Grant Agreement 621023

Europeana Food and Drink

Semantic Demonstrator M21 Progress Report

Deliverable number	*D3.20b*
Dissemination level	*CO*
Delivery date	*9 Oct 2015*
Status	*Final*
Author(s)	*Vladimir Alexiev, Laura Tolosi (ONTO)*

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 progress on developing the EFD Semantic Demonstrator for the 3 months from 1 Jul 2015 to 1 Oct 2015. We describe all work performed, the achieved results and project management considerations.

This is not a formal deliverable but a periodic progress report. It should be read in conjunction with D3.20a, which describes the work performed between 1 April 2015 and 30 June 2015.

Revision History

Rev	Date	Author	Org	Description
v0.1	6/10/2015	Vladimir Alexiev	ONTO	Initial version
v0.2	7/10/2015	Laura Tolosi	ONTO	Additions
v0.3	8/10/2015	Susie Slattery	CT	Review
v0.4	9/10/2015	Vladimir Alexiev	ONTO	Revision

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

This document describes the development progress on the EFD Semantic Demonstrator (semantic application or "semapp") for the second 2.5 months of development.

Work on the semapp started in April 2015 and the previous development period is described in D3.20a [Alexiev 2015e]. This report was submitted to EC on 10 Jul 2015, was reviewed by the project reviewers, and received positive feedback on 15 Jul 2015.

We describe the work performed between 1 Jul 2015 and 15 Sep 2015, the achieved results, and project management considerations. This is not a formal deliverable but a periodic progress report.

1.1 Structure of the Document

This document is structured in the following sections:

Work done: describes work done to date:

- Collection metadata conversion to EDM
- BG metadata conversion and submission
- Elaborating the FD Classification tree
- Manual curation: internal and in Wikipedia
- Semapp design and UI mock-up
- Bottom-up relevance propagation
- Semantic enrichment of collections
- Evaluation of enrichments
- Discovery of Europeana objects
- Leveraging LOD (Getty AAT, DBtax)
- Participation in task forces
- Dissemination and Publications

Project management: describes next steps and resource considerations

- Immediate next steps
- Scope for Oct 2015
- Extended scope until EFD project finish
- Potential future applications

1.2 Abbreviations

Abbrev	Description
AAT	Getty Art and Architecture Thesaurus
API	Application Programming Interface
BG	Bulgaria or Bulgarian language
CH	Cultural Heritage
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
JSON	JavaScript Object Notation
KB	Knowledge Base
LIDO	Lightweight Information Describing Objects, a museum object XML schema
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
TSV	Tab Separated Values
UI	User Interface
URL	Uniform Resource Locator
UTF-8	The most commonly used Unicode Transformation Format

2 Work Done

2.1 Metadata conversion to EDM

EFD experienced significant delays with collecting and converting metadata to EDM, which was the main concern of the project mid-term review. In order not to block development of the semapp, ONTO spent a lot of effort to collect metadata samples from most of the content providers.

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).

The next table shows the number of English-language objects to be used by the first iteration of the semapp (Oct 2015).

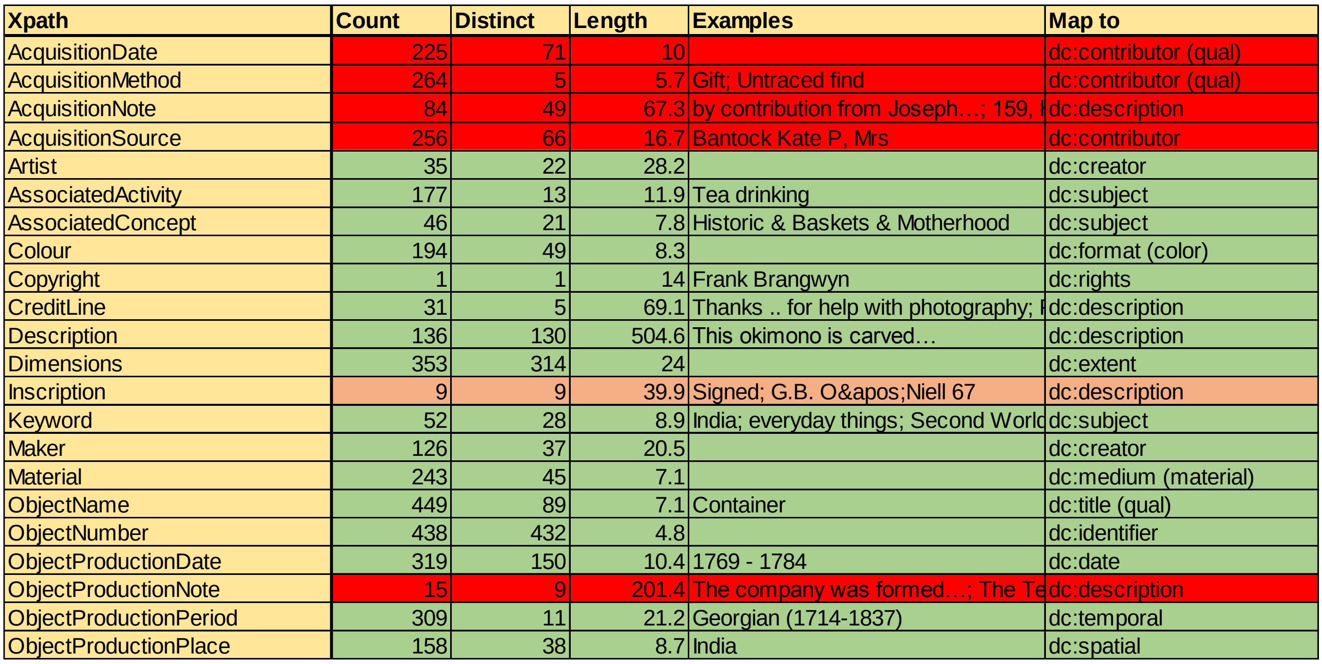
Table 1 English Collections as of 15 Sep 2015

Collection	Obj	Notes
IT-Alinari	498	All have images[1], 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[2], 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[3]. Most are English/Victorian objects.
UK-TopFoto	1814	All have images[4], many are monochrome. A lot of keywords, but a moderate number are about FD. This is a preliminary release, TopFoto has submitted 6119 objects to the EFD Photo Library and we asked them on 2 Oct 2015 to send these objects over.[5]
Total	**7103**	

2.1.1 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



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, I	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; I	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&apos;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 1 Simple Mapping Table for Wolverhampton

[1] e.g. http://images.alinari.it/img/480/ACA/ACA-F-022924-0000.jpg

[2] e.g. http://www.horniman.ac.uk/media-collection/413/media-413331/feature.jpg

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

[4] e.g. http://img04.pars04.fr.topfoto.co.uk/imageflows/imagepreview-if3/t=topfoto&f=EUFD001241

[5] https://basecamp.com/2069212/projects/8450098/messages/39521744#comment_337565031

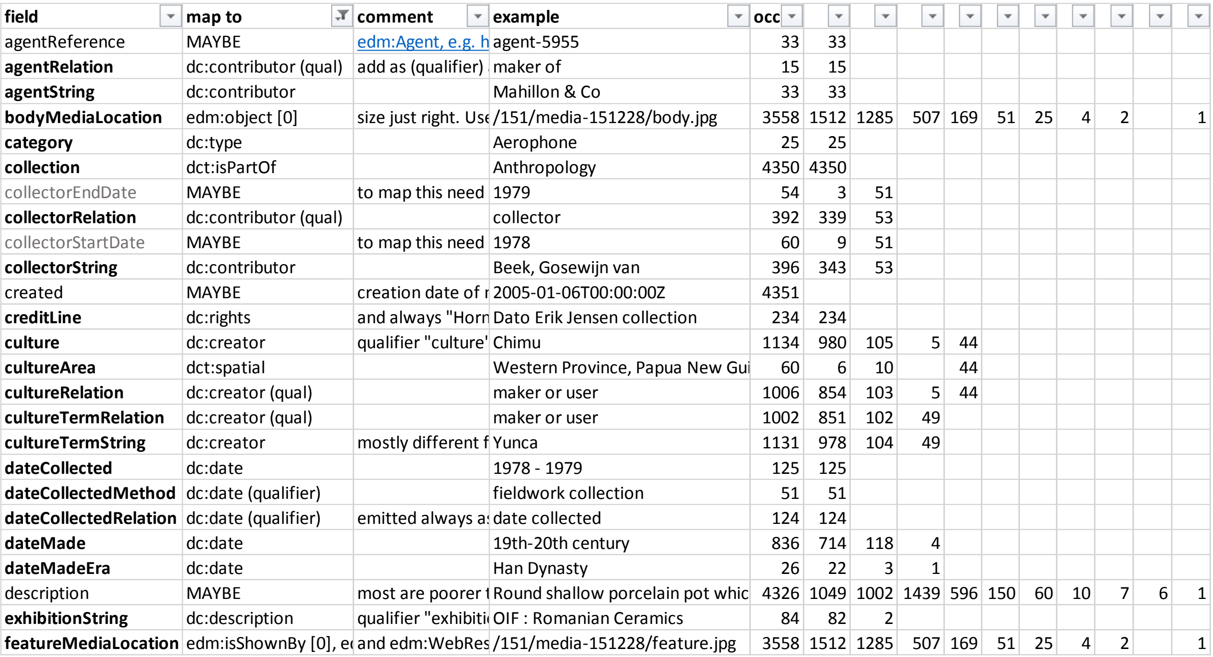
- Then we implement the mapping by fetching fields from the converted object, and putting them into an RDF::Trine graph (model):

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});
```

This takes care to emit proper language tags (always "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 Gui	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 "exhibiti	OIF : Romanian Ceramics	84	82	2								
featureMediaLocation	edm:isShownBy [0], e	and edm:WebRes	/151/media-151228/feature.jpg	3558	1512	1285	507	169	51	25	4	2		1

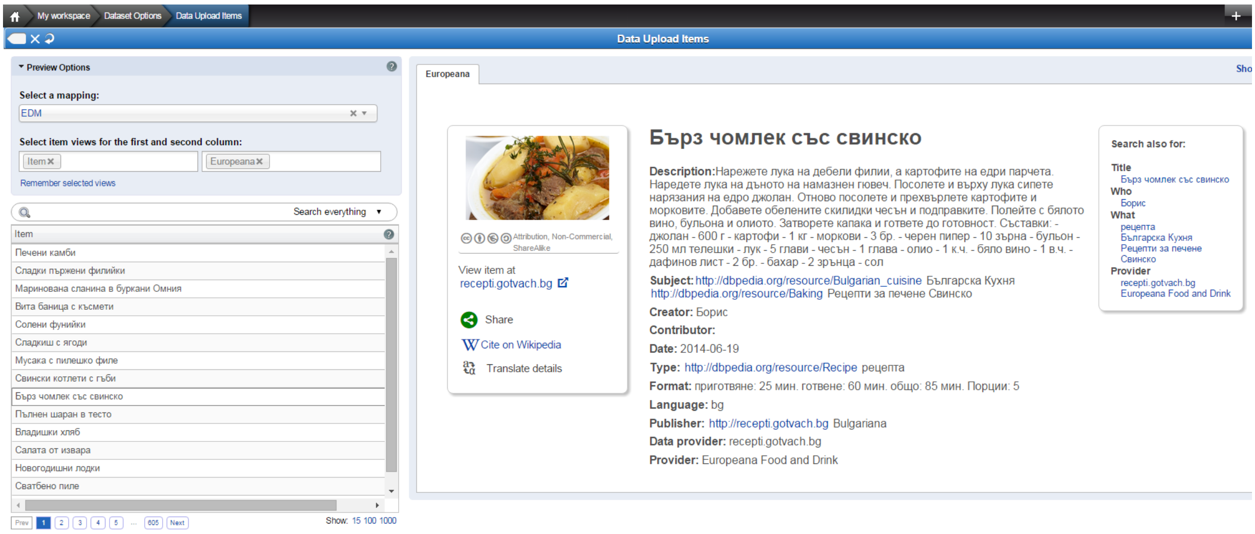
Figure 2 Complex Mapping Table for Horniman

2.1.2 BG Metadata Conversion and Submission

The conversion of Bulgarian Traditional Recipes (ONTO) was described in the previous report and the EDM was submitted to NTUA and EF in June. We also used a Perl script, but simpler than the ones described above.

It turned out that we need to do some additional work:

- EF prefers to get the data from MINT rather than a zip file, so 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



Item

Печени камби
Сладки пържени филийки
Маринована сланина в буркани Омния
Вита баница с късмети
Солени фунийки
Сладкиш с ягоди
Мусака с пилешко филе
Свински котлети с гъби
Бърз чомлек със свинско
Пълнен шаран в тесто
Владишки хляб
Салата от извара
Новогодишни лодки
Сватбено пиле

Бърз чомлек със свинско

Description:Нарежете лука на дебели филии, а картофите на едри парчета. Наредете лука на дъното на намазнен гювеч. Посолете и върху лука сипете нарязания на едро джолан. Отново посолете и прехвърлете картофите и морковите. Добавете обелените скилидки чесън и подправките. Полейте с бялото вино, бульона и олиото. Затворете капака и гответе до готовност. Съставки: - джолан - 600 г - картофи - 1 кг - моркови - 3 бр. - черен пипер - 10 зърна - бульон - 250 мл телешки - лук - 5 глави - чесън - 1 глава - олио - 1 к.ч. - бяло вино - 1 в.ч. - дафинов лист - 2 бр. - бахар - 2 зрънца - сол

Subject: http://dbpedia.org/resource/Bulgarian_cuisine Българска Кухня http://dbpedia.org/resource/Baking Рецепти за печене Свинско

Creator: Борис

Contributor:

Date: 2014-06-19

Type: http://dbpedia.org/resource/Recipe рецепта

Format: приготвяне: 25 мин. готвене: 60 мин. общо: 85 мин. Порции: 5

Language: bg

Publisher: http://recepti.gotvach.bg Bulgariana

Data provider: recepti.gotvach.bg

Provider: Europeana Food and Drink

Search also for:

Title
Бърз чомлек със свинско
Who
Борис
What
рецепта
Българска Кухня
Рецепти за печене
Свинско
Provider
recepti.gotvach.bg
Europeana Food and Drink

Figure 3 Preview of BG-ONTO Object Shown in MINT

- 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 some 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/Boiling_in_cooking
- http://dbpedia.org/resource/Stew
- http://dbpedia.org/resource/Microwave_oven
- http://dbpedia.org/resource/Batter_(cooking)
- http://dbpedia.org/resource/Baking
- http://dbpedia.org/resource/Frying
- http://dbpedia.org/resource/Steaming

The objects should be ingested by Europeana in mid-Sep 2015.

2.2 Elaborating the FD Classification Tree

Elaborating the EFD Classification by refinement of the FD categories is the main task of the semapp. We continued this task in the current period. More details are provided in [Alexiev 2015e] and [Tagarev 2015].

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.

By removing only 314 categories and their connections we were able to reduce the categories by 98%: from 880k to 17.5k FD-relevant categories.

In this period we continued this process of refinement:

- 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
- On the other hand, we added to the FD tree some branches that bottom-up evidence showed are needed (see next).

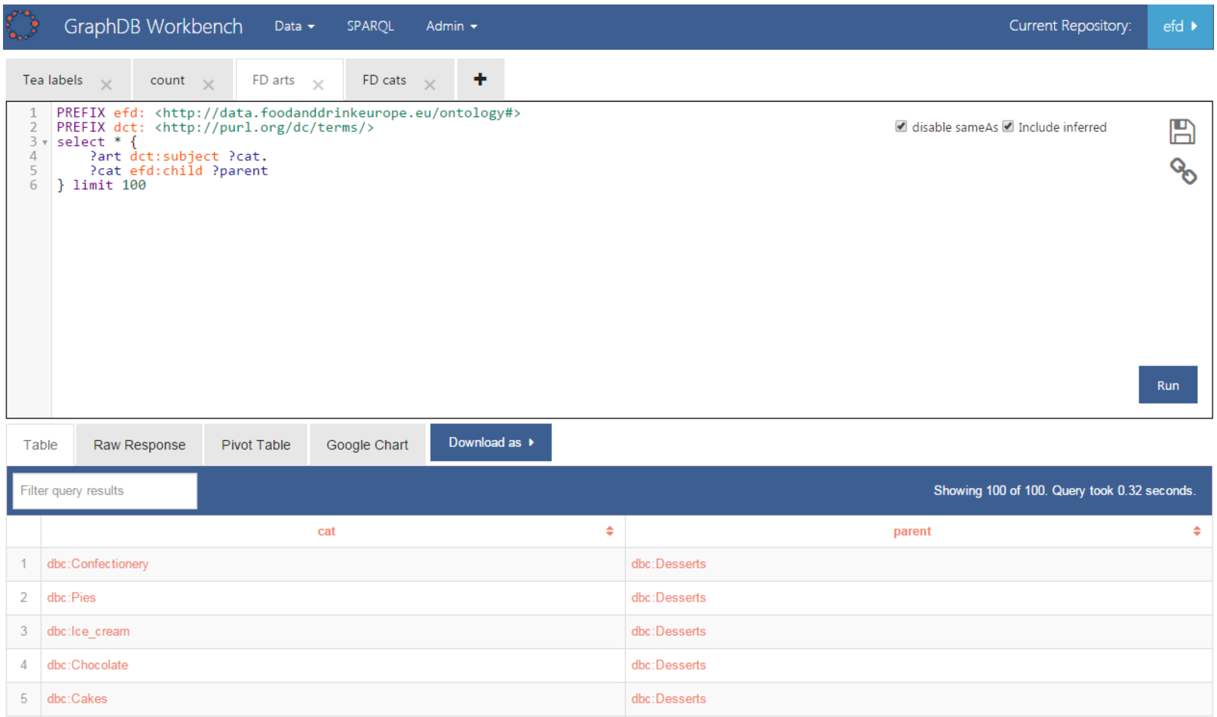
As of late Aug 2015, we have these statistics:

- FD categories: 13,275
- FD articles: 152,160
- cat<cat relations (parent categories): 21,008: 1.58 per cat
- art<cat relations (categorizations): 233,855: 1.53 per art, 17.6 per cat

2.2.1 Example SPARQL Queries

Let's use the EFD SPARQL endpoint[6] 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:

[6] http://efd.ontotext.com/sparql



```
PREFIX efd: <http://data.foodanddrinkeurope.eu/ontology#>
PREFIX dct: <http://purl.org/dc/terms/>
select * {
    ?art dct:subject ?cat.
    ?cat efd:child ?parent
} limit 100
```

☑ disable sameAs ☑ Include inferred

Showing 100 of 100. 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 4 FD Articles Query in Ontotext GraphDB Workbench

This simple query returns FD articles with their categorizations:

```
PREFIX efd: <http://data.foodanddrinkeurope.eu/ontology#>
PREFIX dct: <http://purl.org/dc/terms/>
select * {
    ?art dct:subject ?cat.
    ?cat efd:child ?parent
} limit 100
```

An even simpler query returns FD categories with their parents.

```
PREFIX efd: <http://data.foodanddrinkeurope.eu/ontology#>
PREFIX dct: <http://purl.org/dc/terms/>
select * {
    ?cat efd:child ?parent
} limit 100
```

It returns results like this:

Table 2 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 category[7] is a curious one, including characters like Mr Potato Head, Cipollino and Bananaman.

2.3 Wikipedia Editing

While we were working on Horniman object enrichment, we needed to add a number of things to Wikipedia to improve its FD coverage. See contribution list.[8]

- **Adding parent categories**. E.g. added major branches under FD: Hunting, Fishing, and Livestock. (The Horniman collection has a lot of Hunting objects)
- **Adding labels (redirects)**. E.g. added Muller (a copper device for mulling beer or keeping it warm) as label of Mulled_wine

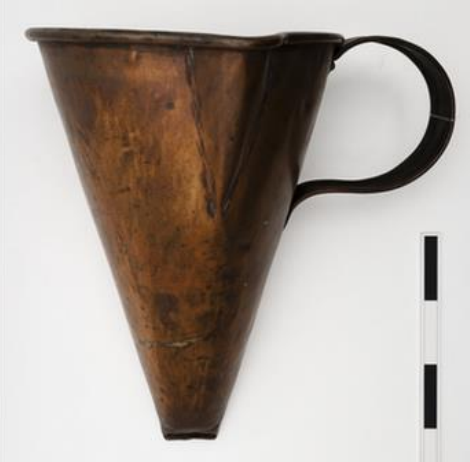


Figure 5 Muller from Horniman[9]

- Creating pages, e.g. **Shepherd's crook** and **Tumbler (glass)** by splitting text from existing pages. Added label "Crook (shepherd)"
- 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"
- Added 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").
- Added sections to pages, e.g. **Lovespoon#Wedding_Spoons**.
- Added categories, e.g. Libation "ceremonial pouring of water, wine, olive oil, etc. Added the category to categories Wine & Olive oil
- Added articles to categories, e.g. **Libation sticks, Rhyton**, **Patera** to **Libation**.
- Added a few illustrations, e.g. a phiala from the Panaguyrishte gold treasure (Used in ceremonial wine drinking or Libation) to article **Patera**. Unfortunately we

[7] https://en.wikipedia.org/wiki/Category:Fruit_and_vegetable_characters

[8] https://en.wikipedia.org/wiki/Special:Contributions/Vladimir_Alexiev

[9] http://www.horniman.ac.uk/object/25.29

couldn’t add illustrations from Horniman because the image license of that museum does not allow it.

2.4 Leveraging LOD

An important part of the EFD Classification approach is bottom-up Evidence propagation (positive feedback), i.e. processing articles or CHOs that are proven FD-relevant by other means to confirm and enlarge the tree:

- **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 (see previous section)
- **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 the tree and 123 were not. We added the appropriate ones to the tree by adding or adjusting categories.
- **DBtax**: this is a heuristic addition of types to DBpedia performed by the Italian DBpedia chapter. The types themselves are not always meaningful (e.g. **Zutho** 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
```

2.5 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.

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 Subjects[10] 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.

This is still work in progress and we may not be able to complete it before end-Oct 2015.

2.6 Place Hierarchy

Since we perform place enrichment (see next section), we want to use a place hierarchy in order to display a hierarchical Place facet.

Surprisingly, it turns out that DBpedia doesn't have a good place hierarchy:

- There is no uniform place hierarchy property. E.g. for dbo:Island, the property dbo:archipelago shows the parent island group, whereas dbo:location is the containing ocean or sea. 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 related YAGO classes, but 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

Therefore we decided to use GeoNames, which has a uniform property gn:parentFeature.

[10] http://id.loc.gov/vocabulary/ethnographicTerms/

- GeoNames has coreferences (links) to other datasets[11], of which 482k are links to Wikipedia (470k to en.wikipedia, 10k ru, 0.6k de).
- dbo:Place (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
- (Conversely, only 5.2% of GeoNames features are linked to en.dbpedia)

We hope that this 62% GeoNames coverage will be enough for the places used in our enrichments, and all their parent places. This is still work in progress.

2.7 Semantic Enrichment of Collections

We have performed enrichment on the 4 selected collections in 2 aspects:

- Topical enrichment using the FD tree. The Horniman collection was enriched semi-automatically: we verified the mapping of each of the 700 object types used by the museum and made appropriate corrections and additions. The other 3 were enriched automatically.
- Place enrichment using places from DBpedia. All 4 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 is highly ambiguous: there are at least 10 such provinces.

The results as of 1 Sep 2015 are as follows:

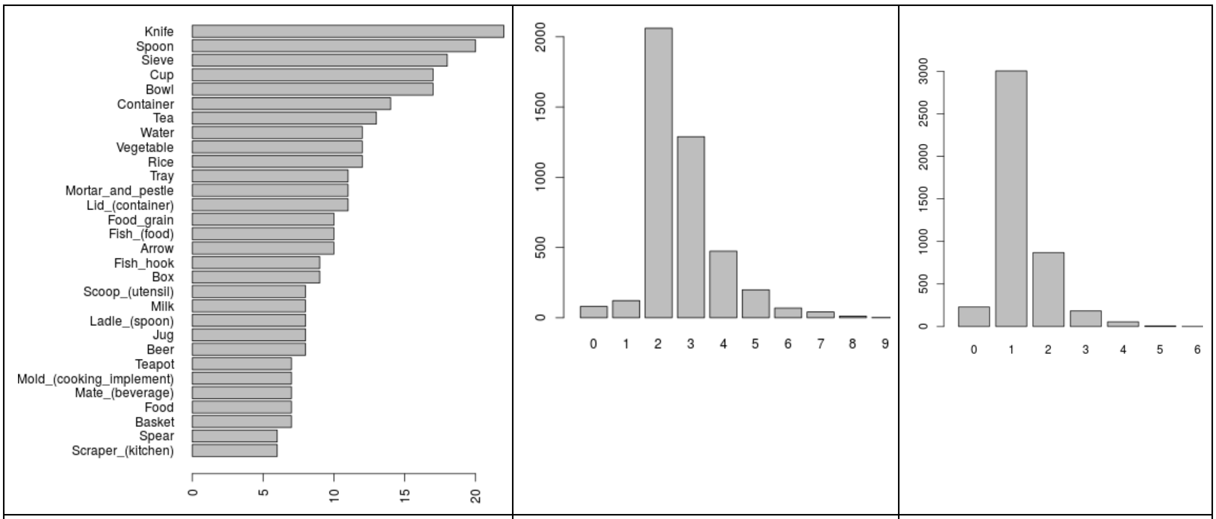
- Total objects from 4 providers: 7103
- Objects with at least one Place tag: 6567
- Objects with at least one FD tag: 5664 (there are some Alinari objects with few if any FD-related words)

Once we complete the Cultures dataset, we'll enrich with it as well. We'll use the same software that applies FD enrichments, so that will not take long.

[11] http://download.geonames.org/export/dump/alternateNames.zip

2.7.1 Horniman FD Statistics

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



Most popular tags

Tags per object

Tags per object (only most-specific)

Figure 6 Statistics of FD Tags in Horniman Objects

2.8 Evaluation of Enrichment

An important question concerns the quality of enrichment. It is estimated on random sample of objects by 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

We achieved the following results:

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 (that is the enrichment we'll use in the semapp).
- Nevertheless we'll inspect CHOs without a single FD enrichment and will add some: there are indeed some Alinari CHOs that have very few or even no FD-related keywords.

Place enrichment:

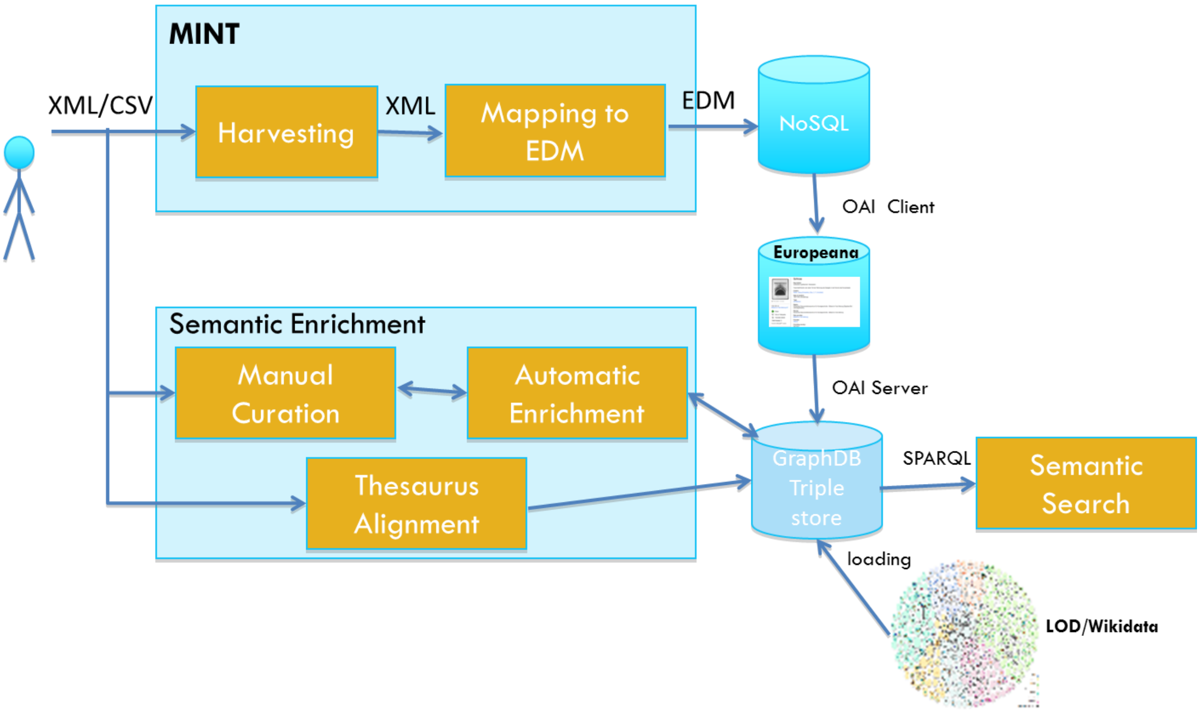
- The F-Measure is very high.
- The factual recall is seven higher because if a parent place is not recognized but its child place is recognized, the parent place will still be activated in the Places facet. E.g. in "Royal Library, Turin, Piedmont" we recognize

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.
- Another imprecision is that "tray" is mis-recognized as "Trayes", which we have corrected manually.

2.9 Semapp Architectural Design

The conceptual architecture of the semapp is shown below:



It is still unclear how to deliver enrichments to Europeana, because:

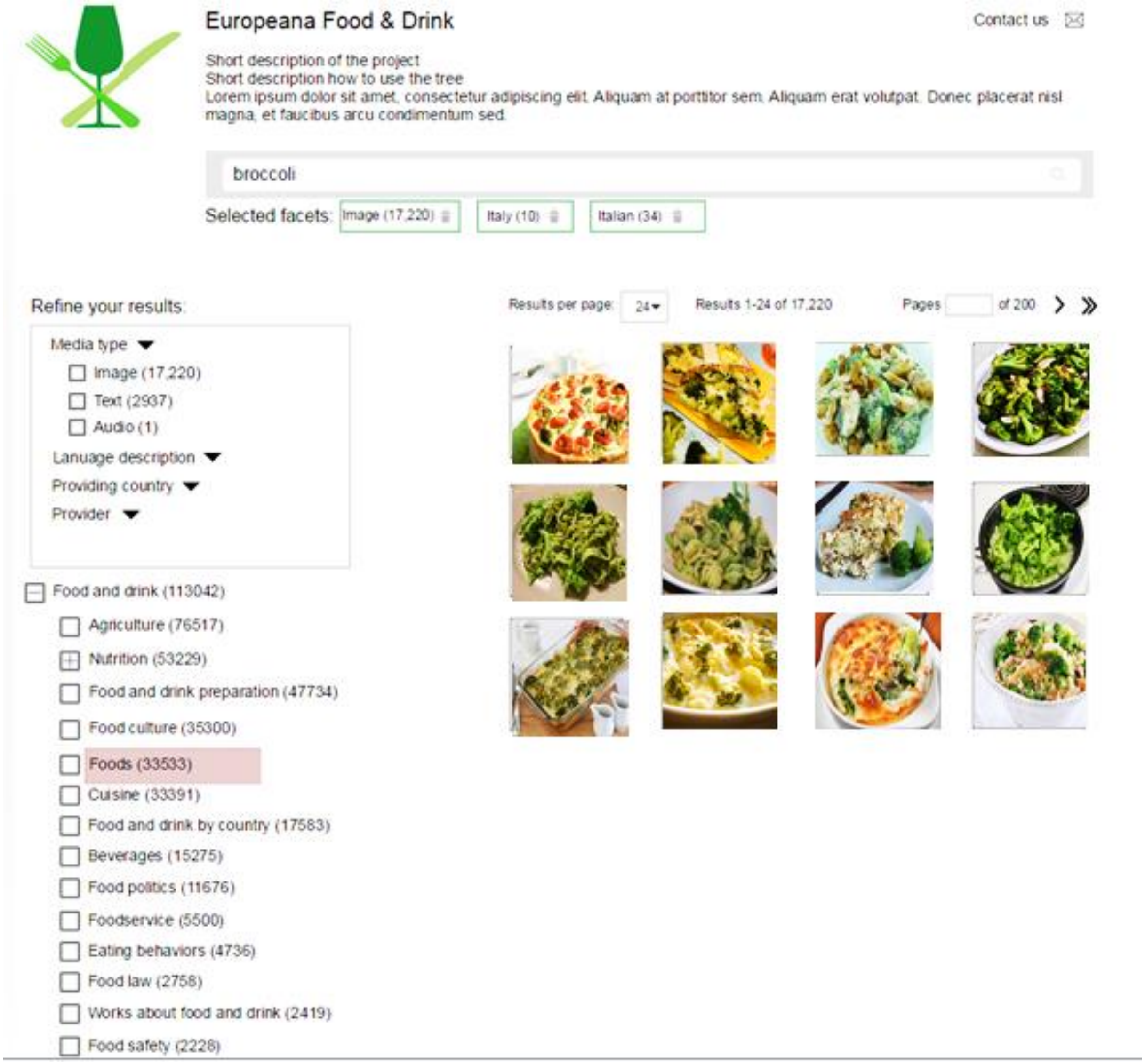
- Neither ONTO nor NTUA can add enrichments in provider collections in MINT
- EF cannot take enrichments for a number of objects at once as a single data file

We raised this question in Jul 2015[12] and are still looking for the easiest solution.

[12] https://basecamp.com/2069212/projects/7205992/messages/45430278

2.10 Semapp UI Design and Web Page

We developed a basic wireframe and mock-up for the semapp.[13] It will be similar to Europeana (search, faceting, pagination, etc), but will provide additional semantic & hierarchical facets.



Europeana Food & Drink

Contact us

Short description of the project
Short description how to use the tree
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam at porttitor sem. Aliquam erat volutpat. Donec placerat nisl magna, et faucibus arcu condimentum sed.

broccoli

Selected facets: Image (17,220) | Italy (10) | Italian (34)

Results per page: 24 | Results 1-24 of 17,220 | Pages of 200

Refine your results:

Media type
- □ Image (17,220)
- □ Text (2937)
- □ Audio (1)

Lanuage description

Providing country

Provider

□ Food and drink (113042)
- □ Agriculture (76517)
- ⊞ Nutrition (53229)
- □ Food and drink preparation (47734)
- □ Food culture (35300)
- □ Foods (33533)
- □ Cuisine (33391)
- □ Food and drink by country (17583)
- □ Beverages (15275)
- □ Food politics (11676)
- □ Foodservice (5500)
- □ Eating behaviors (4736)
- □ Food law (2758)
- □ Works about food and drink (2419)
- □ Food safety (2228)

We also created a webpage together with CT[14] and a detailed description of the semapp[15].

2.11 Europeana CHO Discovery

A very 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 2015c sec.2.12]. Focusing on the technical side, this presents significant challenges:

- We identified 152k articles in 13k 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).

[13] https://live.uxpin.com/3adf4c6d0e75ed13bef0408a09adc837c228824b#/pages/25651569

[14] http://foodanddrinkeurope.eu/professional-applications/semantic-demonstrator/

[15] http://foodanddrinkeurope.eu/wp-content/uploads/2015/09/EFD-Semantic-Demonstrator.pdf

- 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.

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

2.11.1 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 OCRed (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 OCRed 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" letter[16] that is related to Tea since it mentions "Brooke Bond".

2.11.2 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.

2.11.3 FD Classifier

We used some machine learning techniques to create a FD Classifier. This module can predict whether an object is FD-related or but 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 (post popular word **stems**) 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

[16] http://www.europeana.eu/portal/record/2020601/attachments_52959_4640_52959_original_52959_jpg.html

- 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.

2.11.4 Europeana Problems

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

Improper Enrichment with Narrower Terms

For example this cylinder jar[17] (also see provider site[18]) has provider terms "Zylinderhalsgefäß"@de = "cylinder jar"@en, "Gefäß"@de="vessel"@en;

[17] http://www.europeana.eu/portal/record/08501/Athena_Update_ProvidedCHO_Bildarchiv_Foto_Marburg_obj_20727191_410_848.html

[18] http://www.bildindex.de/dokumente/html/obj20727191#|home

"Angewandte Kunst"@de = "applied art"@en. It is correctly enriched with concept "vessels (containers)"@en = "Gefäß (Behälter)@de from the Partage vocabulary[19].

- However, it is incorrectly enriched with 26 AAT **narrower** terms of "vessels": esker, bokser, ... samovars.
- Also, it is irrelevantly enriched with 2 GEMET **broader** 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"[20] are not samovars. We raised appropriate issues to Europeana.[21]

In contrast, the semantic approach provides multiple attested labels for the concept: Samovar; Electric samovar; Semaver; Samowar; Zavarka. We found 960 objects with "Samowar"[22]. 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"[23] 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 onbject 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.

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.

2.12 Task Forces

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

- Evaluation and Enrichments[24]. 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

[19] http://partage.vocnet.org/html/part00083

[20] http://www.europeana.eu/portal/search.html?query=cc_skos_prefLabel:samovar

[21] http://www.assembla.com/spaces/europeana/tickets/2044-enrichment-shouldn--39-t-add-narrower-broader-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

[22] http://www.europeana.eu/portal/search.html?query=samowar

[23] http://www.europeana.eu/portal/search.html?query=beer

[24] http://pro.europeana.eu/get-involved/europeana-tech/europeanatech-task-forces/evaluation-and-enrichments

enrichments were evaluated and compared against 5 other trial submissions, by projects such as LoCloud[25]. 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 Education[26] 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) and second meeting (6-7 Oct 2015 in Warsaw).

2.13 Publications and Presentations

We delivered 1 presentation and 1 paper in the current period (see References below):

[Alexiev 2015b] This presentation shows the work that ONTO completed as part of Europeana Creative to establish 2 new access channels to Europeana data: EDM SPARQL repository and OAI PMH server. 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.

3 Project Management

This section describes scoping, timing, and resource considerations for the semapp task.

3.1 Scope for Oct 2015

For 31 Oct 2015 we will implement the semapp in the scope outlined in the previous periodic report (D3.20a), with the following exceptions:

- Correlation from X Cuisine to place/culture X
- We'll limit the semapp to semantic search, browse and "lightbox" (simple gallery), deferring geographic search or timeline for later.

We will also publish all enriched datasets, the FD category tree, and the simple EFD ontology that we've developed for this data.

3.2 Extended Scope

After submitting D3.20, ONTO plans to continue work on the semapp, as explained in D3.20a.

- At the Project Management Board (PMB) meeting on 15 April 2015 in London it was agreed that ONTO can use 40k EUR in its budget (originally slated for subcontracting) to continue core development on the semapp. On 23 Sep 2015 we wrote an Extension Request to that effect, providing detailed justification.
- On 30 Sep 2015 it was decided that ONTO should continue work on the semapp and Europeana Discovery as part of its allocation in WP5, helping other product partners use more FD-related CHOs.

[25] http://locloud.eu/Resources/LoCloud-enrichment-services

[26] http://pro.europeana.eu/publication/europeana-for-education-policy-recommendations

4 Conclusions

This document has described all work done on the semapp (D3.20) in the first 5.5 months of development, and the progress achieved. We are on track to achieve the scope for Oct 2015, and are ready to continue with further development.

5 References

[Alexiev 2015a] 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

[Alexiev 2015b] Vladimir Alexiev and Dilyana Angelova. O is for Open: OAI and SPARQL interfaces for Europeana. In *Europeana Creative Culture Jam*, Vienna, Austria, July 2015

[Alexiev 2015c] 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

[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