umbel:relatesToProductType rc:PackagingContainerProduct
   127 umbel:relatesToProductType rc:Windmill
   492 umbel:relatesToSubstance rc:Alcohol_Compound
    60 umbel:relatesToTopic rc:AgriculturalEconomics
  1104 umbel:relatesToTopic rc:Agriculture
    85 umbel:relatesToTopic rc:Agronomy
   114 umbel:relatesToTopic rc:FoodScience
   359 umbel:relatesToTopic rc:Gastroenterology
   333 umbel:relatesToTopic rc:Nutrition
    93 umbel:relatesToWrittenInfo rc:CookBook
```

All these are related to FD, with possible exceptions for rc:Plant, rc:HouseholdAppliance and rc:PackagingContainerProduct (maybe we were a bit liberal that these RCs relate to FD).

So we extracted from UMBEL 121 DBpedia categories, 37811 individuals and 44237 related to FD. This is quite a catch!

4 Visualisation

With such large amounts of data it is essential to find effective visualization methods. Visualisation helps data researchers understand the data, and helps content providers navigate the classification, cut off inappropriate branches, and perform other crowd-sourcing actions

4.1 A Menagerie of Visualisations

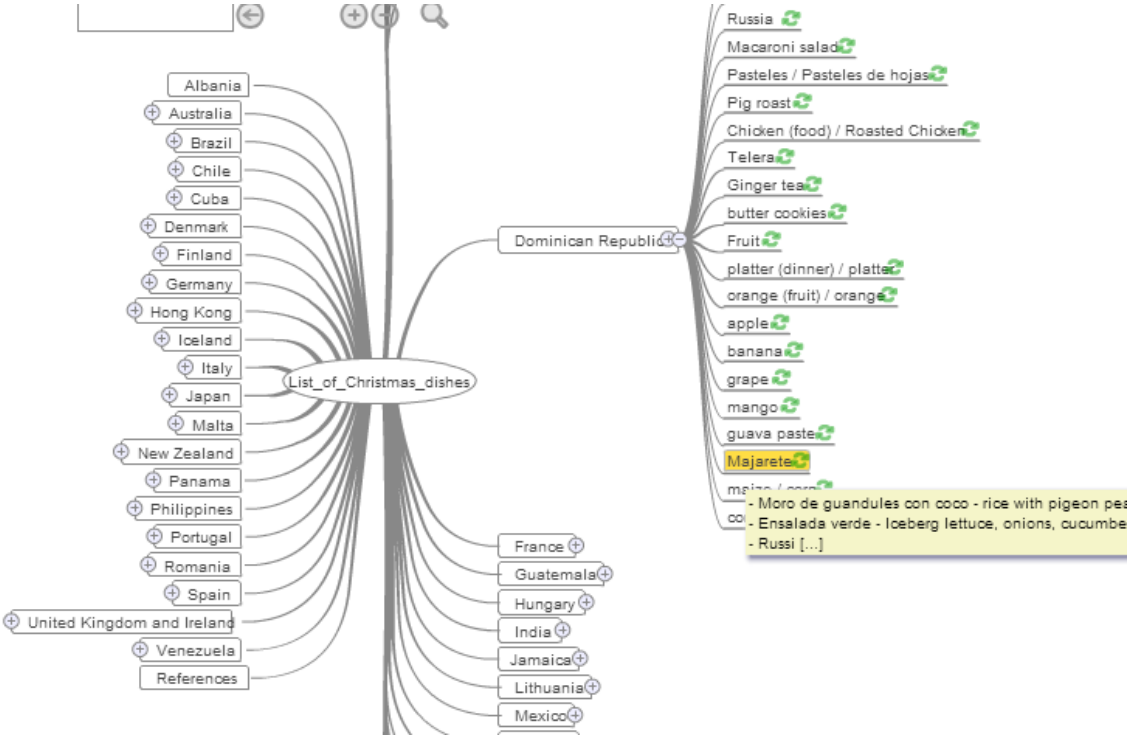
We researched a number of various approaches for visualisation of Wikipedia categories and similar structures, and finally settled on d3.js. A lot of examples and pointers are available at http://www.visualcomplexity.com/

4.1.1 WikiMindMap

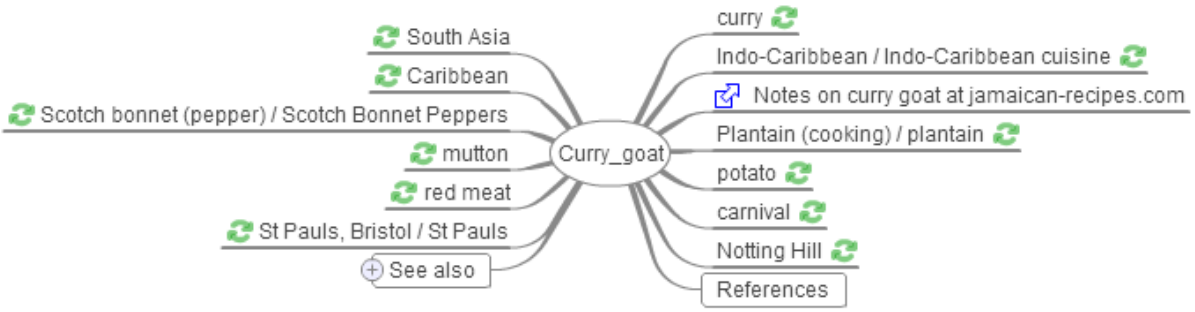
WikiMindMap[103] is very useful to explore a page and its connections

- Picky about spelling. Wait a good time after you enter an article name
- Works on articles only, not categories.
- Parses out page sections and links. Shows the sections expandable

E.g. explore Christmas dishes.[104] Here the section on Dominica is expanded



Or explore the ingredients of Curry Goat:



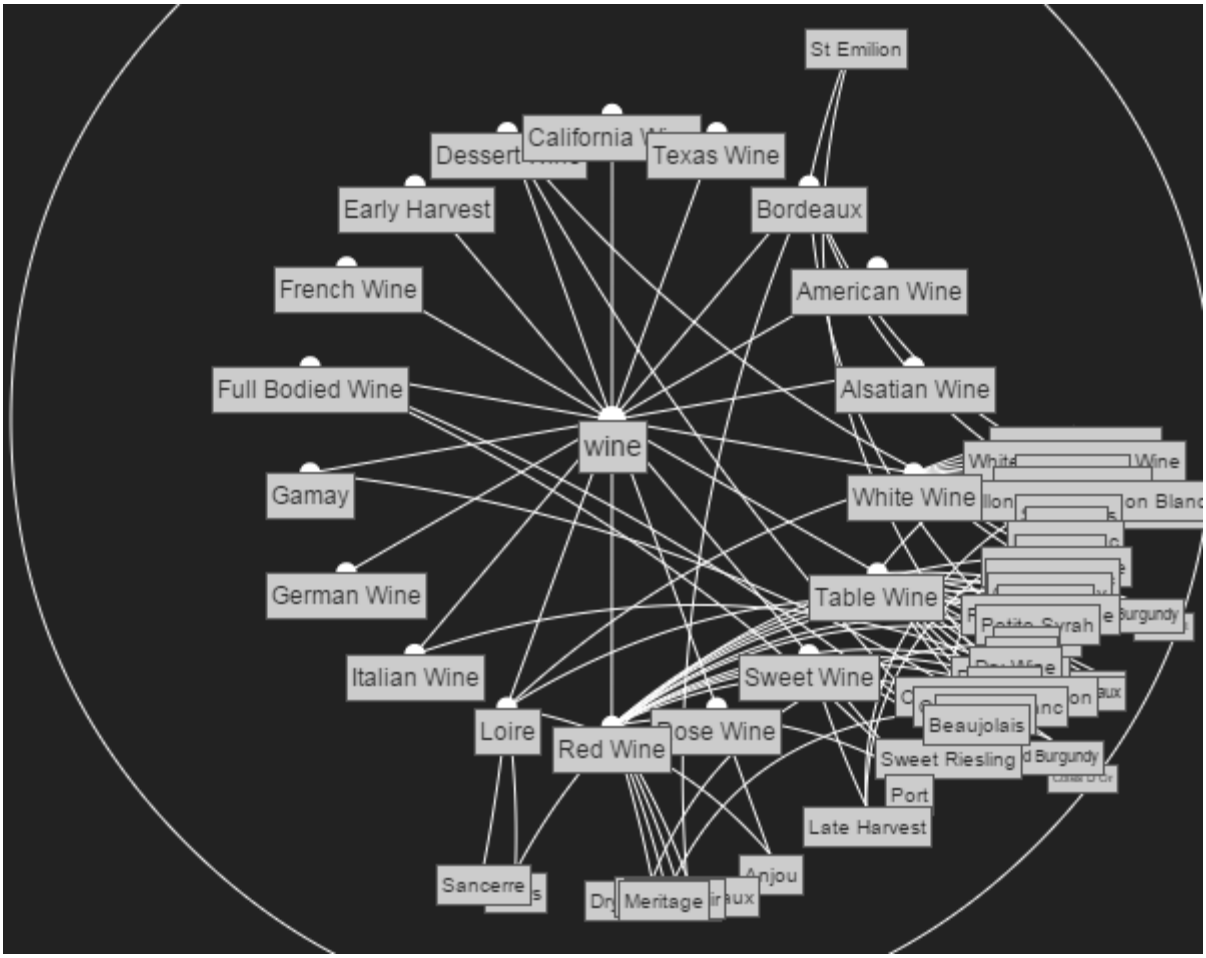
Export to Freemind doesn't work[105]

---

[103] http://www.wikimindmap.org/
[104] http://www.wikimindmap.org/viewmap.php?wiki=en.wikipedia.org&topic=List+of+Christmas+dishes
[105] https://github.com/nyfelix/wikimindmap/issues/3

4.1.2 jOWL Hyperbolic Tree

A hyperbolic tree[106] showing the well-known Wine ontology. A bit sluggish and refreshes unpleasantly.



4.1.3 WikiStalker

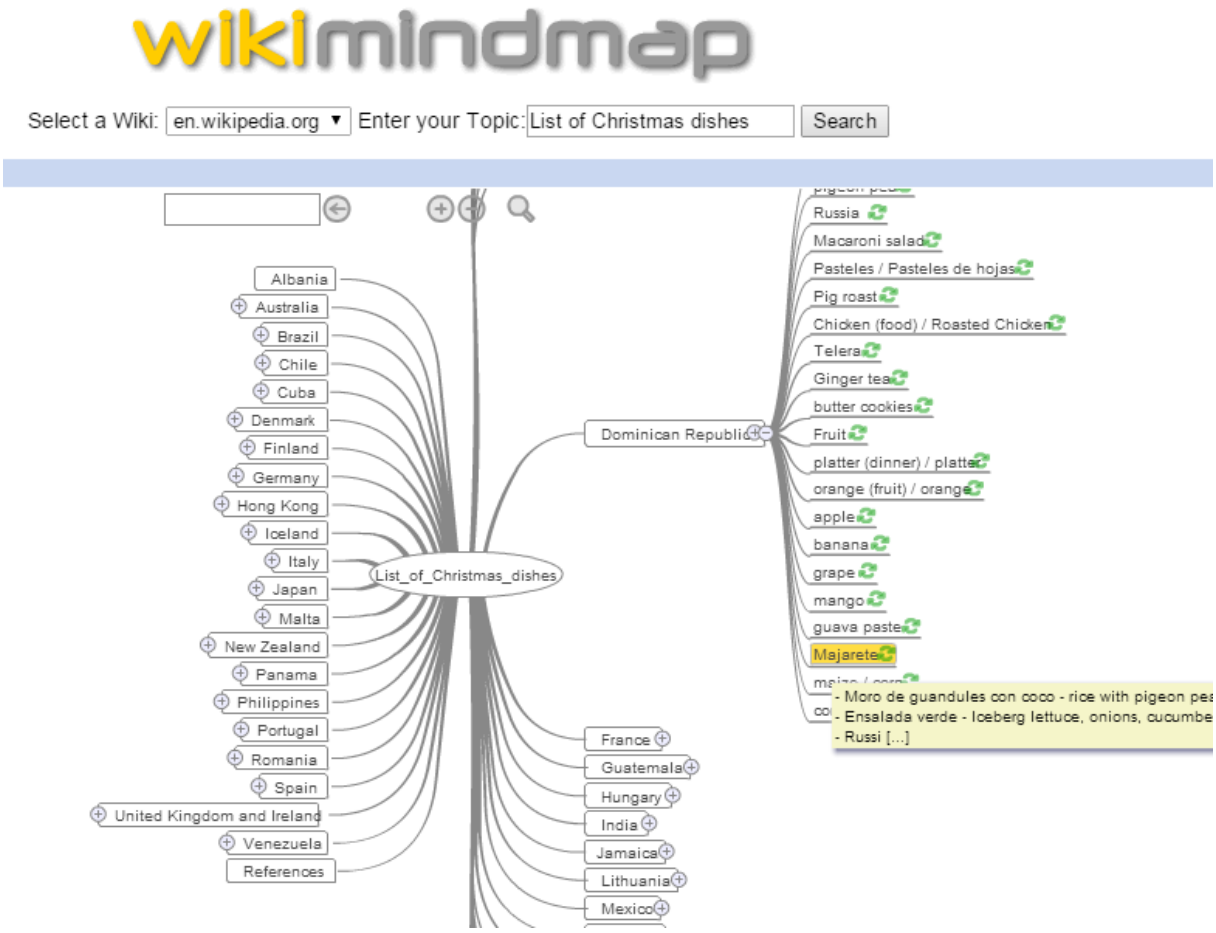
Wiki Stalker[107] provides Wikipedia Category Structure Visualization.

- Not very readable for a page with a lot of links (eg List_of_Christmas_dishes)
- Shows the categories on the right
- Can filter by % relevance (connectedness)
- Can order by relevance (becoming a spiral shape) or alphabetically
- Can show the links of a second selected article

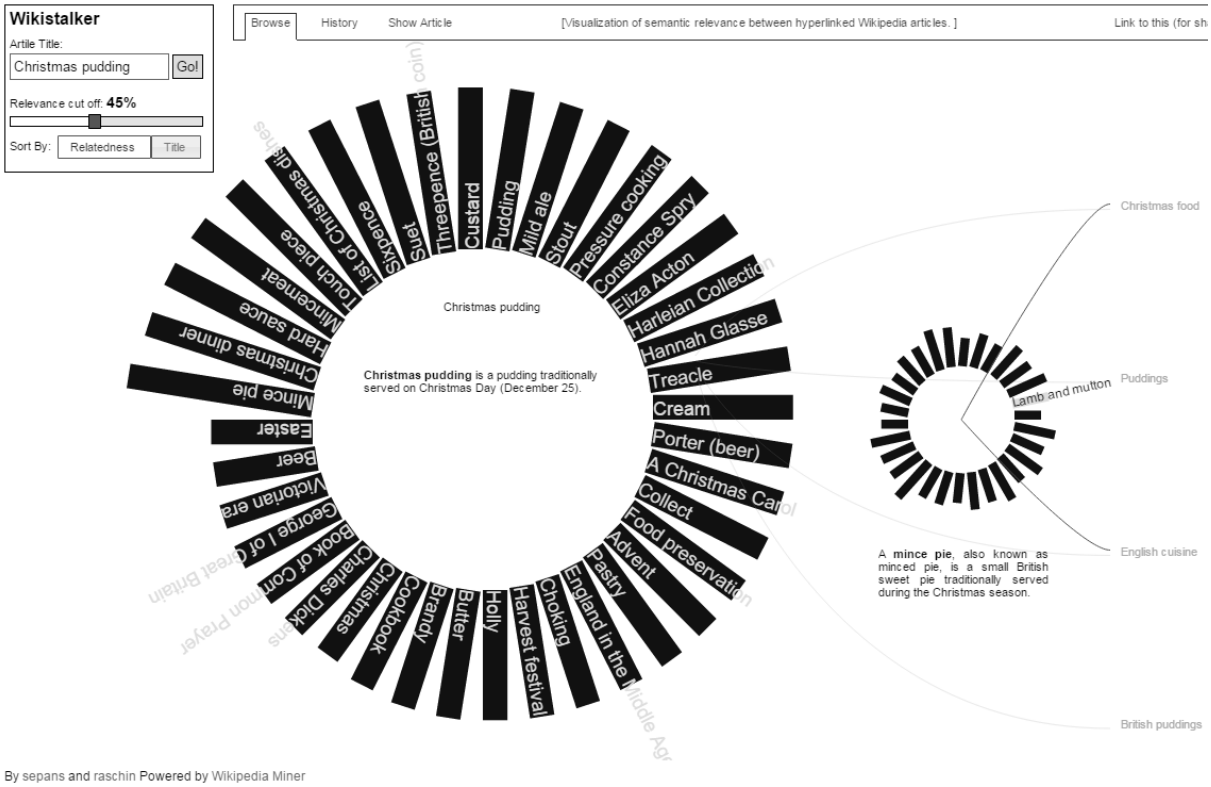
[106] http://jowl.ontologyonline.org/HyperBolicTree.html
[107] http://sepans.com/sp/works/wikistalker/

2.1 Europeana Food and Drink Classification Scheme

List of Christmas dishes:



Christmas Puddings[108]



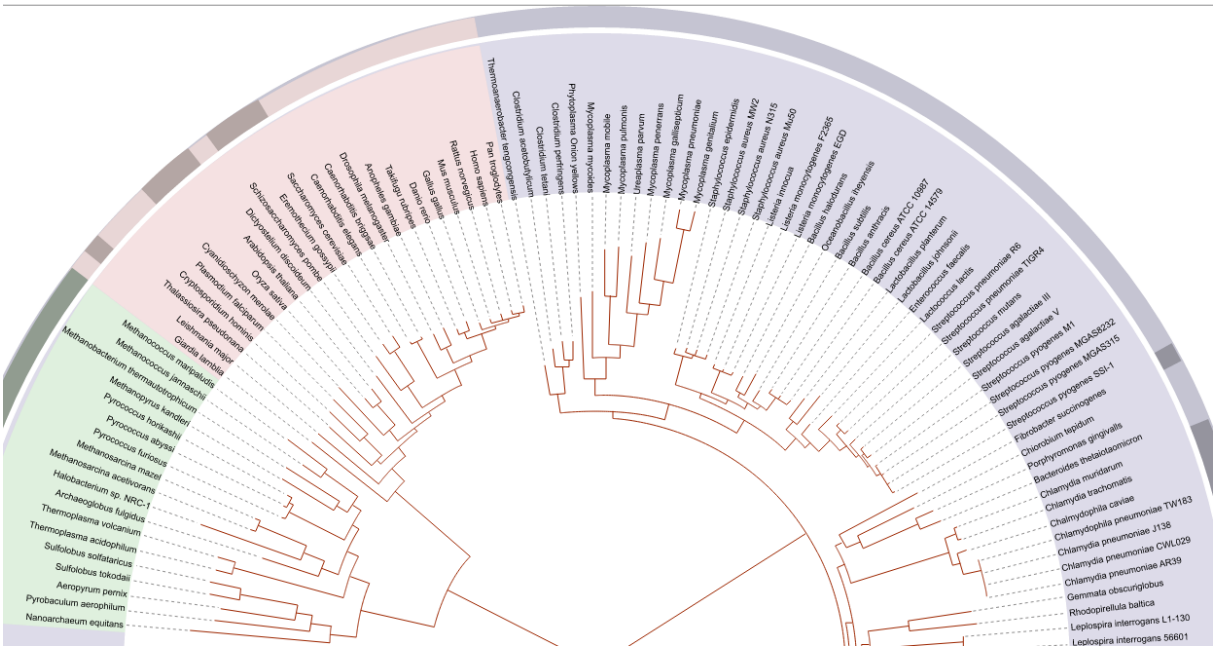
4.1.4 Interactive Tree Of Life

Interactive Tree Of Life[109] is a highly optimized visualisation of biological taxonomy. Provides several modes: circular (normal or inverted), normal-tree, unrooted. The easiest way to explore it (most non-intrusive for your browser) is to see an SVG capture[110]. Or you can play with the tool itself and try different options.

It scales well and is a very promising approach for visualising categories. However, it may be harder to obtain and customize.

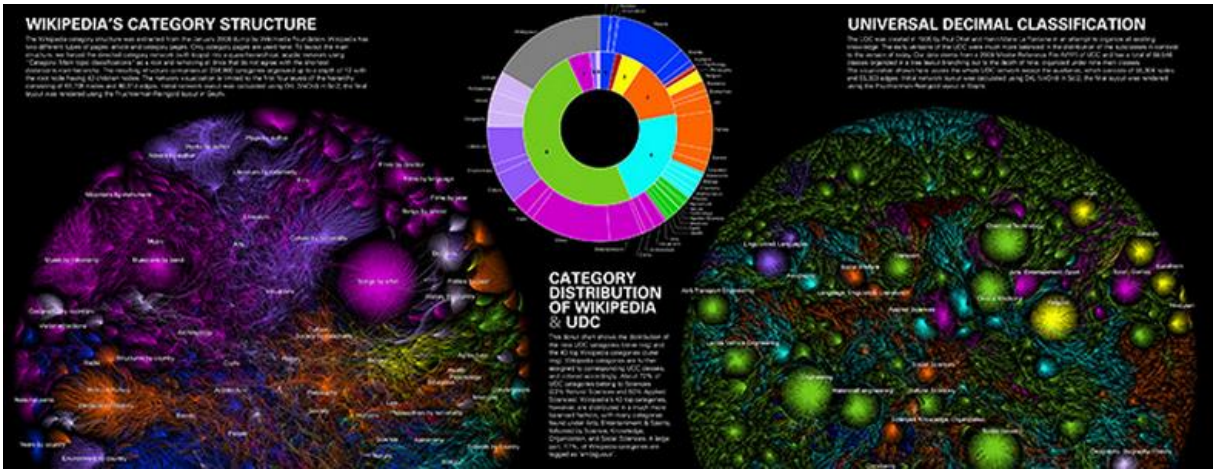
---

[108] http://sepans.com/wikistalker/?title=Christmas%2Bpudding&rel=0.45&sort=relatedness
[109] http://itol.embl.de/
[110] https://commons.wikimedia.org/wiki/File:Tree_of_life_SVG.svg



4.1.5 Wikipedia vs UDC

Wikipedia categories vs Universal Decimal Classification[111] is a high-end artistic visualization created with Sci2/DlR/VxOrd, Gephi, MagnaView. It was created by the Knowledge Space Lab, e-humanities group, RKD and submitted to SciMaps, a curated collection of scientific visualisations.



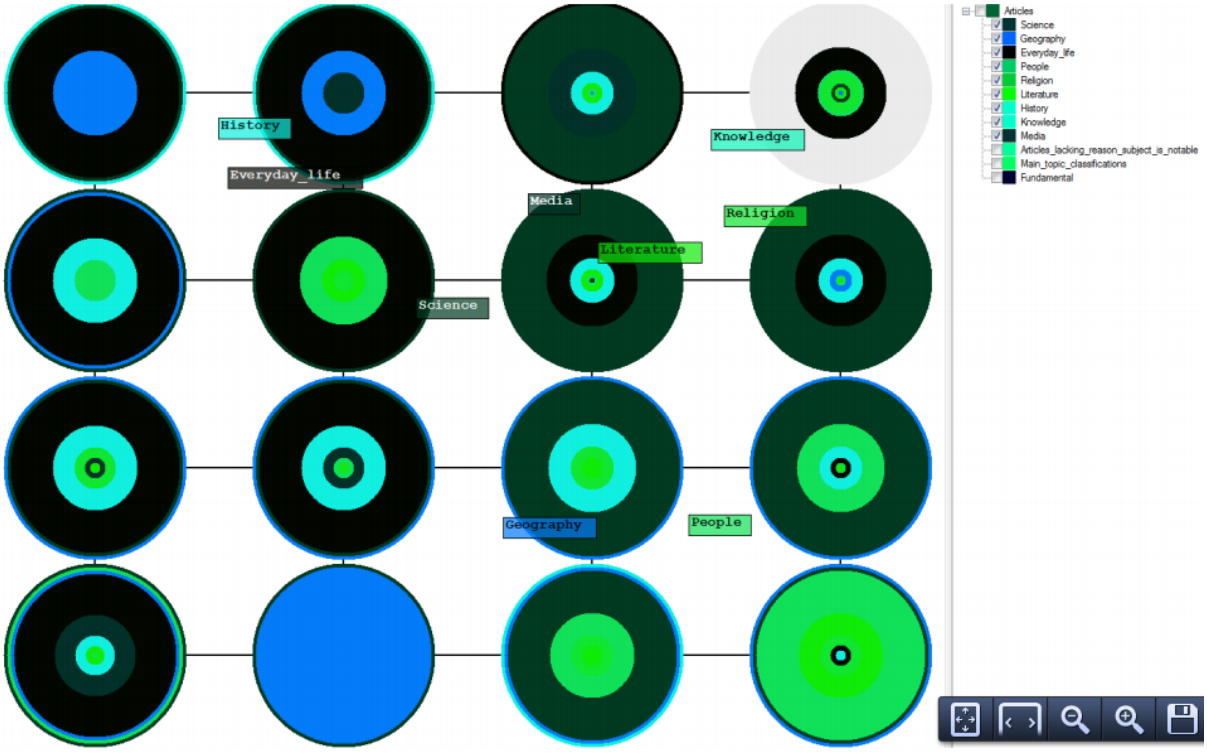
Eliminated cycles from Wikipedia categories by removing edges that don't agree with the shortest distance to root. As of 2008, EN Wikipedia had 235k categories, depth 12 (now has 5.5 times more). The visualization shows 61k categories, depth 4. Food doesn't even appear.

We have included it here for its beauty, but it has no practical application for EFD.

[111] http://scimaps.org/mapdetail/design_vs_emergence__127

4.1.6 Self-Organizing Maps

Self Organizing Maps were applied to visualise Wikipedia categories.[112] It shows an overview of categories and their relations, helping to narrow down search domains. Selecting particular neurons this approach enables retrieval of conceptually similar categories.



This shows some promise, but would be harder to customize and deploy.

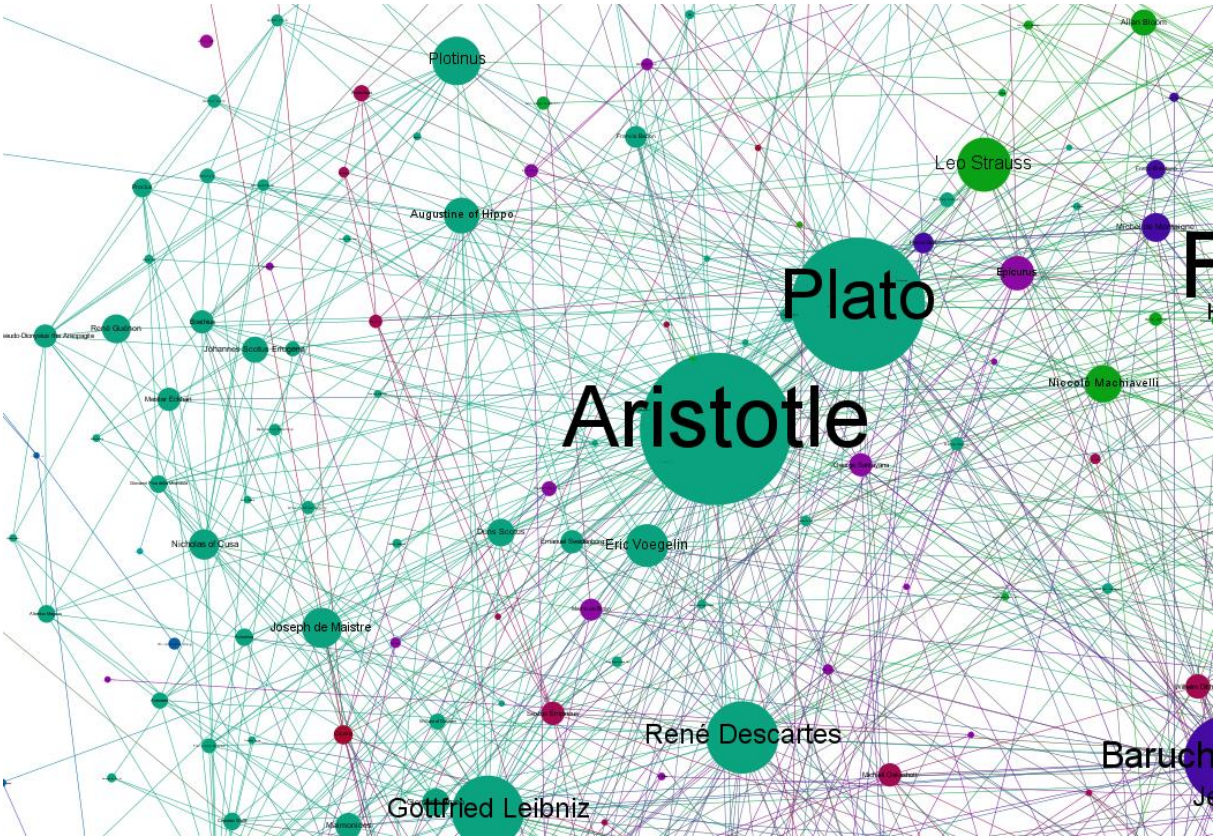
4.2 Gephi Visualisation

Gephi is powerful open source desktop software for graph visualisation, dubbed "Photoshop for Graphs". It has a number of plugins for: working with RDF graphs, SPARQL querying, R statistical calculations, a number of built-in graph algorithms, etc.

E.g. below is a map of influences in philosophy[113] produced from DBpedia data: class Philosopher and property influencedBy.

Gephi has a steep learning curve, requires a lot of dedication, and is perfect for complex graphs, or finding the right visualisation for a graph through experimentation. It's likely too complex for our purposes

[112] http://www.fizyka.umk.pl/publications/kmk/12-SOM-categories.pdf
[113] http://www.coppelia.io/2012/06/graphing-the-history-of-philosophy/



4.3 d3 Visualisations

d3 is a JavaScript library for visualisations completely in the browsers, using SVG and CSS. The charts are fully zoomable. Interactivity may also be programmed using standard browser events. Some books are available:

- Beginning JavaScript Charts
- Getting Started with D3 (2012)
- Interactive Data Visualization for the Web (2013)

A number of papers and tutorials are available:

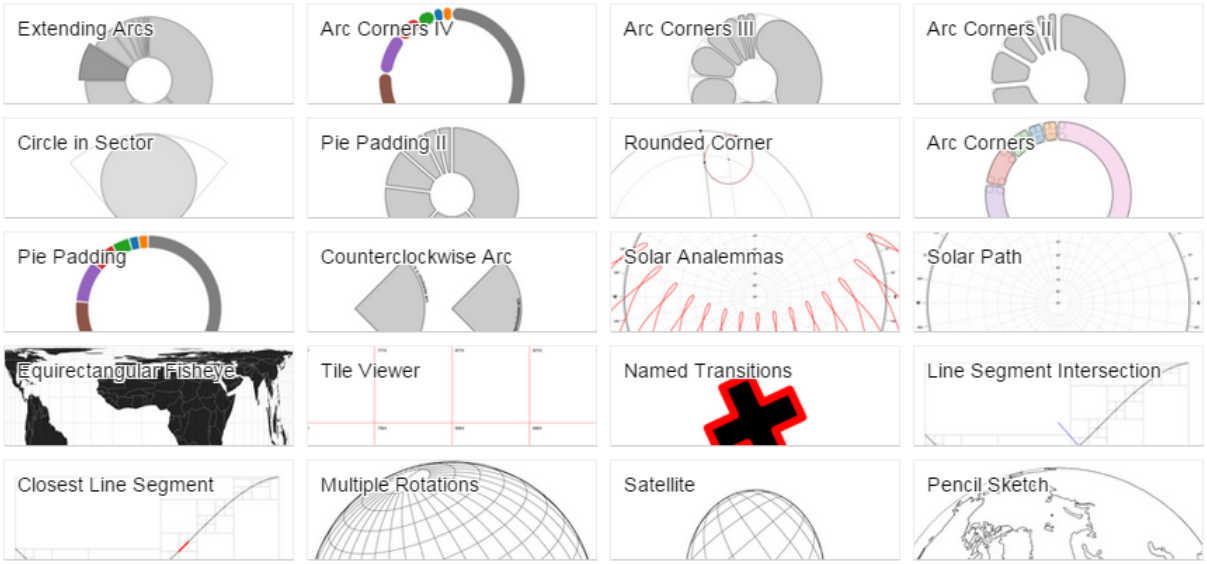
- D3: data driven documents (2011 InfoVis): original paper by d3's author Mike Bostock
- D3.js tutorial (Strata conference 2013)

Its biggest strengths are:

- A site for easy publishing of examples http://bl.ocks.org/ using the "gist" (dode snippet) approach

2.1 Europeana Food and Drink Classification Scheme

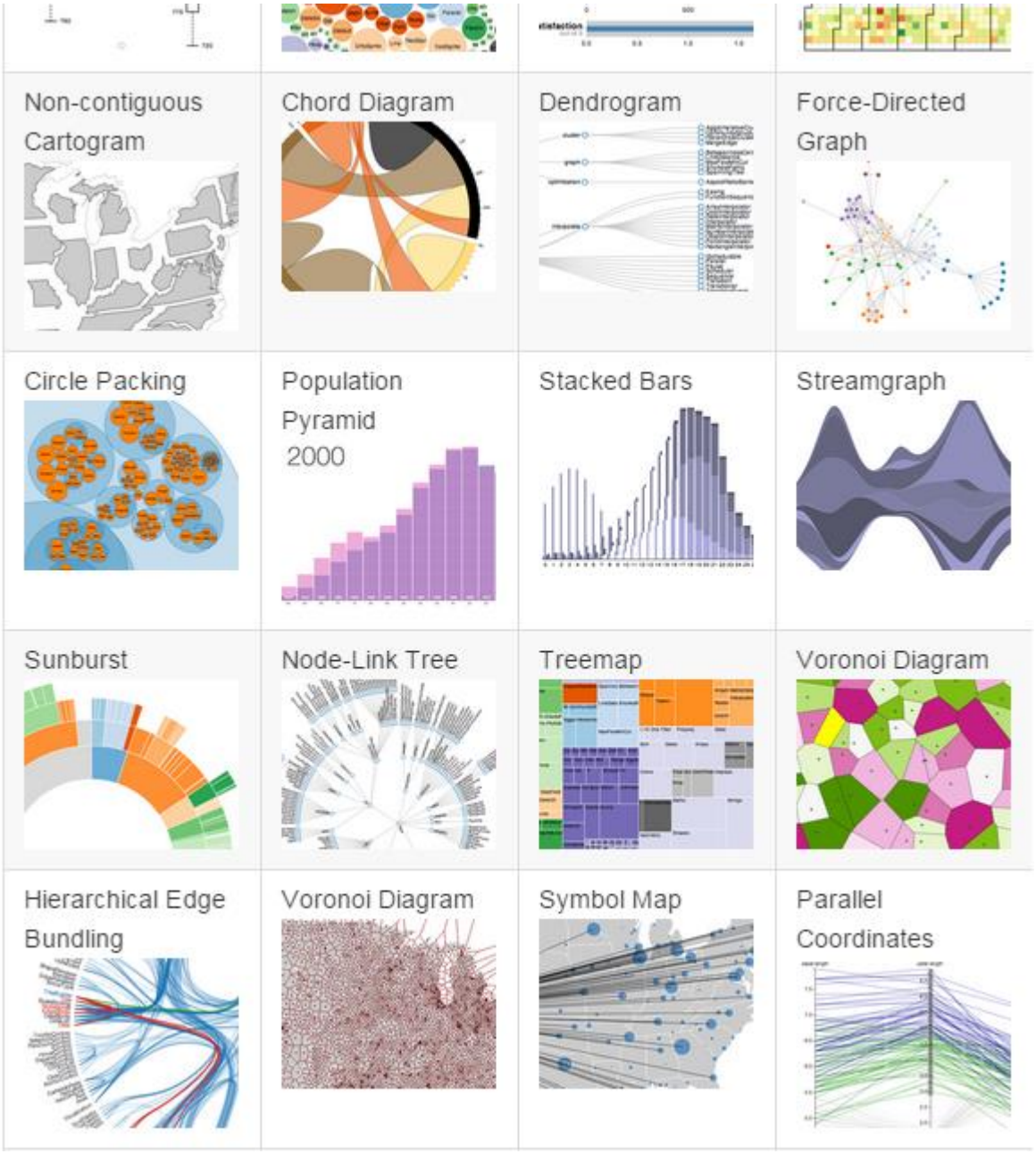
- A huge number of examples by Bostock http://bl.ocks.org/mbostock (849) and many others



2.1 Europeana Food and Drink Classification Scheme

4.4 d3 Visual Index

A visual index/wiki of different diagram types[114]. 270 visual examples, and a number of links



4.4.1 RAW Design Tool

There are some dedicated d3 tools:

[114] https://github.com/mbostock/d3/wiki/Gallery

2.1 Europeana Food and Drink Classification Scheme

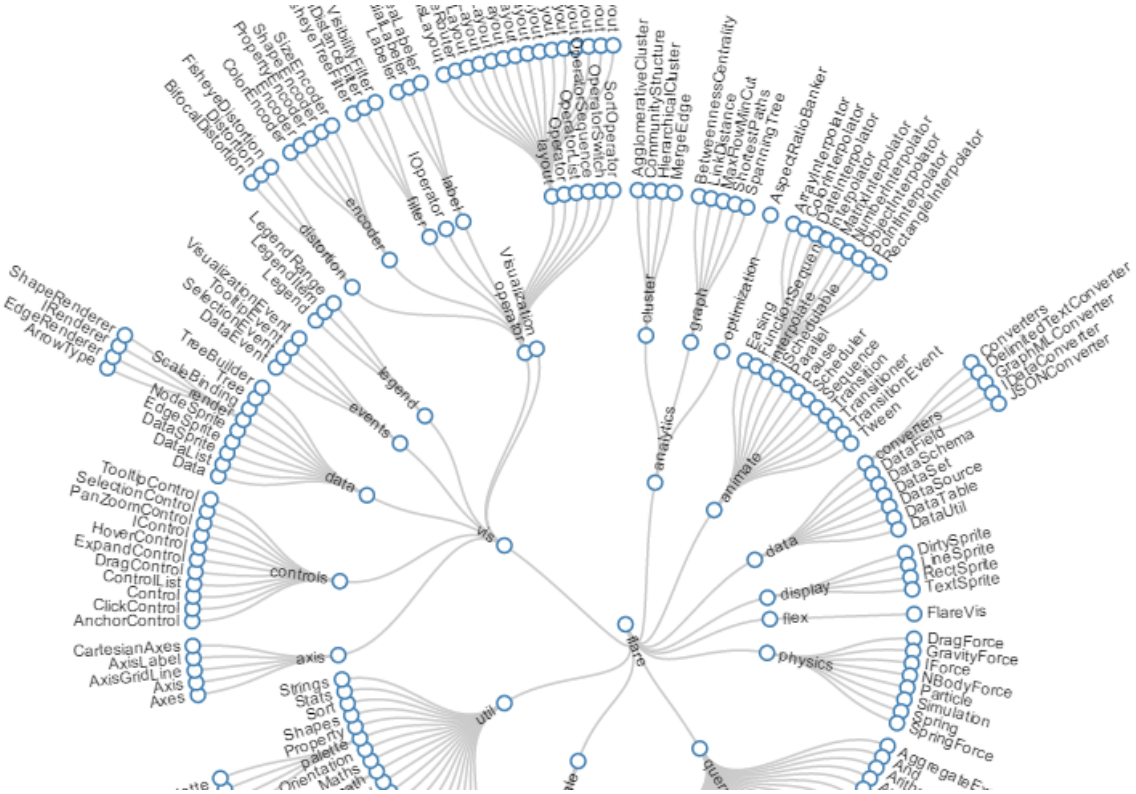
- http://bl.ocks.org/shancarter/raw/4748131/: console for learning and experimenting with d3.js data nesting.
- http://nvd3.org: 12 re-usable charts and chart components for d3.js. With source and fiddle/hack console
- http://app.raw.densitydesign.org/: visualize any CSV data: pick from a gallery of 10 D3.js charts, configure interactively

The following dendrogram (cocktails and their ingredients) was generated with RAW from a simple table:

4.4.2 Radial Tree

In the rest of this section we show the diagram types that are most appropriate for visualizing categories. First is a Radial Reingold–Tilford (Node-Link) Tree or

dendrogram)[115]. Notice how the interleaving of layers saves space: this can fit 250 nodes.



Variations of the dendrogram include:

- Interactive (collapsible) tree[116]. Nature magazine tried it successfully for visualising the complete Nature taxonomy[117]
- Fully interactive tree[118] with pan, zoom, collapse/expand, drag & drop. However, this is a Cartesian not Radial tree and takes more space

4.4.3 Radial FD Tree

We generated a radial tree for the first 2 levels of the Food and drink category hierarchy.[119] It took a 35 line Perl script that uses the Graph::Directed module:

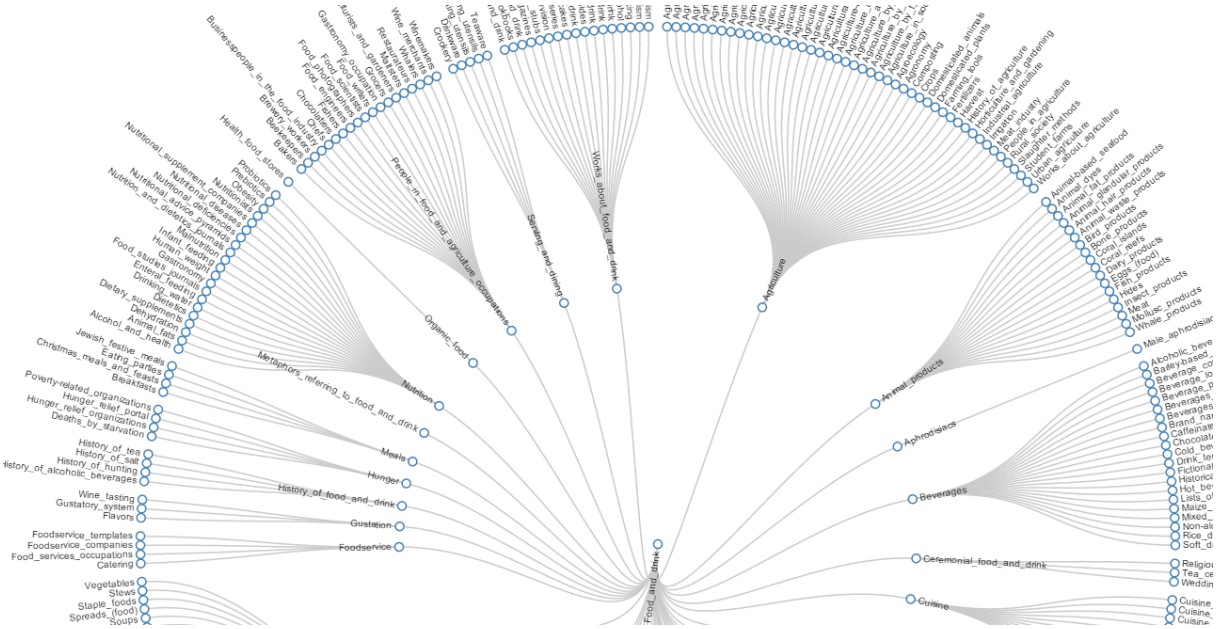
- Load a CSV file to a graph using add_edge()
- Remove loops and redundant edges by using single_source_shortest_paths() to convert to tree
- Dump to JSON using a custom depth-first-search function

---

[115] http://bl.ocks.org/mbostock/4063550
[116] http://bl.ocks.org/mbostock/4339083
[117] http://michelepasin.org/blog/2013/06/21/messing-around-wih-d3-js-and-hierarchical-data/
[118] http://bl.ocks.org/robschmuecker/7880033
[119] http://bl.ocks.org/VladimirAlexiev/raw/1aa55bbdf3b20f8f08d9/

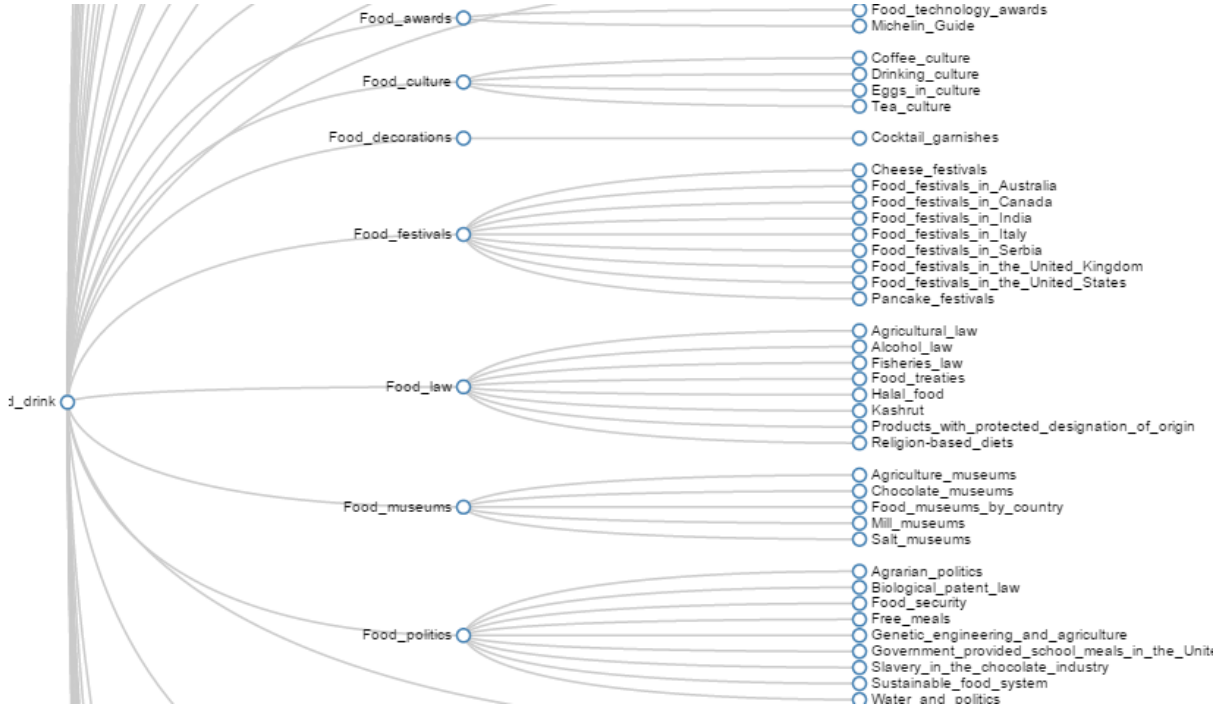
2.1 Europeana Food and Drink Classification Scheme

Try the URL, it's an interactive SVG that you can zoom



4.4.4 Cartesian Tree

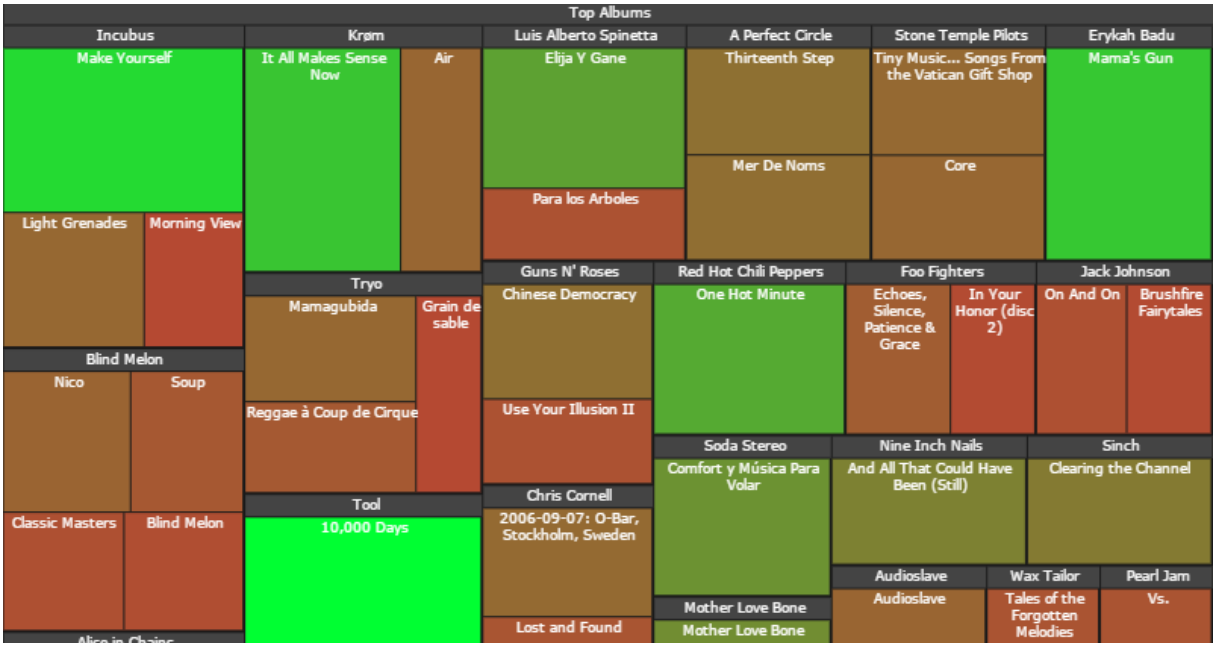
Below is a Cartesian clustered tree[120] of the same data. It takes a lot of space



[120] http://bl.ocks.org/VladimirAlexiev/raw/a6263336b0cdbb586d40/

4.4.5 Tree Map

A tree map shows category "size" (e.g. number of descendants categories and/or pages) by area. This example[121] shows some albums. 2-3 levels can fit.

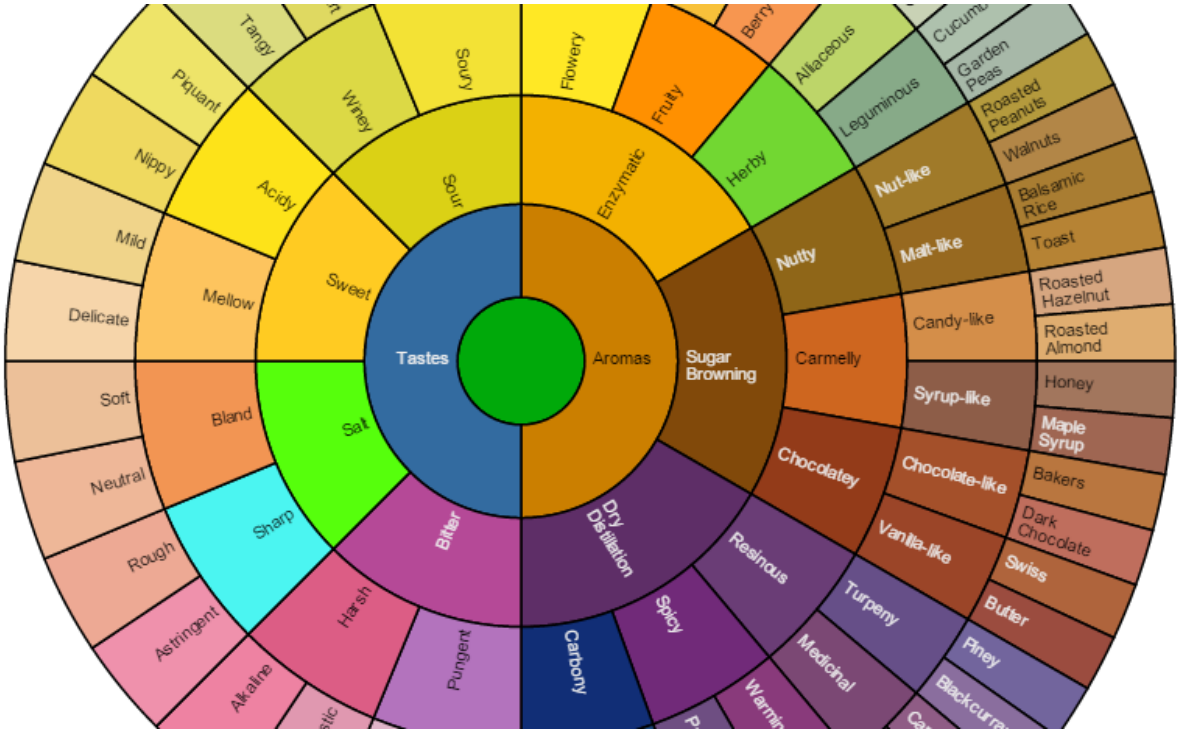


A zoomable treemap[122] allows you to drill down by clicking on a nested category, and go back up by clicking on the header.

[121] http://philogb.github.io/jit/static/v20/Jit/Examples/Treemap/example1.html
[122] http://bost.ocks.org/mike/treemap/

4.4.6 Sunburst

A Sunburst diagram shows sub-categories in sectors, sized per "size" of the sub-cat. Here is an example with coffee styles/flavors[123]



These are all similar:

- Zoomable Sunburst[124] with d3
- Bilevel Partition[125] (same thing)
- Sunburst Partition[126] on 4 levels (but may be hard to fit text on it

5 Annex 1: Classification Examples

In this section we provide a number of classification examples based on objects listed in the Content inventory, and from additional research of EFD content. We provide URLs usually to Wikipedia, but keep in mind that these items are also available in structured LOD.

[123] http://www.jasondavies.com/coffee-wheel/
[124] http://bl.ocks.org/mbostock/4348373
[125] http://bl.ocks.org/mbostock/5944371
[126] http://bl.ocks.org/mbostock/4063423

5.1 Coral Food Pounder

	Food Pounder Cut From Coral, Penu, Austral Islands, Central Polynesia. Penu food pounders of this horned, concavely conical form are found with several variations in style throughout Central and Eastern Polynesia. The design is highly ergonomic - adapted over centuries to fit the hand perfectly and allow exactly the right kind of mechanical action to be applied to the food in the wooden bowl. The selection of a heavy slab of coral from the fringing reef created a working surface of regular pits and ridges that mashed the cooked root vegetables quickly and easily. In general, such pounders were used to make poi, a pudding of mashed taro, yams or breadfruit, moistened and sweetened with coconut milk, and steamed on hot rocks in an earth oven

- Object: http://vocab.getty.edu/aat/300024725 pestles
  - https://en.wikipedia.org/wiki/Mortar_and_pestle is similar, but is a group of 2 objects
- Material: http://vocab.getty.edu/aat/300250925 corals (animals)
  - Same as: https://en.wikipedia.org/wiki/Coral
- Style/Period: http://vocab.getty.edu/aat/300021975 Austral Islands.
  Culture and nationality of the inhabitants of the Austral Islands. Body decoration was very popular with the people of the Austral Islands, specifically shell jewelry, necklaces, grass skirts, earrings, and breast ornaments made from pearl shells combined with string made of human hair. They also created highly crafted sculpture and carved artwork
- Spatial (made and used): http://en.wikipedia.org/wiki/Austral_Islands
  - Same as: http://vocab.getty.edu/tgn/1009851 Îles Tubuaï
  - Same as: http://www.geonames.org/4033250/iles-tubuai.html
  - Penu is not found. Geonames knows 10 places named Penu[127], but neither of them is on the Austral Islands

[127] http://www.geonames.org/search.html?q=penu

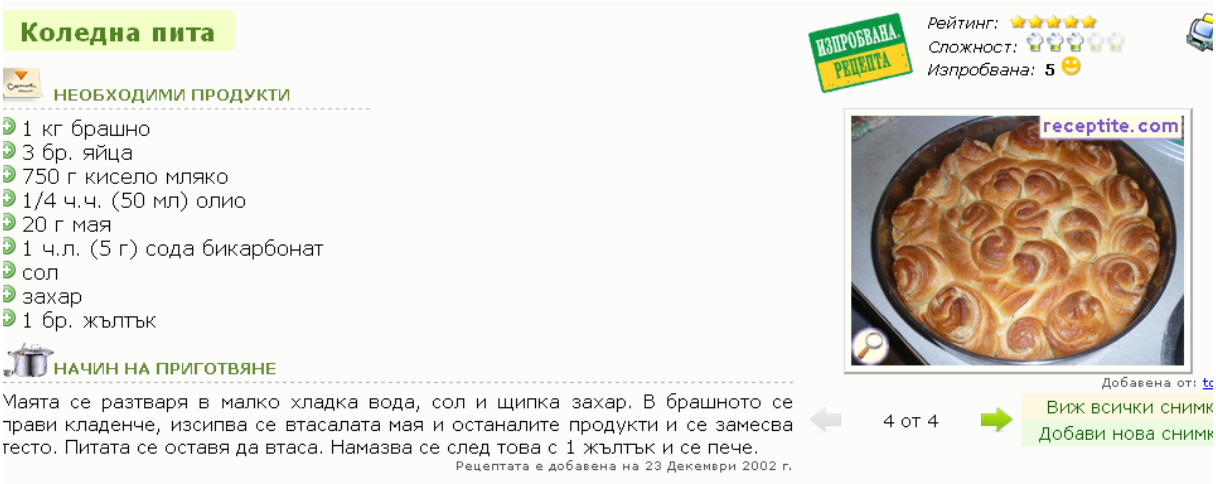
5.2 Lemco Consomme

CAG (Centre for Agriarian History)

- Object: http://vocab.getty.edu/aat/300028730 labels (identifying artefacts)
  - https://en.wikipedia.org/wiki/Nutrition_facts_label is similar but has narrower meaning
- Spatial (made and used): https://en.wikipedia.org/wiki/France.
- Subject (food):
  - https://en.wikipedia.org/wiki/Consommé
  - https://en.wikipedia.org/wiki/Chicken
  - https://en.wikipedia.org/wiki/Portable_soup (bouillons en tablettes)
  - https://en.wikipedia.org/wiki/Meat_extract (Extrait de viande)
- Brand: https://en.wikipedia.org/wiki/Liebig's_Extract_of_Meat_Company (Lemco)

Interest: nostalgia, vintage design, decorative function.

5.3 Christmas Bread



- Object: http://vocab.getty.edu/aat/300027043 recipes
  - Same as: https://en.wikipedia.org/wiki/Recipe
- Subject (food): https://en.wikipedia.org/wiki/Česnica (enwiki, Serbian)
  - Same as: https://ru.wikipedia.org/wiki/Рождественский_хлеб (ruwiki, Russian)
  - Same as: https://ru.wikipedia.org/wiki/Погача (ruwiki, Bulgarian)
- Spatial: Bulgaria & Macedonia (погача, пита), Serbia (чесница), Greece (basilopitta)

Interests:

- Christmas foods
- Trans-regional foods and distribution of recipes

5.4 Woven Drinking Cup

- Object: https://en.wikipedia.org/wiki/Cup
  - Same as: http://vocab.getty.edu/aat/300043202 cups (drinking vessels)
- Material: …
- Spatial (made, used): https://en.wikipedia.org/wiki/North_America
  - Same as: http://vocab.getty.edu/tgn/7029440 (North America)

- Technique: https://en.wikipedia.org/wiki/Weaving
  - Same as: http://vocab.getty.edu/aat/300053642 weaving
- Culture/Style: http://vocab.getty.edu/aat/300017442 Native North American styles

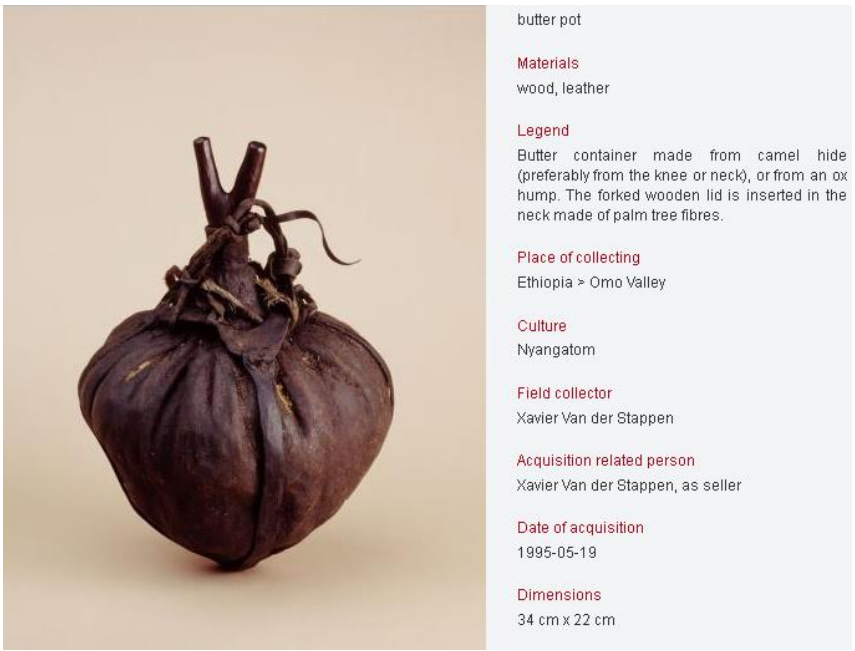
Interest: unusual application of a technique

5.5 Machete

http://www.horniman.ac.uk/object/13.7.56/61

- Spatial (made, used): https://en.wikipedia.org/wiki/Belgian_Congo
- Culture/Style: https://en.wikipedia.org/wiki/Mbuti_people
  - Same as: http://vocab.getty.edu/aat/300260713 Mbuti (Pygmies of Zaire)
- Material: https://en.wikipedia.org/wiki/Iron, https://en.wikipedia.org/wiki/Wood

5.6 Butter Pot



2.1 Europeana Food and Drink Classification Scheme

RMCA

- Object:
- Subject (made for): http://en.wikipedia.org/wiki/Butter
- Materials: http://en.wikipedia.org/wiki/Wood, http://en.wikipedia.org/wiki/Leather
- Material (origin): http://en.wikipedia.org/wiki/Camel, http://en.wikipedia.org/wiki/Ox
- Spatial (found) https://en.wikipedia.org/wiki/Omo_River (Ethiopia)
  - Same as: http://sws.geonames.org/8692910. TGN doesn't have such entry
- Culture/style: https://en.wikipedia.org/wiki/Nyangatom_people
  - Surprisingly, AAT doesn't have such entry (also searched for the alias Donyiro)

Interest: unusual material

5.7 Samovar

ICIMSS

- Object: https://en.wikipedia.org/wiki/Samovar
- Material: https://en.wikipedia.org/wiki/Copper
- Spatial (made, used): https://en.wikipedia.org/wiki/Russia

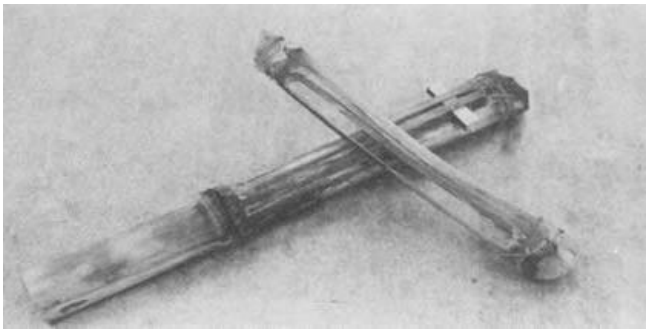
5.8 Cheese Horse

http://www.horniman.ac.uk/get_involved/blog/keeping-food-in-our-collections

- Spatial (made, used): https://en.wikipedia.org/wiki/Poland
- Material: https://en.wikipedia.org/wiki/Cheese
- Object Type: https://en.wikipedia.org/wiki/Doll, https://en.wikipedia.org/wiki/Figurine
- Subject (depicted): https://en.wikipedia.org/wiki/Horse

Interest: unusual material

5.9 Cornstalk Fiddle



https://en.wikipedia.org/wiki/Cornstalk_fiddle,
http://www.banjohangout.org/topic/292931,
http://www.philipblackburn.com/CornstalkFiddles.html

- Object: https://en.wikipedia.org/wiki/Cornstalk_fiddle
- Material: https://en.wikipedia.org/wiki/Maize
- Classification: String instruments, Bowed instruments, Maize products
- Spatial (use):
  - https://en.wikipedia.org/wiki/United_States
  - https://en.wikipedia.org/wiki/Serbia

Interest: an unusual artefact

- Unusual material. A Cornstalk Fiddle is a musical instrument made from a maize stalk.
- Lest you think it's just a joke, we've shown a photo of a Cornstalk Fiddle Sextet.
- Immortalized in folksongs such as "Cotton Eye Joe" that refer to a "cornstalk fiddle and a shoestring bow"
- Paul Laurence Dunbar's poem *The Corn-Stalk Fiddle* describes the construction of the fiddle and playing it at a square dance.

5.10 Shark Hook

http://www.horniman.ac.uk/collections/browse-our-collections/object/136887

- Object Type: https://en.wikipedia.org/wiki/Fish_hook
- Spatial (made, used): https://en.wikipedia.org/wiki/New_Britain
- Material: https://en.wikipedia.org/wiki/Wood

Interest: Looking at the linear scale, this hook is 50 cm long, so perhaps it's intended for Great White sharks

5.11 Cake-Sculpture-Painting

http://collections.britishart.yale.edu/vufind/Record/1670022

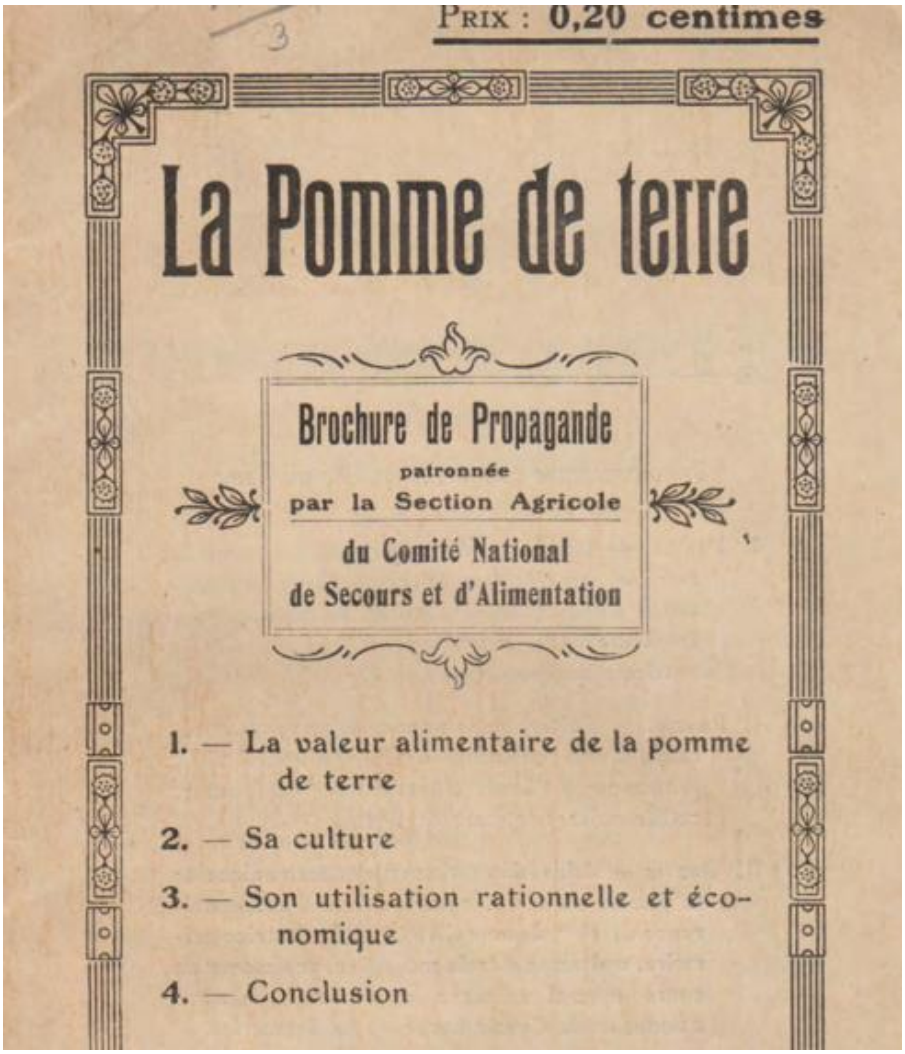
- Creator: http://vocab.getty.edu/ulan/500115024 Lucas, Sarah, (English sculptor, installation artist, and photographer, born 1962)
- Object Type: https://en.wikipedia.org/wiki/Cake, https://en.wikipedia.org/wiki/Sculpture, https://en.wikipedia.org/wiki/Painting
- Material/Technique:

2.1 Europeana Food and Drink Classification Scheme

- https://en.wikipedia.org/wiki/Inkjet_printing
- https://en.wikipedia.org/wiki/Fruitcake
- http://vocab.getty.edu/ulan/300215468 cartons (containers)
  - Similar to: https://en.wikipedia.org/wiki/Cardboard_box

Interest: Unusual multifunctional art. It uses innovative 3D perspective.

5.12 Compère - La pomme de terre



A propaganda brochure from 1916 about the nutritional value of potatoes.

- Object: https://en.wikipedia.org/wiki/Brochure
- Extent: 14 pages
- Spatial (made/used): https://en.wikipedia.org/wiki/France
- Creator: http://en.wikipedia.org/wiki/Comité_National_de_Secours_et_d'Alimentation
- Subject:
  - https://en.wikipedia.org/wiki/Potato
  - https://en.wikipedia.org/wiki/Propaganda

Interest: a lot of inferred (not directly appearing) subjects or topics could be added, e.g. World War 1, War-time food advice, etc.

5.13 Weighing Machine for Infants

Science Museum London, through http://wellcomeimages.org/indexplus/image/L0058869.html and http://www.europeana.eu/portal/record/9200105/BibliographicResource_3000006155251.html

- Object: https://en.wikipedia.org/wiki/Weighing_scale
- Subject: https://en.wikipedia.org/wiki/Infant
- Spatial (made): https://en.wikipedia.org/wiki/England
- Created: 1890-1910
- Creator: https://en.wikipedia.org/wiki/Henry_Pooley_%26_Son_Ltd
- Agent (used): http://viaf.org/viaf/257727778 Mellin's Food Company (Boston, Mass.)

Interest: measures the effect of food on infants. Used as advertising device for a pioneering company in the baby food industry. Science Museum London has more artefacts from that company.

- This is the first time we've had to use VIAF (record came from Library of Congress). VIAF is much bigger than Wikidata in terms of persons/orgs, see [Alexiev 2015 sec.1 and sec 2.4.2].
- Mellin's Food was founded by Gustav Mellin in 1866. Google and the Science Museum London say it is UK-based while VIAF/LoC claim it's in Boston. But it's very likely the US company is a branch of the UK company.
- This company is mentioned 5 times in Wikipedia (in relation to advertising and baby foods), but has no article (either nobody has written it yet, or the company is not notable enough)

- It is not in Wikidata either, but we could create it very easily: much less effort than an article, and no restrictive notability criteria.
- IMDB even has a short film about it "Mellin's Food Baby and Bottle (1904)[128]

6 Annex 2: FD in Europeana

In this annex we show a few examples of FD-related CHOs that already exist in Europeana[129] as of Feb 2015. Below are counts of items stemming from a relatively small set of keywords, and a rough estimate how many of the returned CHOs are actually relevant to FD. The initial EFD Challenge page[130] also described several queries for sourcing FD content from Europeana.

keyword	CHOs	notes
food	32948	
drink	10259	
drinken	3233	
wine	9384	wine AND drink: 437
beer	3071	30714 but many are person names, even "bear" misspelt; so we assume 10%
bread	3067	
recipe -medical	1210	many are medical, thus the negation. "recipe book": 467
container	5035	107192 but most are irrelevant, so we assume 5%
dish	8425	some are irrelevant, eg toilet dish, Petri dish, ...
cup	31963	
jar	13852	some are non-food, eg apothecary
jug	7436	but some are coins with such depiction...
krug	1236	6185, but many are irrelevant so we assume 20%
drinkglas	950	
drinkbeker	584	
samovar	276	
cutlery	1150	lots of photos of cutlery production, grinding…
fork -tuning	5589	lots of tuning forks, thus the negation
Total	**74757**	

The results are not orthogonal, so it is not correct to sum them up. But even if you discount for that factor, the numbers are impressive. These numbers have more than **tripled** since Aug 2014 !!!

The FD content sampling[131] provided for the initial EFD challenge is an excellent document that shows a variety of content covering:

- Cook books

[128] http://www.imdb.com/title/tt1092382/
[129] http://europeana.eu/
[130] http://labs.europeana.eu/blog/food-drink-challenge-content-sourcing/
[131] https://docs.google.com/document/d/1zs_B-xOiQ2faCHR4vSPYdAPQvN84uph4WDImNNgF0E4/edit

- Drinking songs (a favourite in any nation!)
- Utensils, such as drinking cups, ancient Greek craters, spouts
- Photographs, e.g. African tribe men eating, Rotterdam's snack culture, Canned pork butchery, Cooking classes, London tower bridge made from sugar
- Advertisements
- Art, including da Vinci's Ultima Cena (the last supper), Adam and Eve (with the apple), The 'Garden of earthly delights' after Hieronymus Bosch
- Infant items, e.g. Infant weighting machine
- Culture and household advice, e.g. The gentlewomans companion
- Groceries, e.g. canned Quaker oatmeal from 1930
- Restaurants

A conservative guess is that between 1 and 10% of Europeana CHOs are related to FD, i.e. between 390k and 3.9M. This is a **lot more** than the 50k objects to be collected in the EFD project, therefore:

- The Semantic Demonstrator should use these existing objects, as well as the EFD objects
- A major goal of the EFD classification is to identify and classify Europeana objects that relate to EFD
- Conversely, keywords appearing in Europeana FD objects can be used to extend the EFD classification. We plan to employ Machine Learning (ML) and Natural Language Processing (NLP) techniques for such extension, as part of work on the Semantic Demonstrator

6.1 Filtering Europeana CHOs

A major concern for any creative/topical application of Europeana CHOs is how to distinguish useful from "useless" objects. For example, the two objects below are not likely to be of interest to anyone but an archaeologist:

- jar fragment[132] from the Petrie Museum of Egyptian Archaeology (UCL)
- cup fragment[133] from the Fitzwilliam Museum, Cambridge, UK

This is not nit-picking, since about half the objects with the keyword "jar" are fragments or shards. Since Europeana does not enforce **object quality** criteria, nor

[132] http://europeana.eu/portal/record/2022347/8BA4652040C28D97167F10C8A07FB03747BCB5B8
[133] http://europeana.eu/portal/record/2022304/1CDFAE9C1AC3F86DE38CAB40B6764324A1CF634F

has **notability** criteria (like in Wikipedia), all kinds of objects that hold only specialist interest have made their way into Europeana.

We hope to tackle this problem in the Semantic Demonstrator by using machine learning approaches. E.g. after searching for "jar" the user is shown all matches, then

- He provides feedback by pointing a few that are useful and a few that are not
- The application performs semantic enrichment of the useful objects against the EFD classification
- The user can correct the matches if disambiguation could not pick the correct match.
- The application augments the score of the classification nodes used in the objects, and their parents (spreading activation)
- The application discovers keywords used in not-useful objects (e.g. "fragment", "shard"), and records these as negatives in the context of the used keywords.

Another problem is the availability of good images. The thumbnails above are not useful for an application. The original page for the left object[134] is currently not available, accessed 24 Feb 2015 (the right object[135] is available). Such concerns are addressed by the Content Reuse Framework developed by Europeana Creative (e.g. the Image checker and the Technical Metadata tool), and the EFD partners should reuse such tools as much as possible.

7 Annex 3: FD in Horniman

The Horniman Museum and Gardens has some concentrated collections of FD items. They can be browsed by object type or subject. Hopefully all of them can be contributed to EFD, and we can try our Machine Learning approaches to augment the EFD Classification. A few of these sub-collections are described below.

URL	Descr	Objects
http://www.horniman.ac.uk/collections/browse-our-collections/authority/subject/identifier/subject-322	Subject: Food and Feasting	4116
http://www.horniman.ac.uk/collections/browse-our-collections/object_type/term-504874	Object: food processing & storage	2555
http://www.horniman.ac.uk/collections/browse-our-collections/object_type/term-504875	Object: food service	1117
http://www.horniman.ac.uk/collections/browse-our-collections/object_type/term-504841	Object: Agriculture and forestry	85
http://www.horniman.ac.uk/collections/browse-our-collections/object_type/term-504883	Object: Hunting, fishing & trapping	684

[134] http://www.ucl.ac.uk/museums/objects/LDUCE-UC47371
[135] http://webapps.fitzmuseum.cam.ac.uk/explorer/index.php?oid=68359

Total		**8557**

It's not quite correct to total the numbers, since the first row (subject) largely overlaps with the rest (object types). We estimate 5k FD objects, which is almost 20% of Horniman's 27886 catalogued objects, a very respectable percentage.

8 References

- [Alexiev 2014a] Vladimir Alexiev, Jutta Lindenthal, and Antoine Isaac. On Compositionality of ISO 25964 Hierarchical Relations (BTG, BTP, BTI). In *13th European Networked Knowledge Organization Systems (NKOS 2014)*, London, UK, September 2014.
- [Alexiev 2014b] Vladimir Alexiev, Joan Cobb, Gregg Garcia, and Patricia Harpring. *Getty Vocabularies Linked Open Data: Semantic Representation*. Manual, Getty Research Institute, 2.0 edition, August 2014.
- [Alexiev 2015] Vladimir Alexiev, Name Data Sources for Semantic Enrichment, part of Europeana Creative deliverable D2.4
- [Aprosio 2013] Alessio Palmero Aprosio, Claudio Giuliano, and Alberto Lavelli. Automatic Expansion of DBpedia Exploiting Wikipedia Cross-Language Information. ESWC 2013
- [Biega 2013] Joanna Biega, Erdal Kuzey, Fabian M. Suchanek. Inside YAGO2s- A Transparent Information Extraction Architecture. WWW 2013
- [de Melo 2010] Gerard de Melo, Gerhard Weikum. MENTA: Inducing Multilingual Taxonomies from Wikipedia. CIKM 2010
- [Erxleben 2014] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez and Denny Vrandecic, Introducing Wikidata to the Linked Data Web, 2014
- [Flati 2014] Tiziano Flati, Daniele Vannella, Tommaso Pasini, Roberto Navigli. Two Is Bigger (and Better) Than One: the Wikipedia Bitaxonomy Project. Association for Computational Linguistics (ACL 2014), Baltimore, Maryland, USA, June 22-27, 2014
- [Giasson 2015] Frédérick Giasson, Michael Bergman. Upper Mapping and Binding Exchange Layer (UMBEL) Specification. v1.20, 16 February 2015
- [Kailus 2014] Angela Kailus (Bildarchiv Foto Marburg). Partage Plus: Enabling Art Nouveau for Europeana, presentation at International Terminology Working Group. Dresden, Germany, Sep 2014
- [Kliegr 2014] Tomáˇ s Kliegr, Ondˇ rej Zamazal. Towards Linked Hypernyms Dataset 2.0: Complementing DBpedia with Hypernym Discovery and Statistical Type Inference. LREC 2014
- [Lehmann 2013] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer, DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 2013.
- [Mahdisoltani 2015] Farzaneh Mahdisoltani, Joanna Biega, Fabian M. Suchanek. Yago3: A Knowledge Base from Multilingual Wikipedias. CiDR 2015
- [Olensky 2012] Marlies Olensky, Juliane Stiller, Evelyn Dröge. Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy. Metadata and Semantics Research 2012. CCIS Volume 343, 2012, pp 252-263, Springer.

- [Paulheim 2013] Heiko Paulheim, Christian Bizer. Type Inference on Noisy RDF Data, ISWC 2013