



europæana
food and drink

Grant Agreement 621023

Europeana Food and Drink

Semantic Demonstrator Delivery

Deliverable number	<i>D3.20</i>
Dissemination level	<i>CO</i>
Delivery date	<i>31 Oct 2015</i>
Status	<i>Final</i>
Author(s)	<i>Vladimir Alexiev (ONTO)</i>



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

Abstract

This document describes the development and delivery of the EFD Semantic Demonstrator. We describe all work performed between 1 April 2015 and 31 October 2015, the achieved results, the created data and enrichments, and the developed application. It is a merging of the periodic progress reports D3.20a (at M18) and D3.20b (at M21), with new information added.

Revision History

Rev	Date	Author	Org	Description
v0.1	29/10/2015	Vladimir Alexiev	ONTO	Initial version
v0.2	30/10/2015	Laura Tolosi	ONTO	Reviewed

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This deliverable contains original unpublished work except where clearly indicated otherwise.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	6
1.1	Structure of the Document.....	6
1.2	Project Team	6
1.3	Abbreviations.....	7
2	Collection Metadata Processing	9
2.1	Metadata Sample Collection.....	9
2.1.1	BG-ONTO (Bulgarian Traditional Recipes from Ontotext).....	10
2.1.2	IT-Alinari (Fratelli Alinari)	11
2.1.3	UK-Horniman (Horniman Museum and Gardens)	11
2.1.4	UK-TopFoto (Topham Partners).....	12
2.1.5	UK-Wolverhampton (Wolverhampton City Council)	12
2.2	BG & EN Metadata Conversion to EDM	13
2.2.1	Conversion Process	13
2.2.2	MINT Handling and Europeana Ingest	15
2.3	Data Directory and SPARQL Endpoint	16
3	Semantic Knowledge Base.....	17
3.1	EFD Classification	17
3.2	EFD Statistics and Queries	19
3.3	Category Tree UI	21
3.4	Statistical Analyses and Visualizations.....	23
3.5	Manual Curation (Internal)	23
3.6	Wikipedia Editing	24
3.7	Refreshing DBpedia	25
3.8	Bottom-up Evidence Propagation	25
3.9	Geonames Parent Places.....	26
3.10	Geonames→DBpedia Links	27
3.11	Geonames→DBpedia Link Quality	27
3.12	sameAs Processing.....	28
4	Semantic Enrichment.....	29
4.1	FD Enrichment.....	29
4.2	Place Enrichment	30
4.3	Enrichment Evaluation.....	31
5	EFD Semapp	34
5.1	Semapp UI Design and Mockup	34
5.2	Semapp Description and Screenshots	34
5.3	Semapp Architecture	38
5.4	Responsive UI	38
6	Dissemination.....	40
6.1	Semapp Website	40
6.2	Task Forces, Workshops	40
6.3	Publications	40
7	Work in Progress and Future Work	42
7.1	Europeana CHO Discovery	42
7.1.1	Tea-Related Objects	42
7.1.2	Restaurants.....	43
7.1.3	FD Classifier.....	44
7.1.4	Europeana Problems.....	45
7.2	Enrichment Web Service	46

7.3 Culture, Ethnicity, Period, Style, Movement	46
7.4 New Language for Enrichment	47
7.5 Handling Lists, Cuisines	47
7.6 Geographical Mapping	48
7.7 AAT-Wikidata Coreferencing	48
7.8 Propagating UMBEL	50
7.9 Propagating Dbtax	50
7.10 Mobile Application	51
8 References	52

Figures

Figure 1 Image of Pounder at Horniman (left & right) vs semapp (right)	12
Figure 2 Simple Mapping Table for Wolverhampton	14
Figure 3 Perl Code for Converting Wolverhampton	14
Figure 4 Complex Mapping Table for Horniman	15
Figure 5 Preview of BG-ONTO Object Shown in MINT	16
Figure 6 Biggest categories reachable from FD: at level 2 (left), at level 5 (right)	18
Figure 7 Category distribution per level, total network (before chopping)	19
Figure 8 Category distribution per level, FD-relevant tree (after chopping)	19
Figure 9 FD Articles Query in Ontotext GraphDB Workbench	20
Figure 10 Category Tree UI	22
Figure 11 Cluster graph of the FD categorization developed with Gephi	23
Figure 12 A Muller from Horniman	24
Figure 13 Decaying propagation	26
Figure 14 Statistics of FD Tags in Horniman Objects	30
Figure 15 FD concepts activated by bottom-up propagation from Horniman terms	32
Figure 16 Semapp UI Mockup	34
Figure 17 Bulgarian traditional recipes	35
Figure 18 Objects From Oceania	36
Figure 19 Object Detailed View	36
Figure 20 Objects Related to Fermented Beverages and Asia	37
Figure 21 Objects from Alinari related to the Roman Empire and Beverages	37
Figure 22 Semapp Conceptual Architecture	38
Figure 23 Responsive UI on Narrow Screen	39
Figure 24 Marker Clusterer Geographic Map	48
Figure 25 Data flow for matching AAT to DBpedia through WordNet and BabelNet	49

Tables

Table 1 Status of Metadata Sample Collection as of 23 June 2015	9
Table 2 Semapp Collections as of Oct 2015	13
Table 3 FD Categories and Parents	21
Table 4 Correct Place Enrichments not in GeoNames Hierarchy	31

Table 5 Incorrect Place Enrichments.....	31
Table 6 Evaluation of Automatic Enrichments	32
Table 7 FD Articles and Labels for Discovery	42
Table 8 AAT to DBpedia matches: EN (AAT) and NL (AATned). * is estimated	49

1 Introduction

This document describes the development and delivery of the EFD Semantic Demonstrator (EFD semapp). We describe all work performed between 1 April 2015 and 31 October 2015, the achieved results, the created data and enrichments, and the developed application. It is a merging of the periodic progress reports D3.20a (at M18) and D3.20b (at M21), with new information added.

- The semapp is available at <http://efd.ontotext.com/app>
- All the data that it uses is at <http://efd.ontotext.com/data> (including the EFD ontology)
- The RDF SPARQL endpoint is at <http://efd.ontotext.com/sparql>

1.1 Structure of the Document

This document is structured in the following sections.

Collection Metadata Processing

- Metadata sample collection
- BG & EN metadata conversion
- Data Directory

EFD Classification and Dataset Processing:

- Building a semantic Knowledge Base
- Building the FD Classification tree, including concomitant statistical analyses
- Creating a FD Tree UI
- Manual curation: internal and in Wikipedia
- Bottom-up relevance propagation

Semantic Enrichment and Evaluation

- FD Enrichment
- Place Enrichment
- Enrichment Evaluation

EFD Semapp

- Semapp design and UI mock-up
- Semapp Description and Screenshots
- Semapp Architecture

Dissemination

- Participation in task forces
- Publications

Work in Progress and Potential future tasks

- Evidence Propagation (UMBEL, DBtax)
- Getty AAT coreferencing
- Ethnic Groups, Cultures, Periods, Styles, Movements
- Discovery of Europeana FD objects
- Geographic Mapping app
- Enrichment in another language besides English

1.2 Project Team

The following project team at ONTO developed the semapp:

- Vladimir Alexiev: project management, requirements, data research and processing, Wikipedia and GLAM engagement

- Andrei Tagarev: customized semantic enrichment, EFD classification development (tree building)
- Laura Tolosi: evaluation, EFD classification, statistical approaches
- Boyan Simeonov: semantic repository, semapp backend
- Rostislav Kirchev: semapp frontend
- Valentin Zhikov: standardized semantic enrichment (ONTO's Concept Extraction Service, CES)
- Joana Tabet: manual data processing

1.3 Abbreviations

Abbrev	Description
AAT	Getty Art and Architecture Thesaurus
API	Application Programming Interface
BG	Bulgaria or Bulgarian language
CH	Cultural Heritage
DBtax	DB Taxonomy
EDM	Europeana Data Model
EFD	Europeana Food and Drink
EN	English language
ESE	Europeana Semantic Elements, XML schema predating EDM
EUROCLIO	European Association of History Educators
FD	Food and Drink
IRI	Internationalized Resource Identifier, a URL with UTF-8 chars
JSON	JavaScript Object Notation
KB	Knowledge Base
LA	Latin language
LIDO	Lightweight Information Describing Objects, a museum object XML schema
NOK	Not OK
OAI	Open Archives Initiative (Protocol for Metadata Harvesting)
RDF	Resource Description Framework, the semantic data format
SPARQL	SPARQL Protocol and RDF Query Language, the semantic query language
TEL	The European Library
UI	User Interface
UK	United Kingdom
UMBEL	Upper Mapping and Binding Exchange Layer
URL	Uniform Resource Locator

Abbrev	Description
UTF-8	The most commonly used Unicode Transformation Format
WMF	WikiMedia Foundation

2 Collection Metadata Processing

2.1 Metadata Sample Collection

The project experienced significant delays in collecting content from the content provider partners, and converting its metadata to EDM. This lack of content was a major obstacle to starting development on the semapp. In order not to block development, ONTO worked with collection providers directly to secure several collections for processing,

We spent a lot of effort to collect metadata samples from most of the content providers. We accepted any metadata format, since we needed the free texts and eventually thesaurus terms. We spent additional effort dealing with the variety of metadata formats, and converted several collections to EDM ourselves.

A snapshot of metadata samples per collection as of June is summarized in the following table.

Table 1 Status of Metadata Sample Collection as of 23 June 2015

Collection (language)	Stat	Img	Obj	Notes
AT-ONB (DE, some LA: not separately marked)	OK	+	30	URLs to OAI records. Good variety: from Latin books to food market photos
BE-CAG (NL)	OK	-	60	Includes thesauri
BE-KMKG (subject: FR, NL, EN; object type: FR)	OK	+	80	Also manually found on museum site.
BG-Onto (BG)	OK	+	9.5k	More than promised. All have images. Some enrichment URLs to DBpedia
CY-CFNM (GR, some EN)	OK	-	40	Some objects have sparse data. No images
ES-CAT (CAT)	OK	+	10k	Different content types
HU-MKVM (EN, HU)	OK	+	53	LIDO records. Most title/descr EN, some terms HU. Geonames & TGN for major places. Encoding & image links fixed
IE-LGMA (EN)	NOK	-	3	Don't seem to be actual records.
IT-Alinari (IT, EN)	NOK	+	50	Many are not FD-relevant. Images are present, though the thumbnail size is quite small.
IT-ICCU (IT, ES, LA)	OK	+	25	EDM RDF URLs (downloaded).
IT-Lombardia ¹ (IT)	OK	+	14k	Not a project partner. Brief and to-the-point descriptions, may be useful for IT machine learning. Got categories we should use for filtering
LT-VUFC (LT)	NOK	-	1	No image
PL-ICIMMS (PL)	NOK	+	1.8k	No FD selection (checked first 3 BIKOP & first 3 PB). Images work but are slow since the size is very large
UK-Horniman (EN)	OK	+	4.3k	Complete records in XML/JSON from Solr API. Also Object Types thesaurus.

¹ http://www.ersaf.lombardia.it/servizi/archiviofotografico/archiviofotografico_en_fase01.aspx

Collection (language)	Stat	Img	Obj	Notes
UK-Wolverhampton (EN)	OK	+	438	59% have images ² , average 3 images per record
UK-TopFoto (EN)	OK	+	32	Keywords (free tags). Also 3 small hierarchical schemes. Images are small previews.

- **Language:** the different languages are very important for semantic enrichment. We need different languages to be demarcated clearly in the record
- **Stat:** shows the status of the collections' sample
- **Img:** shows whether images are available (we have often discovered them on related sites). While images are not used by the enrichment, they are important to display in the semapp
- **Obj:** the number of objects collected.

Details about all collections are available in progress report D3.20a. Below we describe the collections that we ended up using in the semapp.

2.1.1 BG-ONTO (Bulgarian Traditional Recipes from Ontotext)

12379 recipes, all in BG, collected from these sources:

- 6420 recepti.gotvach.bg
- 823 www.gotvetesmen.com
- 5136 www.receptite.com

9483 have images (we submitted only them). Additional work:

- Fixed various URL encoding issues
- Fixed image links for one of the 3 sites, which has changed its image storing system
- Selected only objects with images.
- Removed 411 duplicate files that described the same recipe

The total is 9071 traditional recipes, much bigger than the commitment of 1000. They already include a few enrichments in the metadata, but more is needed (if we can extend the semapp towards handling Bulgarian):

- <http://dbpedia.org/resource/Recipe>
- http://dbpedia.org/resource/Bulgarian_cuisine
- <http://dbpedia.org/resource/Barbecue>
- [http://dbpedia.org/resource/Blanching_\(cooking\)](http://dbpedia.org/resource/Blanching_(cooking))
- http://dbpedia.org/resource/Boiling_in_cooking
- <http://dbpedia.org/resource/Stew>
- http://dbpedia.org/resource/Microwave_oven
- [http://dbpedia.org/resource/Batter_\(cooking\)](http://dbpedia.org/resource/Batter_(cooking))
- <http://dbpedia.org/resource/Baking>
- <http://dbpedia.org/resource/Frying>
- <http://dbpedia.org/resource/Steaming>

² e.g. http://cdn.collectionsbase.org.uk/wagmu/wams/m244_7_p1%20.jpg

2.1.2 IT-Alinari (Fratelli Alinari)

Alinari is a long-time established Italian photo agency

- Available data: 498 objects, provided in English
- Most are photos of paintings and works of art.
- All have images, many are monochrome.
- Many have only a couple of FD-related words, some even without any

2.1.3 UK-Horniman (Horniman Museum and Gardens)

- Available data: 4352 objects in the Food and Feasting subject³ from Solr API⁴ (uses standard Solr syntax)
- Single file in XML (each object in element <doc>) or JSON (add parameter "&wt=json")
- Uses consistent Object Types thesaurus. Uses full place qualification (e.g. "Oceania › Melanesia › New Guinea › Papua New Guinea › Western Province". Most are ethnographic objects
- This collection has the most elaborate metadata
- 3559 objects with images, available in different sizes.

Each object can have multiple images (views). Images are available in 6 sizes:

- <http://www.horniman.ac.uk/media-collection/413/media-413331/preview.jpg>: too small
- <http://www.horniman.ac.uk/media-collection/413/media-413331/body.jpg>: suitable for Europeana preview (edm:object)
- <http://www.horniman.ac.uk/media-collection/413/media-413331/413/media-413331/mid.jpg>: a bit bigger
- <http://www.horniman.ac.uk/media-collection/413/media-413331/feature.jpg>: best for displaying (edm:isShownBy)
- <http://www.horniman.ac.uk/media-collection/413/media-413331/large.jpg>: a bit too large
- <http://www.horniman.ac.uk/media-collection/413/media-413335/413/media-413335.tiff>: maximum size, zoomable. Available only for some views of some objects

The above URLs are suitable for viewing.

- URLs like this are better for image download:
<http://horniman.ac.uk/download/image/media/413/media-413331/body.jpg>
- The Image links in the data are only partial URLs, e.g. /413/media-413331/body.jpg

This is the first collection that we started semantic enrichment for, because it is in English, is quite large, and Horniman uses a thesaurus, which makes the enrichment task a bit easier.

In mid-October Horniman sent us feedback on the mapping, and advised us to download an updated set of objects from their API. However, it was too late to perform this data refresh for the end-Oct delivery. As a result, a few objects in the

³ <http://www.horniman.ac.uk/collections/browse-our-collections/authority/subject/identifier/subject-322>

⁴ <http://collections.horniman.ac.uk/api/solr/select?q=type:object%20AND%20subjectReference:subject-322&rows=5000>

semapp are older than on the Horniman site. E.g. this stone pounder (pestle) has one image in the semapp⁵, and two images at Horniman⁶.



Figure 1 Image of Pounder at Horniman (left & right) vs semapp (right)

2.1.4 UK-TopFoto (Topham Partners)

- 1814 objects, all have images, many are monochrome.
- A lot of keywords, but a moderate number are about FD.
- TopFoto has submitted 6119 objects to the EFD Photo Library and we asked them on 2 Oct 2015 to provide these objects as well⁷
- Image links are clearly marked but are quite small, e.g.
<PhotoURI><http://www.topfoto.co.uk/imageflows/imagepreview/f=EU056268>
</PhotoURI>
- Keywords such as restaurant, canning machine, grapefruit, slimade, diet aid, drinking, drink, eat eating, table, cup of tea
- Keywords form a sort of hierarchy, but it's not usable as a thesaurus
- We got the objects as an Excel table and still have some minor UTF-8 encoding issues (e.g. the copyright symbol © is shown as two symbols)

2.1.5 UK-Wolverhampton (Wolverhampton City Council)

- Available data: 439 objects in a single xml file
- Most are English/Victorian objects
- 260 have images (59%). Number of images: 788. Images per object: 3.04 (for those that have at least one)
- All data is in English although language isn't indicated anywhere
- Encoding is ISO-8859-1 instead of UTF-8

Images

- Files include partial URLs that are resolved against collectionsbase.org.uk, e.g. WAMS/oa76_p1.jpg⁸
- Images resolve regardless of upper/lowercase and forward/backward slash, e.g. WAMS\op231.jpg is the same as wams/OP231.jpg
- Server handles spaces in filename, e.g. wams/m244_7_p1%20.jpg

⁵ <http://efd.ontotext.com/app/resource/http%253A%252F%252Fefd.ontotext.com%252Faggregation%252Fhttp%253A%252F%252Fcollections.horniman.ac.uk%252Fobjects%252F67347?query=pestle&limit=24&offset=0>

⁶ <http://www.horniman.ac.uk/collections/browse-our-collections/object/67347>

⁷ https://basecamp.com/2069212/projects/8450098/messages/39521744#comment_337565031

⁸ http://cdn.collectionsbase.org.uk/wagmu/wams/oa76_p1.jpg

2.2 BG & EN Metadata Conversion to EDM

Since English is the first language to be tackled by the semapp, we ended up converting several English collections to EDM, to be used internally by the semapp. We presented the results to the respective content partners to decide whether they want to submit this EDM or use a different channel. (Alinari did their conversion using MINT). We also converted the BG-ONTO collection to EDM.

The next table shows the number of English-language objects used by the semapp as of Oct 2015.

Table 2 Semapp Collections as of Oct 2015

Collection	Obj	Notes
BG-ONTO	9071	Traditional recipes, all have images ⁹ . Much bigger than the commitment of 1000. Include a few enrichments, which however are not the result of NLP processing
IT-Alinari	498	All have images ¹⁰ , many are monochrome. Most are photos of paintings and works of art. Many have only a couple of FD-related words, some even without any.
UK-Horniman	4352	3559 with images ¹¹ , available in different sizes. Uses consistent Object Types thesaurus. Uses full place qualification (e.g. "Oceania › Melanesia › New Guinea › Papua New Guinea › Western Province". Most are ethnographic objects. This is the most developed collection
UK-Wolverhampton	439	260 have images ¹² . Most are English/Victorian objects.
UK-TopFoto	1814	All have images ¹³ , many are monochrome. A lot of keywords, but a moderate number are about FD.
Total	16174	We asked TopFoto for the total 6119 objects in the EFD Photo Library, which would increase the semapp set to 20479

2.2.1 Conversion Process

We did the conversion using simple Perl scripts.

- First we input the data using functions such as XML::Simple->XMLin, JSON::XS->decode_json, or split (for simple TSV).
- Then we determine the fields to be mapped by counting & analysing all input fields, then agreeing a mapping table with the provider, e.g. like this

⁹ <http://gradcontent.com/lib/600x350/bean-salad.jpg>

¹⁰ e.g. <http://images.alinari.it/img/480/ACA/ACA-F-022924-0000.jpg>

¹¹ e.g. <http://www.horniman.ac.uk/media-collection/413/media-413331/feature.jpg>

¹² e.g. http://cdn.collectionsbase.org.uk/wagmu/wams/m244_7_p1%20.jpg

¹³ e.g. <http://img04.pars04.fr.topfoto.co.uk/imageflows/imagepreview-if3/t=topfoto&f=EUFD001241>

Xpath	Count	Distinct	Length	Examples	Map to
AcquisitionDate	225	71	10		dc:contributor (qual)
AcquisitionMethod	264	5	5.7	Gift; Untraced find	dc:contributor (qual)
AcquisitionNote	84	49	67.3	by contribution from Joseph...; 159,	dc:description
AcquisitionSource	256	66	16.7	Bantock Kate P, Mrs	dc:contributor
Artist	35	22	28.2		dc:creator
AssociatedActivity	177	13	11.9	Tea drinking	dc:subject
AssociatedConcept	46	21	7.8	Historic & Baskets & Motherhood	dc:subject
Colour	194	49	8.3		dc:format (color)
Copyright	1	1	14	Frank Brangwyn	dc:rights
CreditLine	31	5	69.1	Thanks .. for help with photography;	dc:description
Description	136	130	504.6	This okimono is carved...	dc:description
Dimensions	353	314	24		dc:extent
Inscription	9	9	39.9	Signed; G.B. O'Niell 67	dc:description
Keyword	52	28	8.9	India; everyday things; Second World	dc:subject
Maker	126	37	20.5		dc:creator
Material	243	45	7.1		dc:medium (material)
ObjectName	449	89	7.1	Container	dc:title (qual)
ObjectNumber	438	432	4.8		dc:identifier
ObjectProductionDate	319	150	10.4	1769 - 1784	dc:date
ObjectProductionNote	15	9	201.4	The company was formed...; The Te	dc:description
ObjectProductionPeriod	309	11	21.2	Georgian (1714-1837)	dc:temporal
ObjectProductionPlace	158	38	8.7	India	dc:spatial

Figure 2 Simple Mapping Table for Wolverhampton

- Then we implement the mapping by fetching fields from the converted object, and putting them into an RDF::Trine graph (model):

E.g. the script for converting UK-Wolverhampton is largely shown below (this is only the ProvidedCHO node, a few more statements make the Provider Aggregation).

```

$rdf->assert_resource ($cho, "rdf:type", "edm:ProvidedCHO");
$rdf->assert_literal ($cho, "edm:type", "IMAGE");
assert_literal ($cho, "dc:creator", $obj->{Artist});
assert_literal ($cho, "dc:creator", $obj->{Maker});
assert_literal ($cho, "dc:date", $obj->{ObjectProductionDate});
assert_lang_literals ($cho, "dc:description", $obj->{Description});
assert_lang_literal ($cho, "dc:description", $obj->{PhysicalDescription});
assert_lang_literal ($cho, "dc:description", $obj->{Inscription});
assert_lang_literal ($cho, "dc:description", $obj->{CreditLine});
assert_lang_literal ($cho, "dc:extent", $obj->{Dimensions});
assert_lang_literal_with_qualifier ($cho, "dc:format", $obj->{Colour}, "color");
assert_literal ($cho, "dc:identifier", $obj->{RecordID});
assert_literal ($cho, "dc:identifier", $obj->{ObjectNumber});
assert_literals ($cho, "dc:identifier", $obj->{OtherNumber});
assert_literal ($cho, "dc:identifier", $obj->{RCN});
assert_lang_literal_with_qualifier ($cho, "dc:medium", $obj->{Material}, "material");
assert_lang_literal_with_qualifier ($cho, "dc:medium", $obj->{Technique}, "technique");
assert_lang_literal ($cho, "dc:rights", $obj->{Copyright});
assert_lang_literals ($cho, "dc:spatial", $obj->{ObjectProductionPlace});
assert_lang_literals ($cho, "dc:subject", $obj->{AssociatedActivity});
assert_lang_literals ($cho, "dc:subject", $obj->{AssociatedConcept});
assert_lang_literals ($cho, "dc:subject", $obj->{Keyword});
assert_lang_literals ($cho, "dc:subject", $obj->{Subject});
assert_lang_literals ($cho, "dc:subject", $obj->{Term});
assert_lang_literals ($cho, "dc:temporal", $obj->{ObjectProductionPeriod});
assert_lang_literal_with_qualifier ($cho, "dc:title", $obj->{Title}, $obj->{ObjectName});
assert_lang_literal ($cho, "dc:type", $obj->{UserText1});

```

Figure 3 Perl Code for Converting Wolverhampton

This takes care to emit proper language tags ("bg" or "en" for these collections), field multiplicity and optionality.

The most complex mapping is for Horniman. Out of **303 fields** in their collection management system, we mapped 82 fields, which provides very rich metadata. E.g. the beginning of the mapping table is shown below. The numbers on the right show in how many objects does the field occur, and in case of multivalued fields, the distribution of the number of values. This was important knowledge that informed our mapping.

field	map to	comment	example	occ														
agentReference	MAYBE	edm:Agent, e.g. h	agent-5955	33	33													
agentRelation	dc:contributor (qual)	add as (qualifier)	maker of	15	15													
agentString	dc:contributor		Mahillon & Co	33	33													
bodyMediaLocation	edm:object [0]	size just right. Use	/151/media-151228/body.jpg	3558	1512	1285	507	169	51	25	4	2						1
category	dc:type		Aerophone	25	25													
collection	dct:isPartOf		Anthropology	4350	4350													
collectorEndDate	MAYBE	to map this need	1979	54	3	51												
collectorRelation	dc:contributor (qual)		collector	392	339	53												
collectorStartDate	MAYBE	to map this need	1978	60	9	51												
collectorString	dc:contributor		Beek, Gosewijn van	396	343	53												
created	MAYBE	creation date of r	2005-01-06T00:00:00Z	4351														
creditLine	dc:rights	and always "Horn	Dato Erik Jensen collection	234	234													
culture	dc:creator	qualifier "culture"	Chimu	1134	980	105	5	44										
cultureArea	dct:spatial		Western Province, Papua New Gu	60	6	10		44										
cultureRelation	dc:creator (qual)		maker or user	1006	854	103	5	44										
cultureTermRelation	dc:creator (qual)		maker or user	1002	851	102	49											
cultureTermString	dc:creator	mostly different f	Yunca	1131	978	104	49											
dateCollected	dc:date		1978 - 1979	125	125													
dateCollectedMethod	dc:date (qualifier)		fieldwork collection	51	51													
dateCollectedRelation	dc:date (qualifier)	emitted always a	date collected	124	124													
dateMade	dc:date		19th-20th century	836	714	118	4											
dateMadeEra	dc:date		Han Dynasty	26	22	3	1											
description	MAYBE	most are poorer t	Round shallow porcelain pot whic	4326	1049	1002	1439	596	150	60	10	7	6	1				
exhibitionString	dc:description	qualifier "exhibit	OIF : Romanian Ceramics	84	82	2												
featureMediaLocation	edm:isShownBy [0], ex	and edm:WebRes	/151/media-151228/feature.jpg	3558	1512	1285	507	169	51	25	4	2						1

Figure 4 Complex Mapping Table for Horniman

2.2.2 MINT Handling and Europeana Ingest

The conversion of Bulgarian Traditional Recipes (ONTO) was done with a Perl script, but simpler than the ones described above. Since Europeana prefers to get the data from MINT rather than a zip file, we had to remap this EDM into MINT EDM, which uses a fixed order of fields. That is a trivial mapping that just copies fields from one XML to another, but still took time to develop.

As a benefit, we could see a preview of our objects in MINT

The screenshot displays the Europeana Data Upload Items interface. On the left, there's a 'Preview Options' panel with a search bar and a list of items. The main area shows a preview of a recipe item titled 'Бърз чомлек със свинско' (Quick chomlek with pork). The item includes a description, ingredients, and a list of related items. The interface is in Bulgarian and shows various metadata fields like title, creator, date, and publisher.

Figure 5 Preview of BG-ONTO Object Shown in MINT

We submitted these recipes in Jul and Europeana attempted to ingest them in Oct. However, this uncovered a serious bug in Europeana ingestion¹⁴: it converts Cyrillic characters in IRIs to an underscore. For example, both of these record IRIs

- <http://www.receptite.com/рецепта/пържени-яйца-с-доматен-сос>
- <http://www.receptite.com/рецепта/супа-от-пиле-със-зеленчуци>

get converted to /2059509/_____ (which is the same length and leads to object ID duplication). Europeana is working to fix this bug. Furthermore, it is unclear how to deliver enrichments to Europeana, because:

- Neither ONTO nor NTUA can add enrichments in provider collections in MINT
- Europeana cannot take enrichments for a number of objects at once as a single data file

We raised this question in Jul 2015¹⁵ and the partners are still looking for the easiest solution. This question is of critical importance for the Crowdsourcing Enrichment application (to be developed by D3.5 Technical Demonstrator and T5.2 Community/crowdsourcing platform).

2.3 Data Directory and SPARQL Endpoint

All data that the semapp uses (both collections described in this section, and datasets described in the next section) is available.

- It can be downloaded from <http://efd.ontotext.com/data>. This includes the EFD ontology¹⁶. We are still working on a detailed description of the files (should be completed in very early Nov 2015).
- The data can be queried at the RDF SPARQL endpoint <http://efd.ontotext.com/sparql> that uses the Ontotext GraphDB semantic repository.
- Note: the same software hosts the Europeana LOD (43M objects) at <http://europeana.ontotext.com>

¹⁴ <https://europeanadev.assembla.com/spaces/europeana-ingestion/tickets/1872>

¹⁵ <https://basecamp.com/2069212/projects/7205992/messages/45430278>

¹⁶ <http://efd.ontotext.com/data/efd-ontology.ttl>

3 Semantic Knowledge Base

The first task to enable semantic enrichment was to create a semantic Knowledge Base (KB) (see [Alexiev 2015a sec.2.2]). We completed the following tasks:

- Installed and configured Ontotext GraphDB and a SPARQL endpoint: <http://efd.ontotext.com/sparql>
- Loaded EN and IT DBpedia, including all articles, labels, categories and category assignments. We have not yet started on IT semantic enrichment, but we wanted to have a second DBpedia in order to evaluate approaches for multilingual category fusion (see [Alexiev 2015a] sec. 2.3.2). Refreshed DBpedia on 20 Oct 2015.
- Loaded Geonames and Geonames-DBpedia links in order to develop a coherent Place hierarchy. It turns out that DBpedia doesn't have such consistent info (e.g. it doesn't include a statement that Bulgaria is in Europe), whereas Geonames has the property `gn:parentFeature` that creates a uniform place hierarchy.
- Performed a number of data corrections and enhancements, e.g. to Wikipedia categories and Geonames parent places (see later)
- Developed a small EFD ontology to hold classification data (e.g. FD-relevant parent category links, "not relevant" judgements, scoring counts, etc).

3.1 EFD Classification

Elaborating the EFD Classification by refinement of the FD categories is the main task enabling the semapp. See technical details in [Alexiev 2015a] and [Tagarev 2015].

As explained in [Alexiev 2015a sec.3.8.3], starting from the root category `Food_and_drink`, one reaches 887k categories, over 26 levels deep, representing 80% of all categories. Most of these are irrelevant to FD. As shown below, all the top 10 most populous categories at level 5 are irrelevant (e.g. Oceanography, Water pollution, Physical exercise, Bodies of water, Natural materials, Country planning in the UK, etc). The reason is "semantic drift": since the meaning of the Wikipedia "parent category" relation is not well-defined, the longer path one follows, the harder it becomes to see any logical connection between the two categories (ancestor and descendant).

D3.20 Semantic Demonstrator Delivery

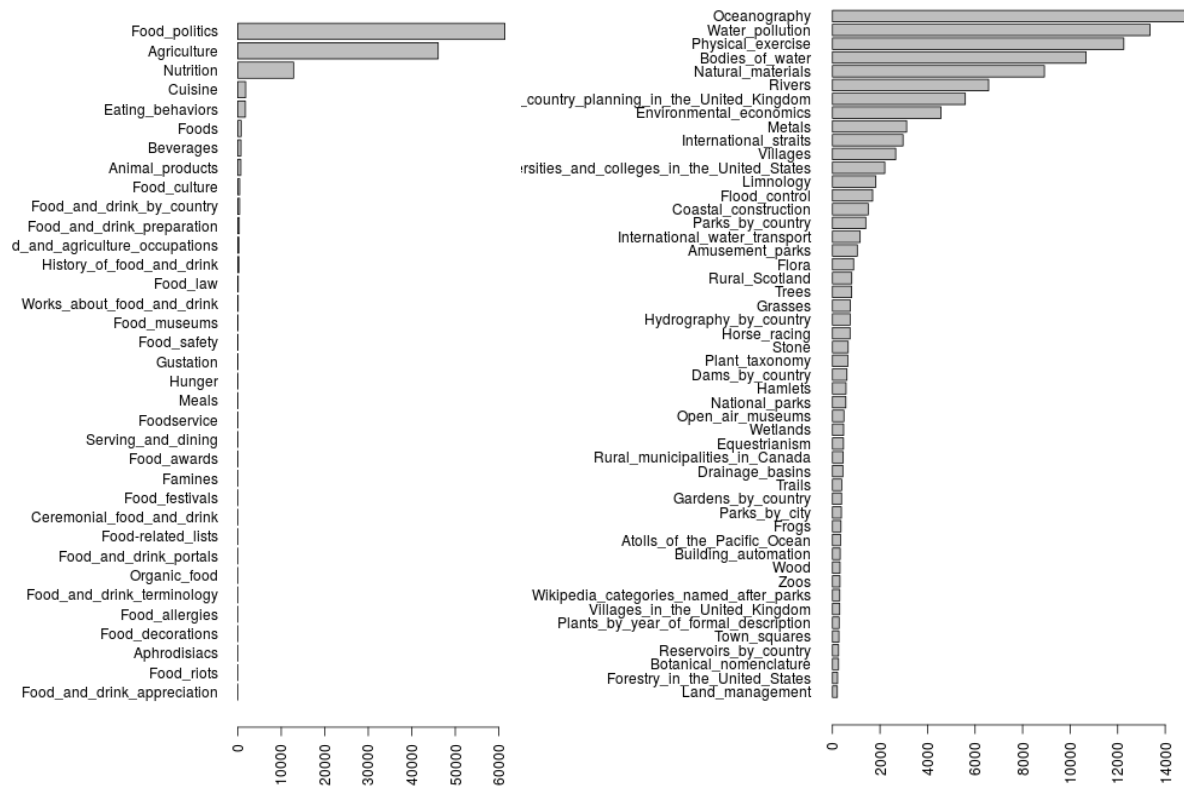


Figure 6 Biggest categories reachable from FD: at level 2 (left), at level 5 (right)

We developed algorithms and software to work with the Wikipedia categorization to build a FD-relevant classification tree. The software is developed in Java, using the Sesame API and Ontotext GraphDB to store the data. The software is reusable, including for domains other than FD. This relies on:

- Statistical analysis of the category network
- A tool for manual curation of the tree (chopping out irrelevant branches)
- Evidence-based feedback

As a result we were able to reduce the categories by 98%: from 880k to **14.7k FD-relevant categories**. This excellent result was achieved by removing only **366** categories and their connections (blacklist).

In addition to improving relevance, the chopping has reduced the distance to the root, confirming the hypothesis that long chains have a lower chance to be meaningful/relevant. As shown on the following figures, the mode of the minimum distance to root was reduced from 16 to 5:

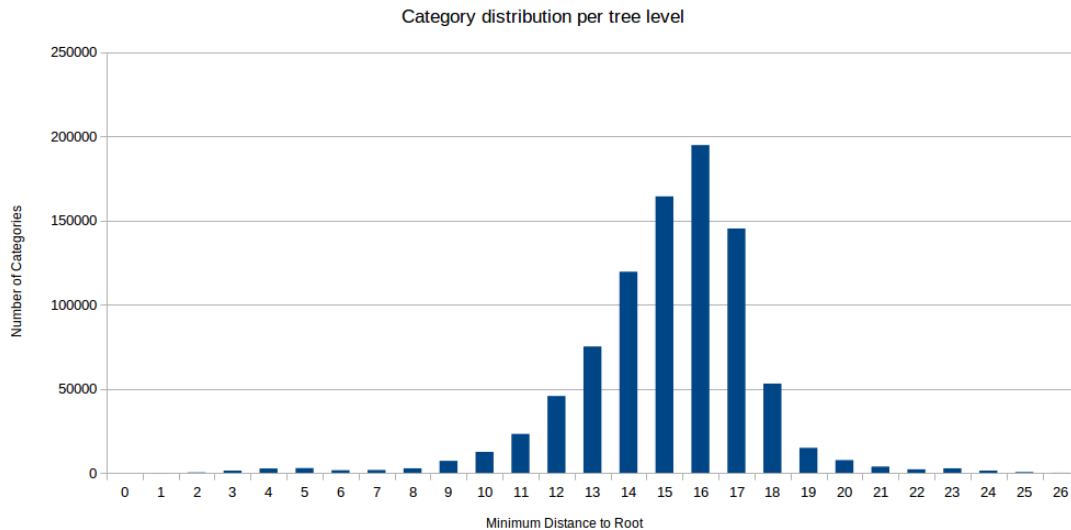


Figure 7 Category distribution per level, total network (before chopping)

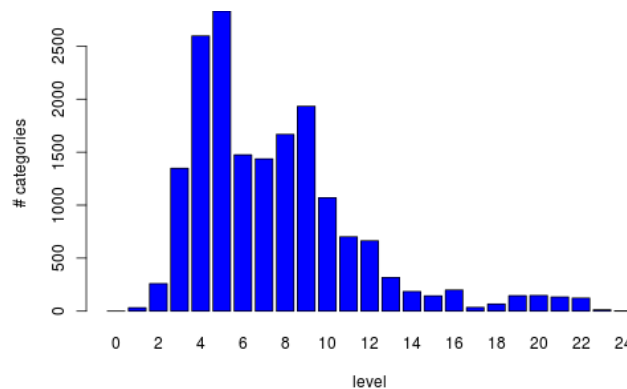


Figure 8 Category distribution per level, FD-relevant tree (after chopping)

The relevant categories include **201k FD-relevant articles**. This confirms our bet that Wikipedia is the largest dataset with FD-relevant items. This number was revised several times due to the following:

- On one hand, removed further irrelevant categories. E.g. Agricultural universities and colleges in the United States includes 58605 articles. But since pretty much any large university has an Agriculture department, this huge list is not really relevant to the topic (see Figure 10). We are currently considering the removal of category Nutrition, which includes a lot of irrelevant articles (e.g. Iron as a micronutrient) or barely relevant (e.g. Eating Disorders).
- Evidence feedback (processing articles or CHOs that are proven FD-relevant by other means) enlarged the tree. E.g. Horniman has a lot of Hunting objects; Hunting was not part of the FD hierarchy in Wikipedia but we added it.

3.2 EFD Statistics and Queries

As of late Oct 2015, we have these statistics:

- FD categories: 14695
- FD articles: 200956
- cat<cat relations (parent categories): 21008: 1.58 per cat
- art<cat relations (categorizations): 312364: 1.56 per art, 21.2 per cat

Let's use the EFD SPARQL endpoint to make a couple of simple queries. We use the [Ontotext GraphDB Workbench](#) to manage queries (load, save), prefixes, autocomplete class and property names, etc:

The screenshot shows the Ontotext GraphDB Workbench interface. The top bar includes the logo, 'GraphDB Workbench', and tabs for 'Data', 'SPARQL', and 'Admin'. The 'Current Repository' is set to 'efd'. Below the tabs, there are tabs for 'Tea labels', 'count', 'FD arts', 'FD cats', and a '+' button. The main area contains a SPARQL query editor with the following query:

```
1 PREFIX efd: <http://data.foodanddrinkurope.eu/ontology#>
2 PREFIX dct: <http://purl.org/dc/terms/>
3 select * {
4   ?art dct:subject ?cat.
5   ?cat efd:child ?parent
6 } limit 100
```

On the right side of the query editor, there are checkboxes for 'disable sameAs' (checked) and 'Include inferred' (checked). A 'Run' button is at the bottom right. Below the query editor, there are tabs for 'Table', 'Raw Response', 'Pivot Table', 'Google Chart', and a 'Download as' button. The 'Table' tab is selected, showing the query results. The results are displayed in a table with two columns: 'cat' and 'parent'. The table shows 5 rows of results, all with a limit of 100. The query took 0.32 seconds.

	cat	parent
1	dbc:Confectionery	dbc:Desserts
2	dbc:Pies	dbc:Desserts
3	dbc:Ice_cream	dbc:Desserts
4	dbc:Chocolate	dbc:Desserts
5	dbc:Cakes	dbc:Desserts

Figure 9 FD Articles Query in Ontotext GraphDB Workbench

Count distinct FD-relevant articles and categories:

```
select (count(distinct ?art) as ?a) (count(distinct ?cat) as ?c)
from onto:disable-sameAs {
  ?art efd:subject ?cat
}
```

Get FD-articles with their categorizations:

```
select *
from onto:disable-sameAs {
  ?art dct:subject ?cat.
  ?cat efd:child ?parent
} limit 100
```

Get FD-relevant categories with their parents.

```
select *
from onto:disable-sameAs {
  ?cat efd:child ?parent
} limit 100
```

It returns results like:

Table 3 FD Categories and Parents

16	dbc:Yogurts	dbc:Desserts
17	dbc:Dessert_templates	dbc:Desserts
18	dbc:Biscuits_(British_style)	dbc:Desserts
19	dbc:Pastries	dbc:Desserts
20	dbc:Dessert-related_lists	dbc:Desserts
21	dbc:Sugar_confectionery	dbc:Desserts
22	dbc:Apples	dbc:Fruit
23	dbc:Melons	dbc:Fruit
24	dbc:Fruit_juice	dbc:Fruit
25	dbc:Pears	dbc:Fruit
26	dbc:Peppers	dbc:Fruit
27	dbc:Citrus	dbc:Fruit
28	dbc:Fruit_and_vegetable_characters	dbc:Fruit

The last category¹⁷ is a curious one, including characters like Mr Potato Head, Cipollino and Bananaman.

3.3 Category Tree UI

We developed a tool to view and manipulate the tree. The tool is written in JavaScript and uses the software described in the previous section as backend server, communicating with it in JSON. The tool is currently not available outside of the ONTO network, until we secure the DELETE operation with username and password (it deletes whole branches of the tree without confirmation). A screen-shot is shown below.

¹⁷ https://en.wikipedia.org/wiki/Category:Fruit_and_vegetable_characters

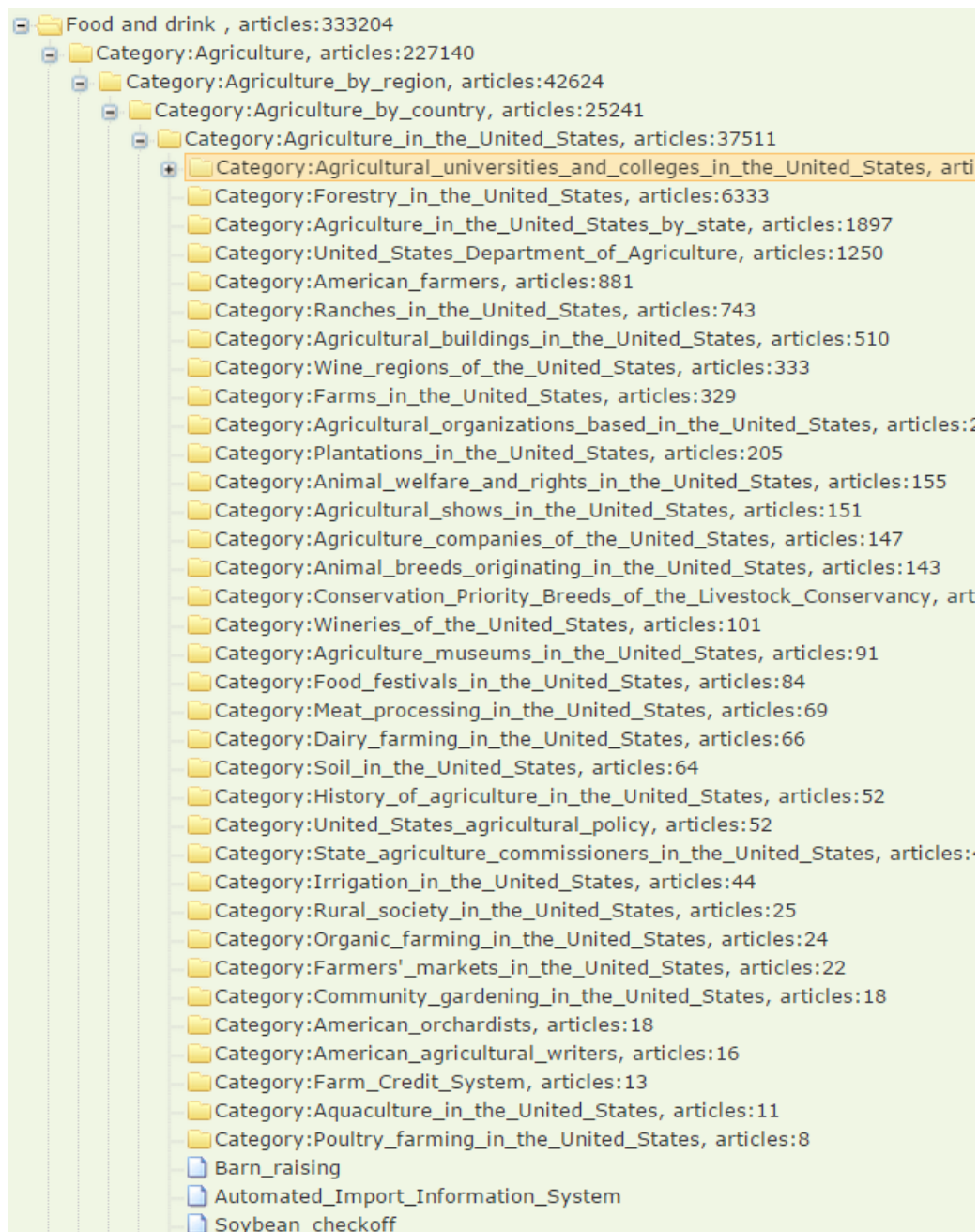


Figure 10 Category Tree UI

The numbers require some explanation. They represent the number of articles per category, prorated to each parent at each level. This may lead to some counter-intuitive numbers. E.g. opening the first branches we see this (remember that Agriculture is not entirely FD-relevant and is subject to further chopping):

- Food and drink: 333204
- Agriculture: 227140
- Agriculture_by_region: 42624
- Agriculture_by_country: 25241
- Agriculture_in_the_United_States: 37511
- Agricultural_universities_and_colleges_in_the_United_States: 58605

The reason that the number in the last line is bigger than the previous line is this: most of the 58k universities/colleges have several category parents, only a few of

which connect to the FD root. So only part of the number 58k contributes towards the number 37k.

3.4 Statistical Analyses and Visualizations

We performed a number of statistical analyses and visualizations that guided our work on the classification tree. For example, sorting categories by prominence so the most populous can be processed first, testing various processing hypotheses, calculating Precision and Recall etc.

We developed data analyses and visualizations using R, Gephi and Excel. All charts and graphs in this document are produced with these tools. The software developed is not reusable.

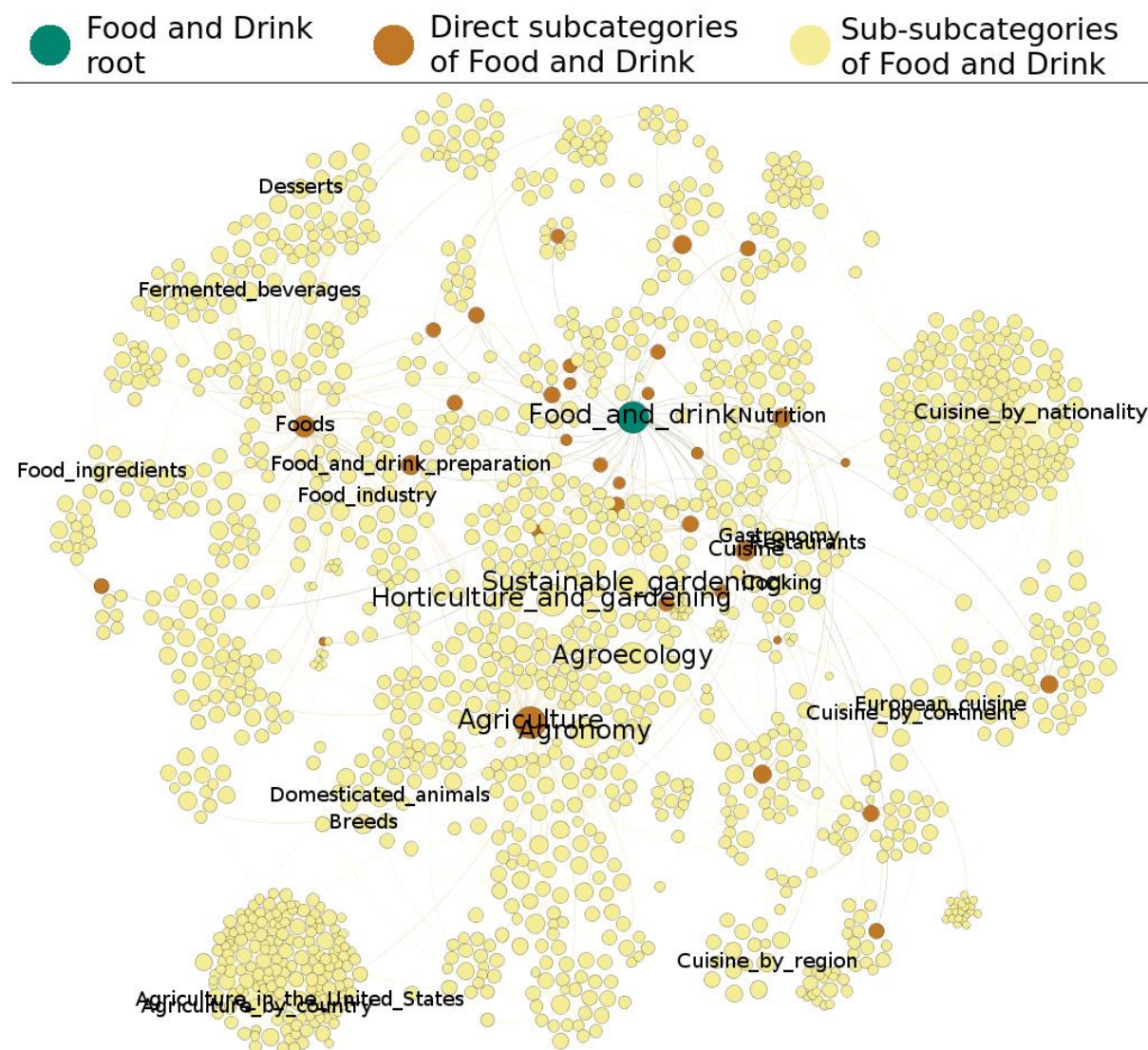


Figure 11 Cluster graph of the FD categorization developed with Gephi

3.5 Manual Curation (Internal)

Significant manual curation work was performed, for example:

- Cutting down the tree (curation to cut out irrelevant parts)
- Assembling black-list of words. For example, the Horniman thesaurus includes object types "X Model" (e.g. "Thresher model"). We recognize "Thresher" correctly, but map "model" to "person who promotes, displays, or advertises commercial products". Since mapping to "X" is good enough, we black list "model"

- Assembling black-list of phrases that appear in every object of a collection. While correct, they don't add information content. (This blacklist is still not applied in the enrichment pipeline). For example:
 - Every Horniman object has the phrase "Food and Drink"
 - Every Horniman object has the phrase "Food and feasting"
 - Every TopFoto object has the phrase "Europeana Food and Drink"
- Adding additional roots (related to Hunting and hunting weapons)
- Manual matching of some Horniman terms

3.6 Wikipedia Editing

Since we use a semantic representations of Wikipedia (DBpedia) as our main source of classification, in many cases the best course of action is to add missing categories and labels to Wikipedia to improve its FD coverage. See contribution list.¹⁸

- **Adding new categories**, e.g. "Libation: ceremonial pouring of water, wine, olive oil, etc". Added to parent categories "Wine" and "Olive oil"
- **Adding parent categories**. Added major branches under FD: Hunting, Fishing, and Livestock. (The Horniman collection has a lot of Hunting objects). E.g. "Bottles" did not have parent "Drink containers". Since most bottles are used in this way, we added it.
- **Adding categories to articles**. E.g. "Gourd" did not have category "Bottles" ("Calabash" or "Bottle gourd" had that category). But since most gourds (even those not of the "Bottle gourd" variety) are used as primitive bottles, we added category "Bottles". E.g. added "Libation sticks", "Rhyton" and "Patera" to "Libation"
- **Adding labels (redirects)**. E.g. "Muller"¹⁹ is a copper device for mulling beer or wine or keeping them warm. We added it as a redirect to "Mulled wine". Even though it represents a different concept, such use of redirects is legitimate and widely used on Wikipedia.



Figure 12 A Muller from Horniman²⁰

- **Creating pages**. E.g. "Shepherd's crook" and "Tumbler (glass)" by splitting text from existing pages. Added label with qualifier "Crook (shepherd)" because Horniman uses this term (without the qualifier).

¹⁸ https://en.wikipedia.org/wiki/Special:Contributions/Vladimir_Alexiev

¹⁹ <http://www.horniman.ac.uk/collections/browse-our-collections/authority/term/identifier/term-503368>

²⁰ <http://www.horniman.ac.uk/object/25.29>

- **Small additions to pages.** E.g. added to "Leash" the note "Leashes are often used to tether domesticated animals left to graze alone" as justification for adding the category "Livestock". E.g. added section "Lovespoon#Wedding_Spoons"
- **Adding references:** to Horniman, Etsy, Gilding, Popular Mechanics to a number of pages, e.g. "Tableware#Place_markers", "Scotch_hands", "Roasting_jack#Bottle-jack", "Lovespoon#Wedding_Spoons", "Corn_on_the_cob" ("Corncob holder from wood made in Kenya").
- **Adding illustrations,** e.g. a phiala from the Panaguyrishte gold treasure (Used in ceremonial wine drinking or Libation) to article "Patera". Unfortunately we couldn't add illustrations from Horniman because the image license of that museum does not allow it.

Wikipedia "edit wars"

E.g. "Cord attacher" is a primitive device for attaching a cord to the rod, or splicing two cords together. It appears often in ethnology museums (e.g. see Horniman²¹ and Burke Museum²²). We added it as a section to article "Fishing Tackle", and as a redirect to that specific section. But our edit was reverted²³ with the comment "not a term commonly used in fishing" and then a suggestion²⁴ to add to article "History of fishing". This matches our conclusion in [Alexiev 2015f] that it's harder to add to Wikipedia than Wikidata, since one needs to learn to work with the editorial guidelines and community of Wikipedia, while Wikidata editing still is not "hindered by bureaucracy".

3.7 Refreshing DBpedia

To ensure that our Wikipedia edits are reflected in the EFD KB, we got a fresh Wikipedia extract on 20 Oct 2015 and regenerated DBpedia from it, using the open source DBpedia Extraction Framework. Since this dataset is very large, it's not yet at the EFD Data directory but can be added on request.

3.8 Bottom-up Evidence Propagation

The top-down tree formation & cleaning described above is only one of the ways in which we elaborate the EFD classification. The other way is based on evidence (contribution) propagation or positive feedback that can confirm and enlarge the tree. There are two general sources:

- Cultural objects and thesauri that are submitted as FD-related and are classified with some articles
- Articles that are proven FD-related through other means, e.g. class Food in DBpedia [Alexiev 2015a sec.3.11.1] UMBEL FD super-type [Alexiev 2015a sec.3.19], DBtax (see below)

We processed these sources:

- **Horniman objects:** we propagated the evidence from Horniman objects and made sure all are present in the FD hierarchy, in many cases enlarging the tree by editing Wikipedia
- **dbo:Food:** there are 6643 en.wiki articles using appropriate Infoboxes (e.g. Prepared Food or Beverage) that are reflected in DBpedia with the class dbo:Food. We checked them against the FD tree: 6520 of them were already in

²¹ http://www.horniman.ac.uk/collections/browse-our-collections/object_type/term-504068

²² <http://collections.burkemuseum.org/ethnology/display.php?ID=46106>

²³ https://en.wikipedia.org/w/index.php?title=Fishing_tackle&type=revision&diff=667166404&oldid=667165709

²⁴ https://en.wikipedia.org/wiki/Talk:Fishing_tackle#Cord_Attacher

the tree and 123 were not. We added the appropriate ones to the tree by adding or adjusting categories.

We implemented two scoring approaches; one is described in [Tagarev 2015 p.6]:

- Decaying propagation of the evidence (contribution) in a bottom-up fashion, following the FD category structure. The contribution decreases with path length going up

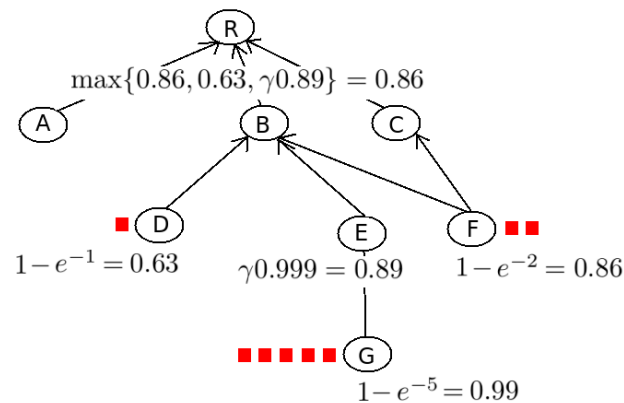


Figure 13 Decaying propagation

- Integral (whole-number) propagation but only towards the FD root, following shortest paths or paths that are longer by a fixed factor

These approaches can serve these purposes:

- Discover new categories that are currently outside the tree but are judged relevant (e.g. Spears as hunting weapons)
- Provides a relevance scoring that can be used for ranking of search results
- After a critical number of CHOs have been enriched and mapped to the tree, one can expect that large categories with very low score (or bottom-scoring categories) should be removed, because of lack of evidence.

3.9 Geonames Parent Places

For the Place semantic facet, we needed a coherent Place hierarchy. Surprisingly, it turns out that DBpedia doesn't have a consistent place hierarchy:

- There is no uniform place hierarchy property. E.g. for `dbo:Island`, the property `dbo:archipelago` shows the island group (physical parent), `dbo:location` is the containing ocean or sea, and `dbo:country` is the owning country (administrative parent). For cities there is `dbo:region` and `dbo:country`.
- There is no property stating that Bulgaria and France are part of Europe. (They belong to several YAGO classes such as "European country" or "European Union country", but there is no statement relating them to `dbr:Europe`. (We cannot fish out all related classes and correlate them to continents and other key places.)

A specific example: `dbr:Andaman_Islands` has:

- `dbo:archipelago dbr:Andaman_and_Nicobar_Islands` (parent, administrative)
- `dbp:countryAdminDivisions dbr:Andaman_and_Nicobar_Islands` (parent, administrative)
- `dbo:location dbr:Bay_of_Bengal` (parent, physical)
- `dbo:country dbr:India` (ancestor, administrative)
- `dbo:capital dbr:Port_Blair` (child, administrative: city)

- `dbo:majorIsland dbr:North_Andaman_Island, dbr:South_Andaman_Island` (child, administrative). Partial inverse of `dbo:archipelago`

In contrast, Geonames has the property `gn:parentFeature` that creates a uniform place hierarchy, so we decided to use GeoNames.

3.10 Geonames→DBpedia Links

Since the ONTO CES (Concept Extraction Service) finds DBpedia places, we needed to first link Geonames to DBpedia.

- GeoNames has coreferences (links) to other datasets²⁵, of which 482k are links to Wikipedia (470k to `en.wikipedia`, 10k `ru`, 0.6k `de`).
- `dbo:Place`²⁶ (the root of the DBpedia place hierarchy) has 167 subclasses. DBpedia has 756k places (resources falling in that type hierarchy), excluding `CelestialBodies`.
- Therefore 62% of `en.dbpedia` places are linked to GeoNames.
- (GeoNames is much bigger with over 9M places. So only 5.2% of GeoNames features are linked to `en.dbpedia`)

Later we found out that Wikidata includes a bigger number of links:

- 625746 Wikidata items have a property GeoNames ID²⁷, which can be checked with a WDK query²⁸
- This is 34% more than the Geonames links. The main reason is that this number is across all Wikipedias while the links described above are to `en.wikipedia` only.
- (Note: WD has parent place links, but they are not uniform and are still quite incomplete)

It would have been better to use the Wikidata→Geonames mapping, but we found it too late. Nevertheless, the Geonames→Wikipedia mapping is quite sufficient for our needs, see sec 4.

3.11 Geonames→DBpedia Link Quality

- The 470738 Geonames→English Wikipedia links are manually curated at the Geonames site (in table Alternative Names). From these links we created DBpedia `sameAs` statements using a DBpedia script²⁹. We call this the "big link set".
- There are also 86547 links generated by a DBpedia contributor through automatic matching using SILK³⁰. We call this the "small link set".
- 79068 places (17%) appear in both link sets, 379432 (81.4%) only in the bigger set, and 7479 (1.61%) only in the smaller set, for a total of 465979 DBpedia places linked to Geonames.
- We estimated the precision of the small link set to 85% (out of 25 samples, 4 are wrong). We assume that the big set has higher precision since it is manually curated. And since the unique contribution of the small set is only 1.61%, we decided not to use it.

²⁵ <http://download.geonames.org/export/dump/alternateNames.zip>

²⁶ <http://mappings.dbpedia.org/server/ontology/classes/#Place>

²⁷ <https://www.wikidata.org/wiki/Property:P1566>

²⁸ [https://wdq.wmflabs.org/api?q=claim\[1566\]&noitems=1](https://wdq.wmflabs.org/api?q=claim[1566]&noitems=1)

²⁹ <https://github.com/dbpedia/extraction-framework/blob/dump/scripts/src/main/bash/process-geonames.txt>

³⁰ <https://github.com/dbpedia/dbpedia-links/tree/master/links/dbpedia.org/www.geonames.org>

We evaluated the injectivity and surjectivity (uniqueness/ambiguity) of the links in the big link set:

- 246 Wikipedia places (0.05%) map to several Geonames
- 12168 Geonames places (2.58%) map to several Wikipedia

We checked some of the Geonames→Wikipedia ambiguities, for example:

- http://dbpedia.org/resource/Aladağ,_Adana maps to both:
 - <http://sws.geonames.org/308998/>. Aladağ. Elevation ca. 896 m. population : 7613. coords: 37.5485, 35.39603. P PPLA2 seat of a second-order administrative division. Hierarchy: Turkey TR » Adana 81
 - <http://sws.geonames.org/8631906/>. Aladağ İlçesi. Elevation ca. 919 m. population : 1722. coords: 37.55854, 35.40196. A ADM2 second-order administrative division. Hierarchy: Turkey TR » Adana 81 » Aladağ İlçesi
Seem to be county and its capital. Close enough
- http://dbpedia.org/resource/Albania,_Santander maps to both
 - <http://sws.geonames.org/3690257/>: P PPL populated place. Population : 810. coords: 5.76139, -73.91778
 - <http://sws.geonames.org/3690263/>: A ADM2 second-order administrative division. population : 4473. coords: 5.77064, -73.86458. Seem to be county and its capital. Close enough

In the opposite direction, an example of Wikipedia→Geonames ambiguity is that Wikipedia has two pages for "Australia" (the country) and "Australia (continent)" whereas Geonames has a single resource for these two coextensive places.

3.12 sameAs Processing

We loaded GeoNames and the GeoNames-DBpedia owl:sameAs links to Ontotext GraphDB with the owl:sameAs optimization³¹.

- This means that the corresponding GeoNames-DBpedia names are merged into one, easing data access.
- The above ambiguities lead to a merged node with several coordinates, GeoNames ID and labels, but because the places are close and are represented in DBpedia as one resource, we decided that is ok.

³¹ <http://graphdb.ontotext.com/display/GraphDB65/GraphDB-SE+Reasoner#GraphDB-SEReasoner-sameAsOptimisation>

4 Semantic Enrichment

Semantic enrichment is the main purpose of the classification and the mainstay information processed by the semapp. We related CH objects to two semantic facets: the FD classification and Places.

Overall statistics as of 1 Sep 2015:

- Total objects: 7103 from 4 English collections, 9071 BG-ONTO
- Objects with at least one FD tag: 5664 in English collections (there are some Alinari objects with few if any FD-related words), 9071 in BG-ONTO (12 simple tags, see sec. 2.1.1)
- Objects with at least one Place tag: 6567 in English collections

4.1 FD Enrichment

The FD topical enrichment uses the FD tree.

- We enriched the Horniman collection through two approaches:
 - Semi-automatic: we mapped automatically the 700 FD object types in the museum's thesaurus, then verified the mapping of each one and made appropriate corrections and additions, then propagated the enrichments from object types to objects.
 - Automatic: based on automatic enrichment using ONTO's Concept Extraction Service (CES), then filtering concepts in the FD hierarchy only.
- The other 3 English collections were enriched only automatically.
- Each object in the BG-ONTO collection has enrichments, which are obtained through simple metadata conversion.

The first collection we selected for semantic enrichment was Horniman (see 2.1.3) because of the following factors:

- It's quite large and we got access to the complete collection
- It's in English, a language that we have most experience with
- Horniman uses a thesaurus, which makes the enrichment task a bit easier

We obtained the Object Types and Materials thesauri from Horniman as excel, and all 4350 FD objects as a JSON file.

- Out of 1400 Object Types relevant to FD (including Hunting and Fishing), 700 have corresponding objects, so we focused on them. It's easier to deal with these 700 concepts than the 4350 individual objects

We performed enrichment/alignment of the Horniman object thesaurus to DBpedia as follows:

- Adapted ONTO CES for working with FD articles & categories
- Since Horniman thesaurus terms lack any description, we formed "pseudo-documents" for the terms in order to provide some contextual information [Tagarev 2015 sec.6.1]
- We did a step of manual curation, because many Horniman terms are over-specified and need to be mapped to more general Wikipedia articles (e.g. "Tribulum" is a type of "Threshing board" that has stone chips.
- Adding some categories to the tree, e.g. Hunting and Spears.

Below are some statistics of FD tags appearing in Horniman objects:

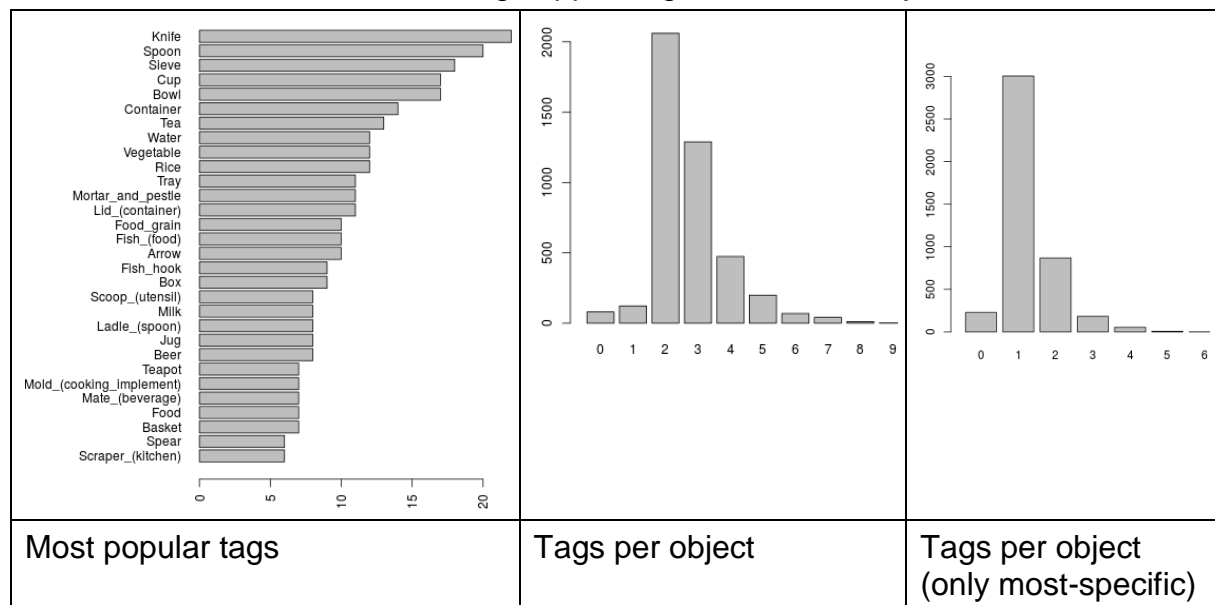


Figure 14 Statistics of FD Tags in Horniman Objects

4.2 Place Enrichment

We did Place enrichment using places from DBpedia and the standard ONTO CES. All 4 English collections were enriched automatically.

- Horniman metadata carries the complete place hierarchy for each place mentioned in an object (e.g. "Oceania › Melanesia › New Guinea › Papua New Guinea › Western Province", which allowed very precise enrichment. The latter is recognized as `Western_Province_(Papua_New_Guinea)`, although "Western_Province"³² is highly ambiguous: there are at least 10 such provinces.
- In many cases an object carries more than one place, e.g. place found vs current location. Also, there are often redundant places, i.e. the higher-up places in the place hierarchy.

1230 unique places and 22k place occurrences were recognized, which represents very high recall (see next section). Nevertheless, when we added the Geonames place hierarchy, we discovered the following:

- Out of 1230 DBpedia places:
- 815 places had a Geonames counterpart, thus are part of the hierarchy. Furthermore, there is full assurance that these are actual places, so we focused on the other 415 (which represent only 1% of the occurrences)
- 234 are legitimate places, but are either not in Geonames, or don't have a link. We added a few links at the Geonames site, and in all cases added a statement to complete the parent hierarchy. E.g. for the Roman Empire that straddled 3 continents.:

```
dbr:Roman_Empire gn:parentFeature dbr:Europe, dbr:Africa, dbr:Asia.
```

³² http://en.wikipedia.org/wiki/Western_Province

Table 4 Correct Place Enrichments not in GeoNames Hierarchy

Kind	Count	Percent	Examples
place	90	21.03%	White Nile, Wealden
historic place	52	12.15%	Roman Empire, British Empire, Soviet Union, Zaire
sub-city place	30	7.01%	Wapping Wall, Tornabuoni Chapel, Royal Arsenal
GLAM	26	6.07%	San Petronio Basilica, Belvedere (fort), Laurentian Library
non-admin place	17	3.97%	Balkans, Greater Antilles, Greater Upper Nile
parent	18	4.21%	Place needed to complete the parent hierarchy
total ok	233	54.44%	All are added to the place hierarchy

- 196 are not legitimate places. We are working with the ONTO CES team to correct these occurrences, which are due to heuristic types obtained by merging DBpedia and Wikidata type info

Table 5 Incorrect Place Enrichments

Kind	Count	Percent	Examples
disambiguation	120	28.04%	Bara, Barrington, Bartlett, Bath, Beloit, Berkeley
place type	33	7.71%	Bay, Bridge, Canton, Chapel, Chemical_plant, Coast, Countryside, Grave, Guildhall
concept	19	4.44%	Column, Flight_into_Egypt, Imperial, Industry, Lateran, Pilaster, Wall
person	6	1.40%	Arabs, Cuvier, Diomedea, Effie, Normans, St_Giles
redirect	11	2.57%	Austin, Bellevue, Louisville, Punjab, Raleigh, Rowland_Heights, Salamanca_(city)
broken	3	0.70%	Broken link due to renamed Wikipedia page
other	4	0.93%	Alder, Melanesian Spearhead Group, Poplar
total nok	196	45.79%	All are removed from enrichments

4.3 Enrichment Evaluation

An important question concerns the quality of enrichment. It is estimated by random sampling of objects and counting true positives (TP: correct matches), false positives (FP: incorrect matches) and false negatives (FN: failure to match). Then the following measures are calculated:

- Precision = $TP/(TP+FP)$
- Recall = $TP/(TP+FN)$
- F-measure = $2*P*R/(P+R)$, i.e. harmonic mean

Automated enrichment using ONTO Concept Extraction Service (CES) was used first, and provided a solid base for improvement via manual curation. ONTO CES achieves state-of-the-art performance of above 90% F-measure for publishing/news items.

When applied to a different domain (mixed metadata fields of cultural objects), the performance decreases. As reported in [Tagarev 2015 sec.6.1], the automated

enrichment achieved an estimated Precision of 0.91 and an estimated Recall of 0.7 on Horniman FD terms, which was raised to 100% by curation. The next figure shows which parts of the whole FD tree are activated by Horniman terms.

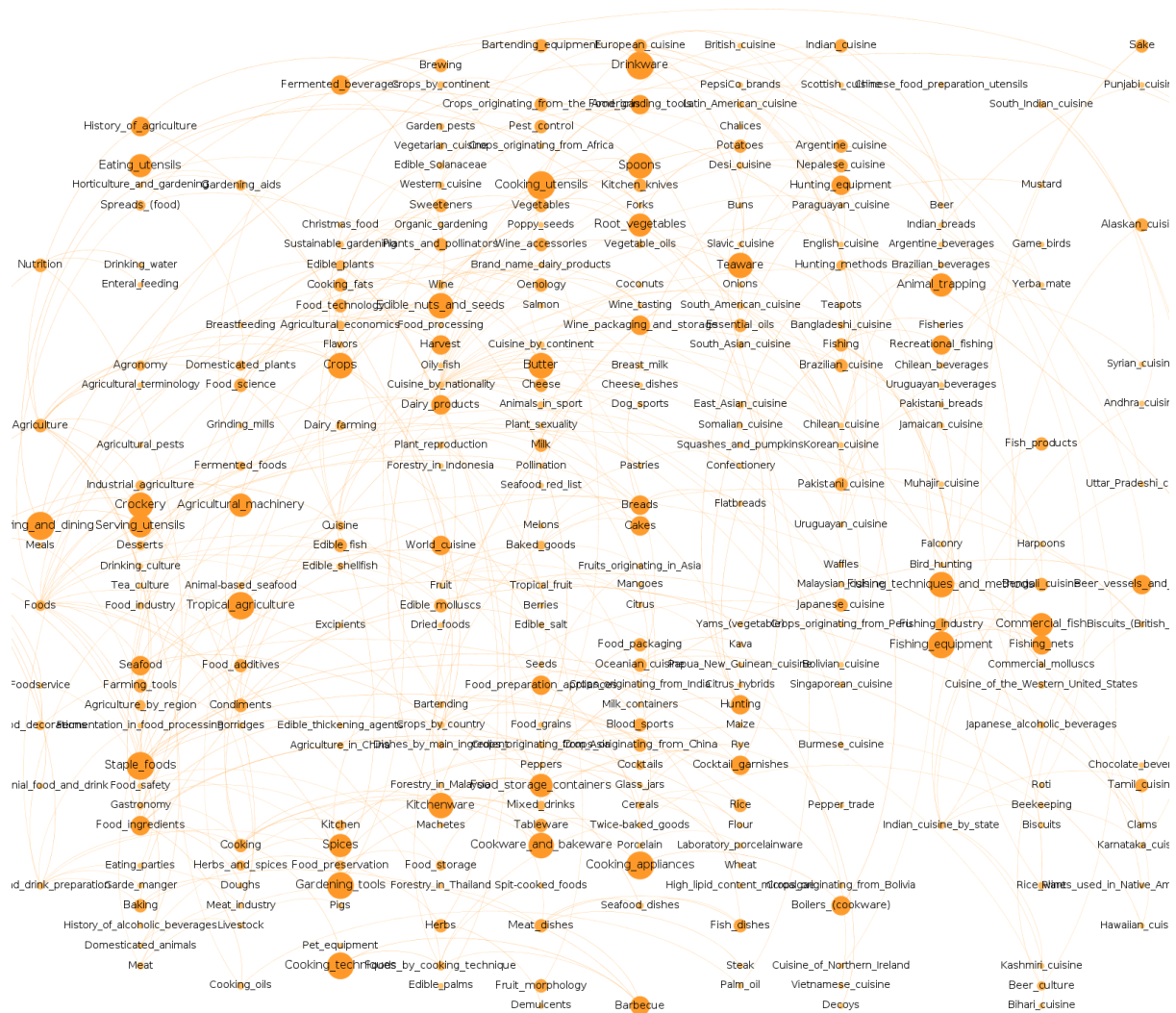


Figure 15 FD concepts activated by bottom-up propagation from Horniman terms
Automatic enrichment results across all English collections are as follows.

Table 6 Evaluation of Automatic Enrichments

Type	Evaluated	TP	FP	FN	Prec	Rec	F-Meas
FD	535	386	15	85	0.96	0.82	0.89
Places	104	306	17	20	0.95	0.94	0.94

FD enrichment:

- Excludes the keyword "Feasting" that appears in all Horniman objects (very unspecific) and is missed.
- The F-Measure of automatic enrichment is high.
- The F-Measure of Horniman objects is even higher since we complemented it with manual curation.

- We inspected CHOs without a single FD enrichment and in some cases added a tag. But there are indeed some Alinari CHOs that have very few or even no FD-related keywords.

Place enrichment:

- The F-Measure is very high.
- The factual recall is even higher because if a parent place is not recognized but its child place is recognized, the parent place will still be activated in the Places facet. E.g. in "Royal Library, Turin, Piedmont" we recognize `Royal_Library_of_Turin` and Turin but not Piedmont. Nevertheless, Piedmont will be activated because it's the parent place of Turin.
- One imprecision that our enrichment service exhibits is related to name inversion: e.g. Charlotte Warrington is written in the Horniman collection as "Warrington, Charlotte" and our pipeline takes that as two separate sub-sentences and mismatches it to "Charlotte, North Carolina"; but this is rare.
- We have removed manually 196 wrong place enrichments, which raised precision by about 1% (see previous section). In some cases we added a FD enrichment instead (e.g. "tray" was mis-recognized as the place "Trayes", which we corrected)

5 EFD Semapp

5.1 Semapp UI Design and Mockup

We developed a basic wireframe and mock-up for the semapp.³³ It is similar to Europeana (search, faceting, pagination, etc), but will provide additional semantic & hierarchical facets.

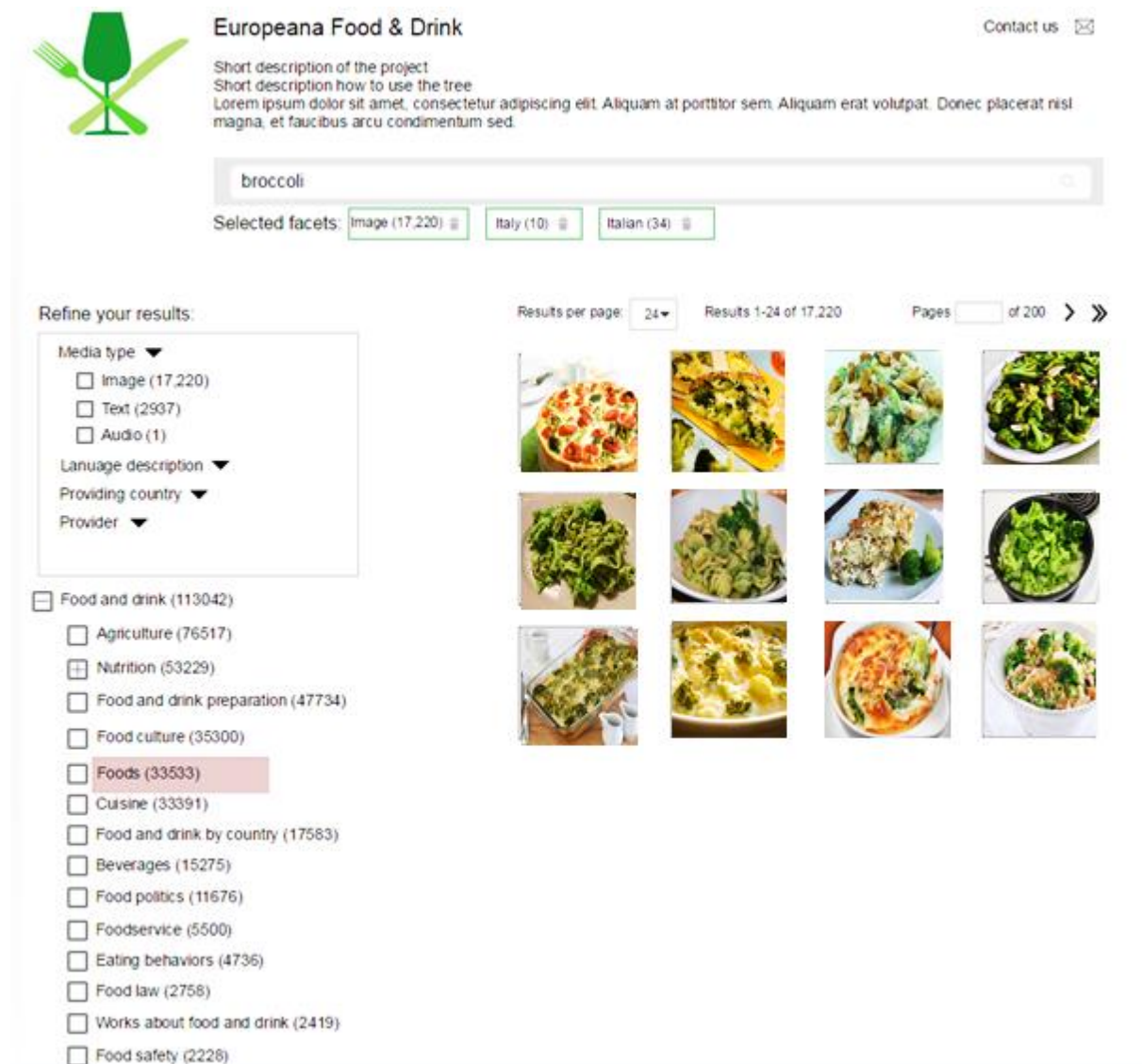



Figure 16 Semapp UI Mockup

5.2 Semapp Description and Screenshots

The semapp is a "mini-Europeana" that provides very similar searches and functionality to Europeana, but uses 2 semantic hierarchical facets (FD and Place). Some examples and actual screenshots.

³³ <https://live.uxpin.com/3adf4c6d0e75ed13bef0408a09adc837c228824b#/pages/25651569>



Europeana Food & Drink

The Semantic Demonstrator shows the use of semantic technologies for classification and discovery of Europeana objects related to Food and Drink. Detailed description, data, SPARQL endpoint.

Food and Drink

Places

Type (resource)

Language

bg 6419


Data provider

recepti.gotvach.bg 6419


Results per page: 24

Results 1 - 24 of 6419


Page 1 of 268




Лазаня на рулца
Корите за лазаня се нарязват на по-малки с ширина около 5 см и дължина колкото е парчето на бекона, отгоре се




Пролетни тиквени кюфтета
Тиквичките се настъргват на едро, посолват се и се изчакават да се отцедят. След това се смесват с




Руло с коприва
Яйцата се разбиват заедно с брашното и сместа се подправя на вкус с черен пипер. Готовото тесто се разпределя в




Сос с гъби за месо
Почистете и нарежете гъбите на парченца. Сложете ги да се сварят до омекване и ги отцедете. Загрейте маслото в




Шарена каша с



Чушки пане




Шарена бобена



Спагети със сос от

Figure 17 Bulgarian traditional recipes³⁴



Europeana Food & Drink

The Semantic Demonstrator shows the use of semantic technologies for classification and discovery of Europeana objects related to Food and Drink. Detailed description, data, SPARQL endpoint.

No active filters

Place: Oceania

Food and Drink

Places

Results per page: 24

Results 1 - 24 of 332

Page 1 of 14

Agriculture 234

Beverages 32

Cuisine 330

Eating behaviors 330

Food and drink by country 12

Food and drink preparation 331

Food and drink terminology 7

Food culture 330

Food industry 124

Food politics 29

Food safety 34

o 2

Africa 23

Americas 2

Asia 136

Europe 79


European Free Trade Association 1

Melanesia 214


North America 12

Oceania 332


Pacific Ocean 1




ladle
Large ladle with a bowl made of coconut, and a wooden handle. The handle has a triangular-shaped midsection, and is topped with an




tray (food service)
Round basket of stiff workmanship constructed from a coil of coconut leaf midribs spliced together which have been wrapped with coconut




fish hook
Two fish hooks with wooden shanks and bends and bone points bound to the hook with plant fibre. The larger of the two fish hooks has a line




arrow
For small game 'Hik'o'. Arrow (Bedamuni: taadi, nigi taadi). An arrow used mainly in war, but also for small game hunting. It is meant to break at




arrow (weapons)



arrow (hunting, fishing)

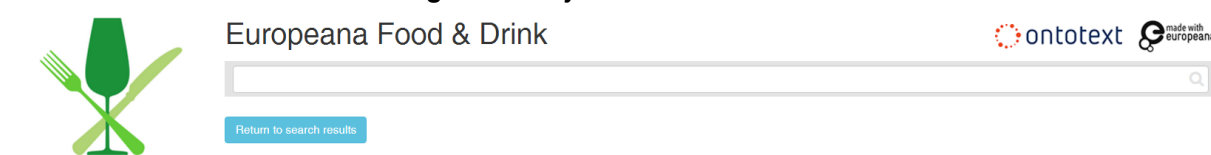


float (hunting, fishing &)



knife (food processing)

³⁴ <http://efd.ontotext.com/app/search?query=&limit=24&offset=0&dataProvider=recepti.gotvach.bg>

Figure 18 Objects From Oceania³⁵

<http://collections.horniman.ac.uk/objects/45472>

CHO: <http://collections.horniman.ac.uk/objects/45472> [EFD PDF](#)

Aggregation: <http://efd.ontotext.com/aggregation/http://collections.horniman.ac.uk/objects/45472> [EFD PDF](#)

float (hunting, fishing & trapping)

Float made of wood, with a human face at one end. The body of the float is roughly cylindrical with a rounded end.


Contributor	Dalton, J. H. (vendor, sale 1914)
Coverage (spatial)	http://dbpedia.org/resource/Papua_New_Guinea ; http://dbpedia.org/resource/Oceania ; Orokelo, Oceania (made or collected); Orokelo, Papua New Guinea (made or collected)
Data Provider	Horniman Museum and Gardens
Identifier	14.60; 19.a7; 45472; 5296; oo
Is Part Of	Anthropology (collection)
Is shown at	http://collections.horniman.ac.uk/objects/45472
Language	en
Medium (material)	wood (material)
Provider	Europeana Food and Drink
Rights	Horniman Museum and Gardens
Rights (URL)	http://creativecommons.org/licenses/by-nc-sa/3.0/
Subject	Food and Feasting , http://dbpedia.org/resource/Wood ; http://dbpedia.org/resource/Food ; http://dbpedia.org/resource/Fishing ; http://dbpedia.org/resource/Hunting ; http://dbpedia.org/resource/Fishing_float
Type (object)	floats (hunting, fishing & trapping)

Figure 19 Object Detailed View³⁶

The detailed view implements an image carousel (the above has 2 images) that is animated after a while or can be activated by the user. It also allows access to the object on the provider's site, and in the semantic repository (CHO and Aggregation data)

³⁵ <http://efd.ontotext.com/app/search?query=&limit=24&offset=0&place=Oceania>

³⁶ <http://efd.ontotext.com/app/resource/http%253A%252F%252Fefd.ontotext.com%252Faggregation%252Fhttp%253A%252F%252Fcollections.horniman.ac.uk%252Fobjects%252F45472?query=&limit=24&offset=0&place=Oceania>



Europeana Food & Drink

The Semantic Demonstrator shows the use of semantic technologies for classification and discovery of Europeana objects related to Food and Drink. Detailed description, data, SPARQL endpoint.

Selected filters: **FD: Fermented beverages** **Place: Asia**

Results per page: 24 - Results 1 - 24 of 114

Food and Drink

- Agriculture 96
- Beverages 114
- Cuisine 112
- Eating behaviors 113
- Food and drink by country 23
- Food and drink preparation 114
- Food culture 113
- Food decorations 2
- Food industry 65
- Food law 1
- Food politics 2

Places

- Africa 4
- Asia 114
- Earth 114
- Europe 27
- Indian Ocean 1
- Melanesia 2
- North America 2
- Oceania 2
- Southern Europe 3
- Western Europe 2

coffee pot; ewer
This object was probably made in the Punjab, perhaps at Silakot. The inscription looks like a verse, but not a religious one since the

water jug (jug (food service))
Sacred Lands (exhibition). Water vessel with a piriform body standing on a pedestal base. The body tapers to a

Ship laden with wine, relief preserved in Treviri. Work displayed at the Augustea Exhibit of the Roman World, held in Rome in 1937-

stopper (containers)
Cork stopper for a carved, roughly egg-shaped container made from a coconut.

food container; stopper

wine strainer (strainer)


churn; lid (food)

beer strainer (strainer)

Figure 20 Objects Related to Fermented Beverages and Asia³⁷

The third object is from the Roman Empire. It appears because we've added multiple parent statements (after all that empire did straddle 3 continents):

```
dbr:Roman_Empire gn:parentFeature dbr:Europe, dbr:Asia.
```



Europeana Food & Drink

The Semantic Demonstrator shows the use of semantic technologies for classification and discovery of Europeana objects related to Food and Drink. Detailed description, data, SPARQL endpoint.

Selected filters: **Category: Beverages** **Place: Roman Empire** **Alinari**

Results per page: 24 - Results 1 - 4 of 4

Food and Drink

- Agriculture 2
- Beverages 4
- Eating behaviors 1
- Food and drink preparation 2
- Food culture 1
- Food industry 1
- Foods 4
- Gustation 1
- Nutrition 2

Places

- Africa 4
- Asia 4
- Earth 4
- Europe 4
- Southern Europe 4
- Western Europe 1

Type (resource)

Language

Milk jugs and other containers for domestic use from Pompeii and Herculaneum, now kept at the National Archaeological Museum

The house of the Dolii, named after its thirty five large terracotta oil and wine jars stored in a pit. Ostia Antica
WARNING: Permission must

Shepherd milking a goat; relief preserved in the Roman National Museum, in Rome. The work was displayed at the Augustea Exhibit of

Ship laden with wine, relief preserved in Treviri. Work displayed at the Augustea Exhibit of the Roman World, held in Rome in 1937-

Figure 21 Objects from Alinari related to the Roman Empire and Beverages³⁸

This illustrates high-precision semantic search.

³⁷ http://efd.ontotext.com/app/search?query=&limit=24&offset=0&category=Fermented_beverages&place=Asia

³⁸ http://efd.ontotext.com/app/search?query=&limit=24&offset=0&dataProvider=Alinari&place=Roman_Empire&category=Beverages

5.3 Semapp Architecture

The conceptual architecture of the semapp is as follows.

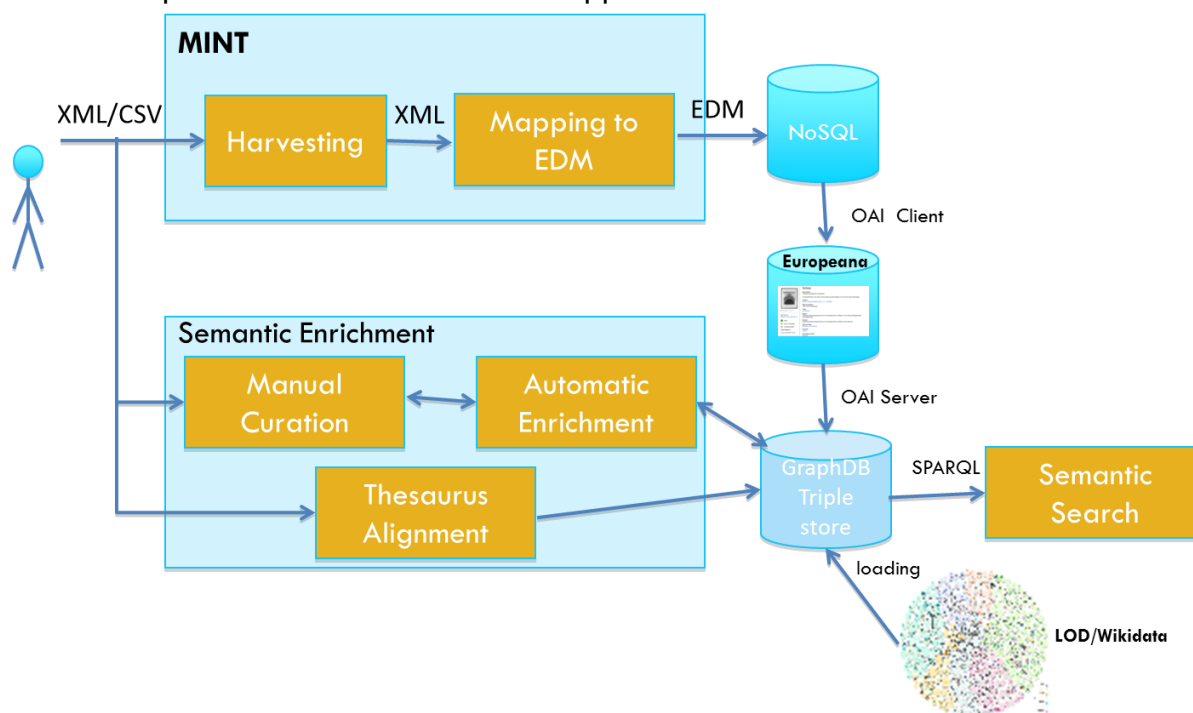


Figure 22 Semapp Conceptual Architecture

The software architecture of the semapp includes the following components and is based on the following principles:

- Ontotext GraphDB as the RDF semantic repository
- ElasticSearch for full-text querying, faceting and faceted querying (this is similar to Europeana's use of SOLR)
- GraphDB Enterprise Connector to ElasticSearch, which synchronizes data from RDF to ElasticSearch
- A Java backend that queries the database
- A JavaScript frontend implemented as a single-page application, and using modern JS technologies such as AngularJS, RequireJS, Bootstrap, UI Bootstrap, Lodash, font Awesome.
- Fully stateless REST communication between the backend and frontend. All information about the current search and facet state is in the URL and can be bookmarked, shared and communicated in email.

The semapp source will be released on GitHub in the near future, after some code reorganization and clean-up

5.4 Responsive UI

Due to the use of Bootstrap and corresponding JS and CSS coding, the application has a responsive design and can adapt to various device sizes. E.g. here is the same object as Figure 21 on a very narrow screen. The title was removed and the number of columns reduced to fit the screen size.



The [Semantic Demonstrator](#) shows the use of semantic technologies for classification and discovery of Europeana objects related to Food and Drink. [Detailed description](#), [data](#), [SPARQL endpoint](#).

Selected filters: [FD: Fermented beverages](#) [Place: Asia](#)

Food and Drink

+ Agriculture 96
+ Beverages 114
+ Cuisine 112
+ Eating behaviors 113
+ Food and drink by country 23
+ Food and drink preparation 114
+ Food culture 113
+ Food decorations 2
+ Food industry 85
+ Food law 1


Places

+ Africa 4
+ Asia 114

Results per page: 24


Results 1 - 24 of 114

Page 1 of 5



coffee pot; ewer

'This object was probably made in the Punjab, perhaps at Silakot. The inscription looks like a verse, but not a religious one since the



water jug (jug (food service))

Sacred Lands (exhibition), Water vessel with a piriform body standing on a pedestal base. The body tapers to a

Figure 23 Responsive UI on Narrow Screen

6 Dissemination

Due to the highly technical nature of the EFD semapp, we have attempted to disseminate semantic techniques mostly to technical audiences, not to the general public.

6.1 Semapp Website

Together with CT³⁹ we created a webpage for the semapp, and a detailed description⁴⁰.

6.2 Task Forces, Workshops

ONTO participated in the following task forces that are relevant to the semapp task:

- Evaluation and Enrichments⁴¹. Continuing the work of the Enrichment Strategy task force, this one will contribute specific recommendations for datasets, exchange formats, tools, and enrichment rules. As part of our participation, ONTO submitted trial enrichments of a selection of 13k objects by TEL. These enrichments were evaluated and compared against 5 other trial submissions, by projects such as LoCloud⁴². ONTO is very active in the task force.
- Europeana for Education. This task force will develop specific steps and recommendations towards implementing the Policy recommendations on using Europeana for Education⁴³ developed by ministries of education from 21 countries. ONTO was invited by Steven Stegers (EUROCLIO), our partner in Europeana Creative. We participated in the task force kick-off (21-22 June 2015 in Paris) second meeting (6-7 Oct 2015 in Warsaw), and ongoing work. The work on semantic enrichment was judged as very important for Education as well.

ONTO will also participate in:

- EDM Workshop (2 Nov 2015), where we will present our semantic work
- The Europeana AGM (3-4 Nov 2015). Vladimir Alexiev is running in the Europeana Council elections with the slogan "More quality and semantics in Europeana's operations".

In addition to Europeana for Education, we've had the following enquiries:

- Antoine Isaac (Europeana) enquired about applying our approach to build the Europeana Arts channel.
- Stefano Caneva (WeLand and Wikipedia, Italy/Belgium) enquired about semantic integration of Italian food resources for cultural path applications

6.3 Publications

We prepared and delivered 2 presentations and 1 paper (see References below):

[Alexiev2015e] A collaboration with Europeana, this presentation outlined the importance of Wikipedia/Wikidata for future Europeana enrichment. It provided examples of using Wikipedia for EFD classification, and

[Alexiev2015f] Prepared for EFD content partners, this presentation shows how easy it is to add labels and items to Wikidata, and somewhat harder to add categories and redirects (labels) to Wikipedia. It emphasizes the recommendations of the

³⁹ <http://foodanddrinkeurope.eu/professional-applications/semantic-demonstrator/>

⁴⁰ <http://foodanddrinkeurope.eu/wp-content/uploads/2015/09/EFD-Semantic-Demonstrator.pdf>

⁴¹ <http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/evaluation-and-enrichments>

⁴² <http://locloud.eu/Resources/LoCloud-enrichment-services>

⁴³ <http://pro.europeana.eu/publication/europeana-for-education-policy-recommendations>

Europeana and Wikimedia task force, and makes it clear that GLAM institutions can use Wikipedia and enrichment to make their collections searchable and discoverable in a multilingual context. Partner PS has taken this to heart and will work with the Cyprus Museum on enriching Greek Wikipedia with relevant FD articles and terms.

[Alexiev2015f] Presented the work that ONTO completed as part of Europeana Creative to establish 2 new access channels for Europeana Labs: OAI PMH server for bulk object download, and EDM SPARQL for semantic querying. We also talked about our experience and tasks in EFD.

[Tagarev 2015] This paper describes our approach to building a domain-specific gazetteer for EFD and includes more scientific details. The paper was accepted and delivered to the International Keystone Conference on semantic keyword search in Portugal in Sep 2015. Furthermore, we were asked to submit an extended version for a journal special issue.

7 Work in Progress and Future Work

In this section we describe various pieces of work that we started but were not able to complete due to limited effort; or ideas for related work. These could be tackled in an extension of the semapp after Oct 2015.

7.1 Europeana CHO Discovery

An important benefit of the FD semantic classification is that we can discover already existing objects in Europeana on the topic of FD. Some approaches are described in [Alexiev 2015b sec.2.12]. Focusing on the technical side, this presents significant challenges:

- We identified 200k articles in 14k categories relevant to FD. Each article has many titles: labels and redirects. There are 3.02 labels per article on average (we have seen items/articles with as many as 40 labels).
- So this makes 456k labels that need to be queried against Europeana. It makes sense to make a query per article, each being a disjunction (OR) of all labels of that article.

Here is some data about articles and labels. It's from a bit older version of the FD tree that includes fewer objects than described above.

Table 7 FD Articles and Labels for Discovery

Category	level	categories	articles	redirects	total labels
Food_and_drink	0	9870	113022	228176	341198
Beverages	1	1487	15262	38417	53679
Caffeinated_beverages	2	103	872	3217	4089
Tea	3	58	617	1562	2179

7.1.1 Tea-Related Objects

We started Europeana discovery for Tea-related objects, since many Horniman objects are related to tea, and one of the EFD products (Tea Trails) is directly related to tea.

We got 658 tea-related articles with 2324 labels. The start and end of this list is:

- 24 flavors; 24 tastes; 24 mei
- ABC tea shops
- A Nice Cup of Tea
- Ahmad Tea
- Akumaki
- Alghazaleen Tea
- Yōkan; 栗子羹; Youkan; 栗羊羹; Lizigeng; Goat liver bar; Yokan; Liyanggeng; Yanggeng; 栗子羊羹; Shioyoukan; Yohkan; Yookan; 羊羹
- Zealong; ZEALONG
- Zenga

We wrote a Perl script that queries the Europeana API:

- Each query is an OR of all labels for one article.
- We drop parenthesized qualifiers (e.g. for "Arare (food)" we query "Arare")
- We use profile=minimal and rows=100 to decrease the load on the server. Nevertheless, we got a number of server errors, e.g. query "Benoist" at start=3401 obtained "500 Internal Server Error"

We discovered several ambiguous words that match many irrelevant objects, so we black-listed them in the script. (We don't filter by language because Europeana language tags are not consistent or exhaustive.) For example:

Blacklist	Comments
(clipper), Ariel, Eleanor, Dartmouth	Clippers that participated in the Boston tea party. The names are generic and fetch many objects
24 mei	"24 May" in Dutch: fetches thousands of newspaper issues
Jamaica	Another name for "Hibiscus tea" or "Karkadé"
Kanten	"lace" in Dutch or "edge" in Nynorsk

We also blacklisted a whole collection: **askaboutireland.ie**. They have scanned tons of Yellow Pages from "Thom's Commercial Directory" from 1975 and submitted every page as a separate CHO. The pages are meticulously OCR'd (the text is perfect), so this collection is a match for pretty much any name you query for (e.g. "Brooke Bond", which is a brand of tea).

In our opinion, this collection should be expunged from Europeana (together with scientific articles submitted by TEL, hand-written census pages, etc). Ironically, many precious texts are not OCR'd at all or not well recognized.

We've only completed the download of 43 queries (out of 658) but already got about 3.5k objects. Some interesting hits:

Hits	Labels
25	"Amacha" OR "Ama-cha" OR "甘茶" OR "あまちゃ"
3259	"Anthemis"
225	"Arare" OR "Kaki mochi" OR "Kakemochi" OR "Mochi crunch" OR "Kakimochi" OR "Norimaki arare" OR "Hurricane popcorn"
48	"Assam tea" OR "Camellia sinensis assamica" OR "Assam Tea"

The main label "Tea" alone matches 9.9k objects. But we are doubtful we'll be able to obtain them from the Europeana API (see error 500 above). So it may be better to use the ONTO Europeana SPARQL endpoint, which also provides keyword search (FTS).

We made some surprising discoveries, e.g. a WW1 "Wounded" letter⁴⁴ that is related to Tea since it mentions "Brooke Bond".

7.1.2 Restaurants

On 25 Aug 2015 we had a call with Shift on the topic of WP5 Engagement. We emphasized that it would be nice for product partners to use some of the semantically enriched or discovered objects in their products.

Shift suggested that instead of Tea objects, we should discover restaurants and similar establishments, because it will be easier to geo-locate them. Then the enriched objects can be placed on HistoryPin as an interesting collection.

We started evaluating queries with "restaurants" but the work is incomplete. We will continue work on Discovery as part of the extended semapp scope.

⁴⁴ http://www.europeana.eu/portal/record/2020601/attachments_52959_4640_52959_original_52959_jpg.html

7.1.3 FD Classifier

We used some machine learning techniques to create a FD Classifier. This module can predict whether an object is FD-related or not by looking at the metadata text of the object. The prediction is based on the Wikipedia text of FD-related articles. The current implementation and possible improvements are described below.

- The available labelled data consists of 4330 **positive** examples (articles used to tag Horniman objects), 106k **maybes** (all other articles in the FD hierarchy) and 3.6M **negative** (articles outside the FD hierarchy). The model was trained using all positive examples and a random sub-sample of size 5000 from the negatives. We should include more articles as positive examples, e.g. from leveraging other LOD datasets that evidence FD relevance (see sec 2.4).
- The most informative features (word **stems** with largest weights in the model) are as follows:
food, fish, cook, cake, agricultur, tree, bread, sweet, type, milk, plant, tradit, dish, common, sugar, shape, cuisin, drink, rice, edibl, coffe, water, fruit, perenni, nativ, popular, tea, hunt, dessert
- We use the article abstracts (i.e. first paragraphs before the Table of Contents of each article).
- We use a simple "bag of words" approach. Performance may be improved by giving special prominence to linked words or key phrases in the articles.
- The classifier should be retrained after updates to DBpedia, the FD classification tree, or amended evidence.

Technical notes:

- A regular maxent model was trained on 80% of the samples.
- Results:
- **Training set:**
pos F1:0.99 Prec:0.99 Rec: 0.99
neg F1:0.99 Prec:0.99 Rec: 0.99
Golden set pos: 3354 samples; neg: 3846 samples;
Macro-F1: 0.99, Micro-F1: 0.99
- **Test set:**
pos F1:0.95 Prec:0.98 Rec: 0.93
neg F1:0.95 Prec:0.94 Rec: 0.98
Golden set pos: 902 samples; neg: 899 samples;
Macro-F1: 0.9572269325637222
Micro-F1: 0.9572459744586341

Other notes:

- The model is biased towards recognizing documents similar to Horniman CHOs, because for the moment the evidence is mostly from the Horniman thesaurus.
- The model could be biased towards popular topics in Wikipedia. There are numerous pages about people in Wikipedia. Then, the negative set, being randomly sampled, may be biased towards biographies of peoples, which makes it easy to separate positives and negatives (food vs. people). So, the accuracy could be too optimistic. A more realistic negative set would lead to a more general model, applicable to any domain.

Once fine-tuned, this classifier can be a very promising module for Europeana Discovery.

- Rather than making queries using specific keywords, we can run it through all Europeana CHOs, predicting **which** are FD-relevant.

- Because there are 43M CHOs, speed is a concern. But extracting features from CHOs is fast because they are small; and prediction for a new case is a fraction of a second
- Then we will run semantic enrichment over the positively predicted objects to find out **why** are they relevant.

7.1.4 Europeana Problems

In experimenting with Europeana Discovery, we found some problems with the data.

Improper Enrichment with Narrower Terms

For example this cylinder jar⁴⁵ (also see provider site⁴⁶) has provider terms "Zylinderhalsgefäß"@de = "cylinder jar"@en, "Gefäß"@de="vessel"@en; "Angewandte Kunst"@de = "applied art"@en. It is correctly enriched with concept "vessels (containers)"@en = "Gefäß (Behälter)"@de from the Partage vocabulary⁴⁷.

- However, it is incorrectly enriched with 26 AAT **narrower** terms of "vessels": esker, bokser, ... samovars.
- Also, it is irrelevantly enriched with 2 GEMET **broadener** terms of "container": miscellaneous product, product.

Because of the first problem, many vessels that are decidedly not samovars, are marked as "samovar" on Europeana. In fact most of the 917 objects found by querying for prefLabel "samovar"⁴⁸ are not samovars. We raised appropriate issues to Europeana.⁴⁹

In contrast, the semantic approach provides multiple attested labels for the concept: Samovar; Electric samovar; Semaver; Samowar; Zavarka. We found 960 objects with "Samowar"⁵⁰. Because this spelling doesn't appear in thesauri (it is used less often), it's free of the "narrower" concepts defect and all hits are relevant.

Multilingual Ambiguity

This problem has been reported widely, but we want to emphasize it. A seemingly unambiguous term like "Beer" is in fact ambiguous when used in different languages. It can refer to "de Beer" (a very common Dutch name) or "Bears. When searching for "beer"⁵¹ you may find that only 1/20 of the objects are relevant.

Improper Person Name Representation

Searching for "Kettle" returns a medal by "Artist: Kettle, Henry, die-engraver". Can enrichment discover that this is not a relevant match? Unfortunately the object metadata has this unreasonable Subject: "Henry; medals; Kettle; medal". Rather than in dc:creator, the name is put in dc:subject, and is split up beyond recognition in two separate dc:subject fields. So there is no easy way to recognize Kettle as a person name in structured fields.

⁴⁵ http://www.europeana.eu/portal/record/08501/Athena_Update_ProvidedCHO_Bildarchiv_Foto_Marburg_obj_20727191_410_848.html

⁴⁶ <http://www.bildindex.de/dokumente/html/obj20727191#|home>

⁴⁷ <http://partage.vocnet.org/html/part00083>

⁴⁸ http://www.europeana.eu/portal/search.html?query=cc_skos_prefLabel:samovar

⁴⁹ <http://www.assembla.com/spaces/europeana/tickets/2044-enrichment-shouldn--39-t-add-narrower-broadener-concepts>,

<http://www.assembla.com/spaces/europeana/tickets/2045-concept-labels-are-mangled>,

<http://www.assembla.com/spaces/europeana/tickets/2046-enrichment-concepts-are-not-connected-to-cho>

⁵⁰ <http://www.europeana.eu/portal/search.html?query=samowar>

⁵¹ <http://www.europeana.eu/portal/search.html?query=beer>

The only way to recognize it is from the free-text field: "Description: Artist: Kettle, Henry, die-engraver". This involves name inversion ("Last, First") that is very common in the library domain, but our enrichment pipeline does not yet handle. But even if the artist name is recognized in Description, that does not provide sufficient warrant to discard object type "Kettle" from the Subject field.

7.2 Enrichment Web Service

In late Oct there was a discussion with Shift and EEA about the creation of a Crowdsourcing Enrichment application, to be developed by D3.5 Technical Demonstrator and T5.2 Community/ crowdsourcing platform. ONTO would establish a web service to:

- Perform semantic enrichment of FD topics and Places to suggest automatic enrichments to curators (people from the content partners)
- Provide interactive search with auto-completion for the same categories of data, to enable curators to select tags themselves.

Important issues to be tackled for this service are:

- Availability, performance, monitoring
- Fine tuning of the enrichment process, leveraging all manual work done to date

The same service could be used to provide semantically enriched content to other application creators as well.

7.3 Culture, Ethnicity, Period, Style, Movement

Culture, Ethnicity, Period, Style, Movement are important aspects of a CHO. Since the boundaries between these categories are not always clear-cut, it makes some sense to treat them uniformly (as Getty AAT does).

We have started a significant effort to compile a master list from the following sources:

- Getty AAT's facet Periods/Styles has 5.5k entries, of which 2.2k are nationalities.
- The British Museum Ethnic Group thesaurus has about 2.5k ethnicities.
- Wikipedia/DBpedia has over 10-15k such articles. We discover them using several approaches:
 - Class `dbo:EthnicGroup`
 - Property `dbp:ethnicGroups` on Region or Place
 - Property `dbp:ethnicity` on Language or Person
 - Property `dbo:movement` on `dbo:Artist`
 - Article titles ending in "people", "tribe", "culture" or their plural variants.
- (We have also evaluated the AFSET Ethnographic Thesaurus published as part of LoC Subjects⁵² but it doesn't have such categories).

They are relevant to the EFD semapp because Horniman has a term Ethnic group (e.g. Ainu) and Wolverhampton has periods (e.g. Victorian). This would make a nice extra hierarchical semantic facet.

Significant cleaning is required to make this data usable. E.g. for articles ending in "culture" we need to remove "Bicycle culture" and "LGBT culture"; for `dbo:movement` we need to remove revolutionary movements, etc.

Our ambition is also to create a merged hierarchy, using the respective AAT and BM hierarchies. DBpedia doesn't have a useful hierarchy for this type of data.

⁵² <http://id.loc.gov/vocabulary/ethnographicTerms/>

Such data will be added as a new semantic facet. It would be useful for EFD, e.g. Horniman uses a lot of Ethnic groups and Wolverhampton uses various English periods/styles.

More importantly, it will be an important contribution to the CH LOD cloud, since no such master list exists today.

7.4 New Language for Enrichment

It would be very useful to extend semantic enrichment to another language (in addition to English), in order to increase the number of collections than it can handle. The considerations for selecting a new language to handle are:

- Size and importance of collection data from EFD content partners
- The richness of the respective national-language Wikipedia
- The availability of language-specific NLP resources
- ONTO's experience with the language

Our current candidates for extension include:

- **Bulgarian:** we have the full BG-ONTO collection (converted to EDM) and some language resources (from collaborations with the Bulgarian Academy of Sciences). However, preliminary assessment of the FD coverage of BG Wikipedia is not very promising
- **Dutch:** we have commercial NLP experience with Dutch (for the Dutch Press Association NDP) and sufficient language resources. We have not assessed NL Wikipedia. The counts in [Alexiev 2015a sec.3.7] are promising (lots of articles), but the counts in [Alexiev 2015a sec 3.8.1] are not promising (poor category structure).
- **German:** we have recent NLP experience in German. the German Wikipedia is very well developed. AT-ONB provides a collection in German
- **Greek:** the Cyprus Museum provides an important collection in Greek, since it's dedicated to the topic of FD. But we don't have NLP experience, and the Greek Wikipedia is very poor.
- **Italian:** we have preliminary NLP experience. Wikipedia is of medium size, and there are 2 EFD collections.

An important sub-task will be to leverage inter-language links, i.e. links across articles and category networks in different Wikipedias (see [Alexiev 2015a sec.2.3.2] for considerations).

7.5 Handling Lists, Cuisines

Wikipedia Lists are a useful classification mechanism in addition to Categories. [Alexiev 2015b sec.2.4] describes the tasks required to use these lists in a way similar to Categories.

- We would need to implement a custom extractor, which we can do either stand-alone (e.g. using the Wikipedia Miner framework) or as part of the DBpedia Extraction Framework (which is written in Scala).
- We have aliased with 4 students from the University of Potsdam who announced they would be working on list extraction at the DBpedia mailing list (July 2015), enquiring about possible collaboration

Other tasks related to enlarging the scope of FD enrichment are:

- Continue elaborating the FD Classification through bottom-up augmentation
- Correlation from X Cuisine to place/culture X, e.g. Bulgarian cuisine→Bulgaria or Cajun cuisine→Acadia, Louisiana

7.6 Geographical Mapping

Now that we have Place enrichment in the semapp, relatively little work is need to implement a Geographical Map tab, in addition to the existing lightbox (thumbnail grid). It will involve the following tasks:

- Eliminate superfluous ancestor places. E.g. if a CHO is tagged with Rome and Italy, we want to remove the parent place, else the same CHO will appear with two different markers on the map
- Complement with ancestors with coordinates: If a CHO is marked with "Fleet Street" and neither GeoNames nor DBpedia have coordinates about it, we need to add its lowest ancestor with coordinates (in this case, "City of London" and not "London" which is a greater area)
- Average coordinate values. Experience shows that many places have several coordinate pairs that differ by little. We need to consider that in the query, and average the coordinates of the same place, to ensure one marker per place.
- Add coordinates to the semapp backend (API). An important consideration is the total response size, and whether we need to limit it somehow.
- Implement geo map display in the semapp frontend. We'd like to use the "marker clusterer" library, which can display many thousands of places by using spots with the number of markers, which upon zoom are split into more fine-granularity spots.



Figure 24 Marker Clusterer Geographic Map

7.7 AAT-Wikidata Coreferencing

We want to use Getty AAT for its strong Styles and Periods facet⁵³, which also includes numerous Cultures and Tribes. It has 5487 terms, which can be checked with this query⁵⁴:

```
select * {?x gvp:broaderExtended aat:300015646}
```

Comparing to Wikipedia's https://en.wikipedia.org/wiki/Category:Ethnic_groups category, this is a more compact and well-defined list. However, we still need to compare to Wikipedia ethnic groups, because we don't know how the coverage of the two datasets compare.

⁵³ <http://www.getty.edu/vow/AATHierarchy?find=periods&logic=AND¬e=&subjectid=300015646>

⁵⁴ At SPARQL endpoint <http://vocab.getty.edu/sparql>, maintained by ONTO

In order to use AAT and Wikipedia in concert, we need to coreference AAT to Wikipedia. One of the best ways to do this is by using the Wikipedia Mix-n-Match tool⁵⁵, as mentioned in [Alexiev 2015b sec.2.5.2]. We started a task for this at the Wikidata project Authority Control⁵⁶ (also initiated by us). AAT is already loaded to Mix-n-Match, however the precision of automatically matched concepts is only 50%⁵⁷. So we devised an approach to salvage old AAT-WordNet 2.0 mappings through WordNet 3.0 and BabelNet:

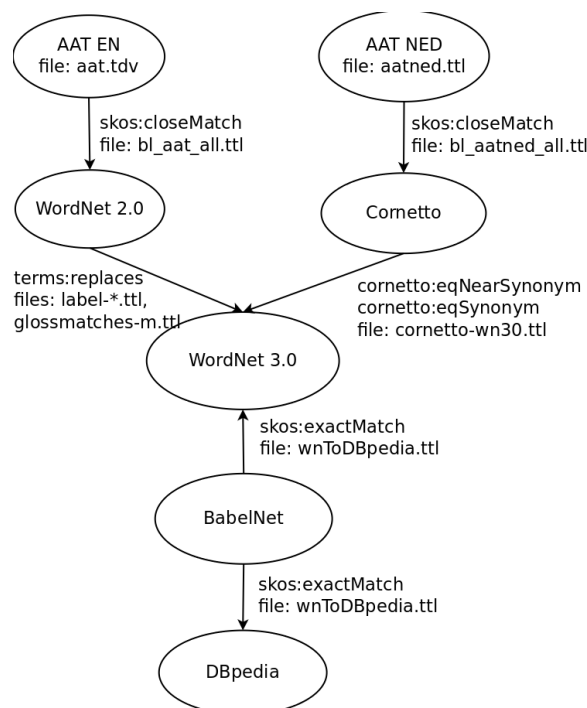


Figure 25 Data flow for matching AAT to DBpedia through WordNet and BabelNet

This automatic matching is completed:

Table 8 AAT to DBpedia matches: EN (AAT) and NL (AATned). * is estimated

	EN	NL	Union
Matches	3841	5949	7570
Matched AAT concepts	3288	5092*	6481*

We matched 6481 AAT concepts to DBpedia. On average, there are 1.168 matches per concept. E.g. AAT "ballistae" and "trebuchets" are matched to Wikipedia "Arbalest", "Ballista", "Catapult", "Mangonel", "Trebuchet": all are closely matching concepts. The matching has high precision (over 95%), but the precision of Cultures/Styles is much lower (about 65%). Some example mismatches:

- Fremont (Pre-Columbian South-western North American style) is mismatched to John C. Frémont (an American military officer)
- "Touchstone" (velvety-black variety of cryptocrystalline quartz) is mismatched to "Technical standard"

⁵⁵ <https://tools.wmflabs.org/mix-n-match/>

⁵⁶ https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control#Coreference_AAT

⁵⁷ https://meta.wikimedia.org/wiki/Talk:Mix%27n%27match#Coreference_AAT

7.8 Propagating UMBEL

UMBEL is a subset of OpenCyc, including a mapping to DBpedia. We have extracted from UMBEL 44237 items that are FD-relevant with a high degree of probability, using the UMBEL FD Super-type (see [Alexiev2015a sec 3.19]). Although the UMBEL→DBpedia mapping is 4 years old, it will be worth to propagate these items against the EFD tree, which will be useful both for enlarging the tree, and précising the FD classifier (see sec 7.1.3)

7.9 Propagating Dbtax

DBtax is a heuristic addition of types to DBpedia performed by the Italian DBpedia chapter. The types themselves are not always meaningful, e.g. **Zutho** (brand name of a soft drink) is classified as dbtax:Beverage but also dbtax:Article, dbtax:Type. But they are a good predictor of article clustering.

We selected all articles relevant to FD using an iterative process: we started from dbtax:Food and dbtax:Beverage and added appropriate co-occurring types.

As a result we fetched 20k articles in two categories: Relevant and Maybe. We still need to evaluate the relevance of the latter category, and to propagate the evidence

```
### RELEVANT
 141 Appetizer
3246 Beverage
 122 Brandy
 307 Breakfast
 184 Chocolatier
 182 Cookbook
2002 Dish
 959 Drink
1665 Farm
 91 Fireplace
 461 Fishery
5744 Food
 83 Gin
 73 Grain
1112 Ingredient
 258 Liqueur
 31 Melon
 153 Nutritionist
 139 Pizzeria
3965 Restaurant
 146 Sausage
 212 Sweetener
 194 Utensil
 218 Vodka
 101 Whisky
 810 Winery
```

```
### MAYBE
 283 Additive
 36 Alcohol
 30 Appliance
1651 Brand
 11 Breed
5152 Company
 5 Diabete
 59 Dietetic
 1 Diuretic
 14 Famine
 80 Fertilizer
 2 Insecticide
 5 Market
 27 Nutrient
 6 Pesticide
 2 Seaweed
 52 Shop
1334 Variety
 4 Venture
```

7.10 Mobile Application

ONTO has not yet done any work on a mobile semantic application due to the large delays in the project related to content delivery. For example, the provider of one of the key collections used in the semapp (Horniman Museum) still has not converted their data to EDM. Since it is impossible to develop a semantic application with a few selected objects, ONTO had to spend a large amount of effort speaking to different content providers, obtaining their data in whichever way it was available, and converting it to EDM. This left less time and effort than desired for the development of the semapp proper. This has been communicated in periodic progress reports D3.20a and D3.20b.

We have not yet subcontracted the development of a mobile application, because we had neither the data, nor the enrichments to enable its development. Furthermore, the consortium feels that the rest of the semapp budget will be best spent on extending the enrichment scope, opening up APIs for other partners to use semantic data, and extending the functionalities of the semapp. This has also been communicated in D3.20a sec 3.4 Extended Scope at the end of June, and in a formal "Request for Budget Reallocation and Time Extension" in October.

Nevertheless, if the EC does not approve this change, ONTO will proceed according to the original DOW and engage a subcontractor to develop a mobile application. It will be similar in scope and functionalities to the current EFD semapp. There is more than enough time to develop such an application before the end of the project.

8 References

- [Alexiev 2015a] Vladimir Alexiev. Europeana Food and Drink Classification Scheme. Deliverable D2.2, Europeana Food and Drink project, February 2015. [http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Classification-Scheme-\(D2.2\).pdf](http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Classification-Scheme-(D2.2).pdf)
- [Alexiev 2015b] Vladimir Alexiev. Europeana Food and Drink Semantic Demonstrator Specification. Deliverable D3.19, Europeana Food and Drink project, March 2015. [http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-Specification-\(D3.19\).pdf](http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-Specification-(D3.19).pdf)
- [Alexiev 2015c] Vladimir Alexiev. Europeana Food and Drink Semantic Demonstrator M18 Progress Report. Progress Report D3.20a, Europeana Food and Drink project, June 2015. [http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-M18-Report-\(D3.20a\).pdf](http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-M18-Report-(D3.20a).pdf)
- [Alexiev 2015d] Vladimir Alexiev and Laura Tolosi. Europeana Food and Drink Semantic Demonstrator M21 Progress Report. Progress Report D3.20b, Europeana Food and Drink project, Sep 2015. [http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-M21-Report-\(D3.20b\).pdf](http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-M21-Report-(D3.20b).pdf)
- [Alexiev 2015e] Vladimir Alexiev, Valentine Charles, and Hugo Manguinhas. Wikidata, a Target for Europeana's Semantic Strategy. In *Glam-Wiki 2015*, The Hague, April 2015. <http://www.slideshare.net/valexiev1/wikidata-a-target-for-european-s-semantic-strategy-glamwiki-2015>
- [Alexiev 2015f] Vladimir Alexiev. GLAMs Working with Wikidata. In *Europeana Food and Drink content provider workshop*, Athens, Greece, May 2015. <http://www.slideshare.net/valexiev1/glams-working-with-wikidata>
- [Alexiev 2015g] Vladimir Alexiev and Dilyana Angelova. O is for Open: OAI and SPARQL interfaces for Europeana. In *Europeana Creative Culture Jam*, Vienna, Austria, July 2015. [http://vladimiralexiev.github.io/pubs/O_is_for_Open_\(CultJam_201507\)_poster.pdf](http://vladimiralexiev.github.io/pubs/O_is_for_Open_(CultJam_201507)_poster.pdf)
- [Tagarev 2015] Andrey Tagarev, Laura Tolosi, Vladimir Alexiev. Domain-specific modelling: Towards a Food and Drink Gazetteer. First International Keystone Conference, Coimbra, Portugal, Sep 2015. <http://vladimiralexiev.github.io/pubs/Tagarev2015-DomainSpecificGazetteer.pdf>