europeana
food and drink

Grant Agreement 621023

# *Europeana Food and Drink*

# Semantic Demonstrator
# M18 Progress Report

| | |
|---|---|
| **Deliverable number** | *D3.20a* |
| **Dissemination level** | *CO* |
| **Delivery date** | *30 June 2015* |
| **Status** | *Final* |
| **Author(s)** | *Vladimir Alexiev (ONTO)* |

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

# Abstract

This document describes the progress on developing the EFD Semantic Demonstrator for the first 2.5 months. We describe all work performed between 1 April 2015 and 15 June 2015, the achieved results and project management considerations.

This is not a formal deliverable but a periodic progress report.

# Revision History

| Rev | Date | Author | Org | Description |
|---|---|---|---|---|
| v0.1 | 22/06/2015 | Vladimir Alexiev | ONTO | Initial version |
| v0.2 | 24/06/2015 | Andrei Tagarev, Nadia Ognianova | ONTO | Review |
| v0.3 | 26/06/2015 | 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This deliverable contains original unpublished work except where clearly indicated otherwise.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 first 2.5 months. Work on the semapp started in April 2015, after completing D3.19 Semantic Demonstrator Specification [Alexiev 2015d]. We describe the work performed between 1 April 2015 and 15 June 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: descries work done to date:

- Metadata sample collection
- BG metadata conversion and submission
- 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 of the Horniman collection
- Getty AAT coreferencing
- Participation in task forces
- 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 |
| 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 |
| UTF-8 | The most commonly used Unicode Transformation Format |
| WMF | WikiMedia Foundation |

# 2 Work Done

## 2.1 Metadata Sample Collection

To date the project has not yet submitted EDM metadata to Europeana. The first collections are scheduled to be submitted in June 2015: AT ONB, BE KMKG, BG ONTO, LT VUFC, PL ICIMSS.

This lack of content was a major obstacle to starting development on the semapp. In order not to block development, ONTO 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 (ultimately the semapp will work only with EDM), but there was no other way around this obstacle.

The status of metadata samples per collection 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's 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 the different collections follow in sub-sections.

### 2.1.1 AT-ONB (Austrian National Library)

- Available data: 30 objects, each object contained in separate xml file
- Downloaded from OAI endpoint. ESE not EDM format
- Language is marked for each entry. If different fields in the same object are in different languages, they seem to be tagged correctly.
- Links to images are given as actual URL and marked as such: <dc:identifier>http://www.bildarchivaustria.at/Preview/307063.jpg</dc:identifier>

AT-ONB/307063.xml

- title@de: Erster Wiener Volksküchenverein (First Wien folk club kitchen). German word formation makes for some very long words
- Creator is the studio that took the photo

### 2.1.2 BE-CAG (Centrum Agrarische Geschiedenis)

- Available data: 6 categories with about a dozen objects each, each category is contained in a separate excel file
- No language tags. Entries are mostly in Dutch but some French and possibly German mixed in.
- No image links
- Also included is the CAG thesaurus (non-hierarchical) and a list of keywords used as browse terms on the public website (broader than a thesaurus). Both are in Dutch
- These lists are filtered for FD-relevance (yes/no). E.g. these are relevant:
  Zwarte bes (thesaurus)
  zurkelpot (keyword)

### 2.1.3 BE-KMKG (Koninklijke Musea voor Kunst en Geschiedenis)

- Website: http://www.kmkg-mrah.be/
- Online catalog: http://www.carmentis.be/
- Available data: objects split in two xml files (China and Patacons) in LIDO
- Data comes with Europeana URLs but these have not been finalized: the file "China" does not contain proper links. The file "Patacons" does.
- Items are tagged with materials from a specific thesaurus: <lido:termMaterialsTech lido:type="Material"><lido:conceptID

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

lido:type="MuseumPlusThiID">51095</lido:conceptID><lido:term lido:addedSearchTerm="yes">Pijpaarde</lido:term></lido:termMaterialsTech>

- Direct image URLs contained in the data: <lido:linkResource lido:formatResource="image/jpeg">http://carmentis.kmkg-mrah.be/eMuseumPlus?service=ImageAsset&module=collection&objectId=197770&resolution=superImageResolution</lido:linkResource>
- 18 Jun 2015: 646 records will be available through OAI[3] after KMKG completes their conversion to EDM

Object 199284:

- Available at carmentis[4]
- All objects in the first collection have object type (FR only): Patacon (application made from Kaolin) or Moule a patacon (application mold)
- It is not this dish: https://fr.wikipedia.org/wiki/Patacón (dish made with pieces of flattened and fried green platain)
- Not sure how is this relevant to FD

### 2.1.4 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: will submit only recipes with images. 6419 have "CUISINE: Българска Кухня". Already converted to EDM (see 2.2)

### 2.1.5 CY-CFNM (Cyprus Food and Nutrition Museum)

- Data is not encoded in UTF-8 but in ISO-8859-7
- Available data: Two sets of data but no explanation what they are. Both come as csv and excel file but they seem to be identical
- Data is in Greek, some category names are in English (ID, TITLE, DESCRIPTION) but others are in Greek.
- Data seems to be sparse: everything has ID and NAME but other fields are not always present
- Basic topics (e.g. Μιλλόπιτα, Πίτα με λαρδί, Ελιόπιτα, Ελιόψωμο) are not found in http://el.wikipedia.org. This is bad news for semantic enrichment.
- No images

### 2.1.6 ES-CAT (Generalitat de Catalunya)

About 10k records of different content types. Counts per collection:

| | |
|---|---|
| 2543 | http://calaix.gencat.cat/oai/request?verb=ListRecords&metadataPrefix=ese&set=col_10687_7709 |
| 4666 | http://calaix.gencat.cat/oai/request?verb=ListRecords&metadataPrefix=ese&set=col_10687_53660 |
| 200 | http://calaix.gencat.cat/oai/request?verb=ListRecords&metadataPrefix=ese&set=col_10687_10990 |
| 2 | http://calaix.gencat.cat/oai/request?verb=ListRecords&metadataPrefix=ese&set=col_10687_23851 |
| 1545 | http://calaix.gencat.cat/oai/request?verb=ListRecords&metadataPrefix=ese&set=com_10687_58331 |
| 670 | http://calaix.gencat.cat/oai/request?verb=ListRecords&metadataPrefix=ese&set=com_10687_53055 |

[3] http://panic.image.ntua.gr:9876/foodanddrink/oai?verb=ListRecords&set=1013&metadataPrefix=rdf

[4] http://carmentis.kmkg-mrah.be/eMuseumPlus?service=ExternalInterface&module=collection&objectId=199284&viewType=detailView

| | |
|---|---|
| 1675 | http://calaix.gencat.cat/oai/request?verb=ListRecords&metadataPrefix=ese&set=col_10687_51373 |
| 336 | http://calaix.gencat.cat/oai/request?verb=ListRecords&metadataPrefix=ese&set=col_10687_10 |

Formats at http://calaix.gencat.cat/oai/request?verb=ListMetadataFormats:

- ese: best choice
- oai_dc: isShownAt is in dc:identifier, but there's another that's a mere string
- qdc: like oai_dc but repeats namespaces in every element
- rdf: is some custom format with root ow:Publication. Same defect as oai_dc

Object 10687/233[5]

- Web page (edm:isShownAt)[6]: works
- Preview (edm:object)[7]: available but a bit small
- Image (edm:isShownBy)[8]: not present in metadata, but can be obtained from object by chopping the double file extension

That DSpace OAI server is quite powerful:

- Allows you to look at objects' metadata even before downloading them, if requested from a browser
- Returns the pure metadata (wrapped in OAI), if requested with curl

### 2.1.7 HU-MKVM (Magyar Kereskedelmi és Vendéglátóipari Múzeum)

- Available Data: 53 LIDO records, each in an individual xml file.
- All titles and many descriptions are in English but many entries are in Hungarian (especially names).
- UTF-8 encoding problems have been fixed
- Image links have been fixed[9]
- Object pages not working[10]

16 June 2015: contacted about obtaining the complete collection because it's also in EN and we can start semantic enrichment:

- How many total records do you plan to provide?
- Can you provide them all at this time, even though in LIDO? (we currently have a sampling of 53)
- How many of your objects have EN translation? E.g. 8460 doesn't have one
- Can you provide indication of language in the textual fields? We're not ready to do text analysis over HU, we can handle EN only for now. If you can't, we can apply a language recognizer component but it's better if you have an explicit indication.

### 2.1.8 IE-LGMA (Local Government Management Agency)

- Available data: only 3 incomplete objects in an excel file
- Image data doesn't include links.

### 2.1.9 IT-Alinari (Fratelli Alinari)

Alinari is a long-time photo agency

[5] http://calaix.gencat.cat/oai/request?verb=GetRecord&identifier=oai:calaix.gencat.cat:10687/233&metadataPrefix=ese

[6] http://calaix.gencat.cat/handle/10687/233

[7] http://calaix.gencat.cat/bitstream/handle/10687/233/afoto_2174_000013_col.jpg.jpg

[8] http://calaix.gencat.cat/bitstream/handle/10687/233/afoto_2174_000013_col.jpg?sequence=1

[9] E.g. http://www.museum-digital.de/hu/portal/images/201504/100h_10133036759.jpg

[10] E.g. http://www.museum-digital.de/hu/portal/index.php?t=objekt&oges=8389

- Available data: 50 objects, provided in both English and Italian (in separate excel files).
- Have names for the data fields.
- No image links

### 2.1.10 IT-ICCU (Istituto centrale per il catalogo unico delle biblioteche italiane)

- Available data: 27 Objects from Bibliotheca Alexandrina in individual xml files.
- Note: data comes with RDF notation.
- Entries are in Italian, Latin, Spanish. Language is not indicated
- Image links[11]: work, "&amp;" needs to be un-escaped

Object RMLE006787.xml

- title@es: Libro d'guisados manjares y potajes intitulado libro de cozina ... por maestre Ruberto!
- description@it: Il nome dell'A. si ricava dal Prologo sul v. del front
- creator: same person given in two creator records
- creator: "Nola , Ruperto : de": the punctuation is weird but significant. Means "Ruperto de Nola"
- creator: "Robert": compared to original record[12] that has "Robert <mestre>", the designation <mestre> is omitted but is significant. This is confirmed by available enrichment URLs:
    - https://www.wikidata.org/wiki/Q3437000: "Robert de Nola", "Ruperto de Nola", "Mestre Robert"
    - https://en.wikipedia.org/wiki/Robert_de_Nola
    - https://es.wikipedia.org/wiki/Mestre_Robert

### 2.1.11 LT-VUFC (Vilnius University)

- Available data: a single object in xml file with RDF notation
- RDF URIs are mostly in Lithuanian
- No image links

Although we have only 1 object from this collection, the institution uses a thesaurus (see image below) and is interested to work on coreferencing and enrichment, which is a very positive sign. They wrote: "We are very interested in semantic enrichment that you described and we would like to contribute in any way we can.

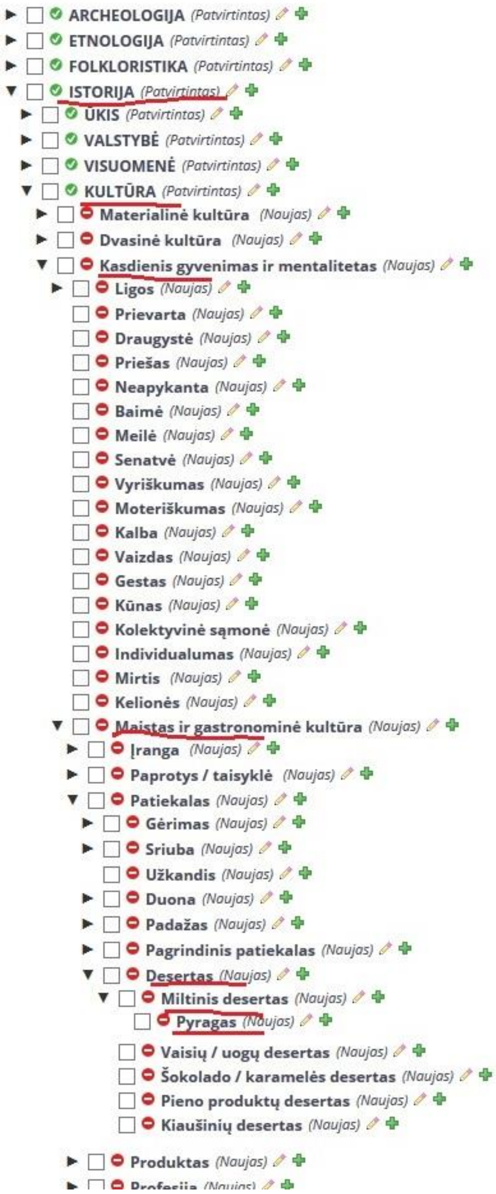
[11] http://iccu01e.caspur.it/ms/thumb.php?size=300&font=0.8&id=oai%3Awww.internetculturale.sbn.it%2FTeca%3A20%3ANT0000%3ABVEE001804

[12] http://www.culturaitalia.it/opencms/viewItem.jsp?language=en&case=&id=oai%3Awww.internetculturale.it%2Fmetaoaicat%3Aoai%3Awww.internetculturale.sbn.it%2FTeca%3A20%3ANT0000%3ARMLE006787

"So far we have about 500 images of scanned recipes with metadata. We can export them as XMLs. The classification for the image is integrated into the xml. We also can export the whole Thesaurus if needed.

"We use local thesaurus with Lithuanian terms for classification. There is a possibility to provide translation in other languages as synonyms adding them to our Thesaurus and then exporting into xml. But in this case, we have to go through all terms and translate them at least into English in order to have multilingual dimension. We can do this of course if this is what is needed. But as I understood, the semantic classification tool could do this automatically linking terms through Wikipedia? If this is the case - it would be a great option for us. At any rate, we would be happy to assist you during the process to have the best result possible.

"The example is a recipe of "Omelette with onions". There are 3 terms (of the lowest level) added to the classification: "Omelette", "Eggs/product from eggs" and "Onion". We also included the hierarchy in XML; every term has a link to the Thesaurus, as well as its unique ID."



- ▶ □ ✓ ARCHEOLOGIJA (Patvirtintas)
- ▶ □ ✓ ETNOLOGIJA (Patvirtintas)
- ▶ □ ✓ FOLKLORISTIKA (Patvirtintas)
- ▼ □ ✓ ISTORIJA (Patvirtintas)
  - ▶ □ ✓ ŪKIS (Patvirtintas)
  - ▶ □ ✓ VALSTYBĖ (Patvirtintas)
  - ▶ □ ✓ VISUOMENĖ (Patvirtintas)
  - ▼ □ ✓ KULTŪRA (Patvirtintas)
    - ▶ □ ⊖ Materialinė kultūra (Naujas)
    - ▶ □ ⊖ Dvasinė kultūra (Naujas)
    - ▼ □ ⊖ Kasdienis gyvenimas ir mentalitetas (Naujas)
      - ▶ □ ⊖ Ligos (Naujas)
      - □ ⊖ Prievarta (Naujas)
      - □ ⊖ Draugystė (Naujas)
      - □ ⊖ Priešas (Naujas)
      - □ ⊖ Neapykanta (Naujas)
      - □ ⊖ Baimė (Naujas)
      - □ ⊖ Meilė (Naujas)
      - □ ⊖ Senatvė (Naujas)
      - □ ⊖ Vyriškumas (Naujas)
      - □ ⊖ Moteriškumas (Naujas)
      - □ ⊖ Kalba (Naujas)
      - □ ⊖ Vaizdas (Naujas)
      - □ ⊖ Gestas (Naujas)
      - □ ⊖ Kūnas (Naujas)
      - □ ⊖ Kolektyvinė sąmonė (Naujas)
      - □ ⊖ Individualumas (Naujas)
      - □ ⊖ Mirtis (Naujas)
      - □ ⊖ Kelionės (Naujas)
      - ▼ □ ⊖ Maistas ir gastronominė kultūra (Naujas)
        - ▶ □ ⊖ Įranga (Naujas)
        - ▶ □ ⊖ Paprotys / taisyklė (Naujas)
        - ▼ □ ⊖ Patiekalas (Naujas)
          - ▶ □ ⊖ Gėrimas (Naujas)
          - ▶ □ ⊖ Sriuba (Naujas)
          - □ ⊖ Užkandis (Naujas)
          - ▶ □ ⊖ Duona (Naujas)
          - ▶ □ ⊖ Padažas (Naujas)
          - ▶ □ ⊖ Pagrindinis patiekalas (Naujas)
          - ▼ □ ⊖ Desertas (Naujas)
            - ▼ □ ⊖ Miltinis desertas (Naujas)
              - □ ⊖ Pyragas (Naujas)
            - □ ⊖ Vaisių / uogų desertas (Naujas)
            - □ ⊖ Šokolado / karamelės desertas (Naujas)
            - □ ⊖ Pieno produktų desertas (Naujas)
            - □ ⊖ Kiaušinių desertas (Naujas)
        - ▶ □ ⊖ Produktas (Naujas)
        - ▶ □ ⊖ Profesija (Naujas)

We examined http://lt.wikipedia.org for coverage of the FD domain and interlanguage links to http://en.wikipedia.org. Looking at the provided example:

1. there is https://lt.wikipedia.org/wiki/Omletas (Omelettes) and https://lt.wikipedia.org/wiki/Kiaušinienė (Fried eggs)
2. both have links to EN
3. both have appropriate category https://lt.wikipedia.org/wiki/Kategorija:Kiaušinių_valgiai (Egg dishes)
4. that category has EN link
5. that category has appropriate super-categories, leading to root https://lt.wikipedia.org/wiki/Kategorija:Valgiai_ir_gėrimai (Food and drink)
6. the root category has EN link

This is the absolute perfect situation:

- Even if 2,3,4,5 were missing, we could still figure out that the two terms in 1 are about Food and drink, because of 6: we'll be merging the cat hierarchies across languages.
- With 2 present we can have cross-language conceptual search of "Omelet"
- Even if 2 was missing, if 3,4 are present: a user can still browse down to "Egg dishes" and find your objects, but will have to read the label "Kiaušinienė" in LT

In other words, we'll be leveraging inter-language links wherever we find them: at term level or at any category level.

The Onion situation is also good:

- https://lt.wikipedia.org/wiki/Svogūnas is a redirect to https://lt.wikipedia.org/wiki/Valgomasis_svogūnas. We catch this, interpreting "Svogūnas" as another label for the same concept.
- It has EN link
- It has categories (translated to EN) Leafy vegetables < Vegetables < Foods < Food and drink
- Plenty of these have EN links

Note: on https://lt.wikipedia.org/wiki/Kategorija:Česnakiniai, Google translates 3 of the articles as Garlik but they are different kinds, with different names in LT. Similar in BG: we have more terms for Garlik than EN

### 2.1.12 PL-IMMS (Institute for Management of IT Systems)

- Two datasets: BIKOP with 1246 and PB with 605 objects. Each object has individual xml file.
- The objects are not relevant to FD. Asked to make a selection of FD items: could just point to them in BIKOP and PB, or post a new dataset.
- Data entries are in English and Polish with each term in an entry marked with the appropriate language.
- Image URLs[13] work but are slow since the size is very large. Need to un-escape "&amp;"

### 2.1.13 UK-Horniman (Horniman Museum and Gardens)

- Available data: 4350 objects in the Food and Feasting subject[14] from Solr API[15] (uses standard Solr syntax)
- Single file in XML (each object in element <doc>) or JSON (add parameter "&wt=json")

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.

[13] http://www.pictures-bank.eu/index.php?action=przegladaj_zdjecie&id=52960

[14] http://www.horniman.ac.uk/collections/browse-our-collections/authority/subject/identifier/subject-322

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

- 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

Object 46798[16] (first one returned):

- Applied terms ("Related subjects") are at bottom:
  theme: Food and Feasting
  object name: goblets (food service)
  material: wood
- We don't need the "theme" (it's fixed in the query)
- objectNamePath: hierarchical object type concepts (3 levels)
  term-504836 term-504875 term-503747.
  Corresponds to: tools & equipment: general> food service> goblets (food service)
- materialTechniquePath: hierarchical material type concepts (3 levels)
  term-1016217 term-1015458 term-1015488
  Corresponds to: wood

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. See 2.10 for details.

### 2.1.14 UK-TopFoto (Topham Partners)

- Available data: 32 records in individual xml files
- 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

Got 3 lists of keywords

- HOM_FoodAndDrink: foods & drinks, e.g. cakes/pastries, buns & scones, wine gadgets, competitions
- IND_Catering: catering industry, e.g. accommodation, boarding houses, cafes, interiors, exteriors
- IND_FoodAndDrink: food & drink industry, e.g. baking, barrels, bottling, cider making, mills, presses

The lists are in PDF, with a hierarchy of 3 levels. Unfortunately they don't constitute a proper thesaurus and cannot be used directly for enrichment, since many labels are incomplete and can only be interpreted in context. E.g.:

- meet>abbatoirs>misc: "misc" alone is not related to FD
- cakes/pastries> finished> 1930s: neither "finished" nor "1930s" alone are related to FD.

Furthermore, composite labels like "cakes/pastries" and "tarts & pies" cannot be used for enrichment if they appear in objects as separate labels (cakes, pastries, tarts, pies)

### 2.1.15 UK-Wolverhampton (Wolverhampton City Council)

- Available data: 438 objects in a single xml file
- All data is in English although language isn't indicated anywhere

[16] http://horniman.ac.uk/collections/browse-our-collections/object/46798

- 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.jpg[17]
- 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
- Objects with images: 259 (59%). Number of images: 788. Images per object: 3.04 (for those that have at least one)

This is the second collection we've selected for enrichment.

## 2.2 BG Metadata Submission

ONTO will be one of the first EFD institutions to submit FD objects to Europeana. See 2.1.4 for details about the BG Recipes collection. We have already converted the collection to EDM, discussed details with the partners, and submitted to NTUA[18]. We hope it will be ingested by Europeana by the end of June 2015. We have submitted a lot more than originally promised: 9.5k recipes.

We have already provided some enrichments in the metadata (but more is needed):

- 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

## 2.3 Semantic Knowledge Base

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

- Installed and configured the Ontotext GraphDB semantic repository 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 2015d] sec. 2.3.2)
- Developed a small EFD ontology to hold classification data (e.g. FD-relevant parent category links, "not relevant" judgements, scoring counts, etc). We will publish it at http://efd.ontotext.com/ontology# in Oct 2015

In the near future we'll be loading Getty AAT (after we complete coreferencing, see 2.11) and perhaps Wikidata (for additional labels).

[17] http://cdn.collectionsbase.org.uk/wagmu/wams/oa76_p1.jpg

[18] https://basecamp.com/2069212/projects/8450098/messages/44028873

## 2.4 EFD Classification

Elaborating the EFD Classification by refinement of the FD categories is the main task of the semapp. It was also the second-most effort-intensive task for the period, after Metadata sample collection.

As explained in [Alexiev 2015c 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).

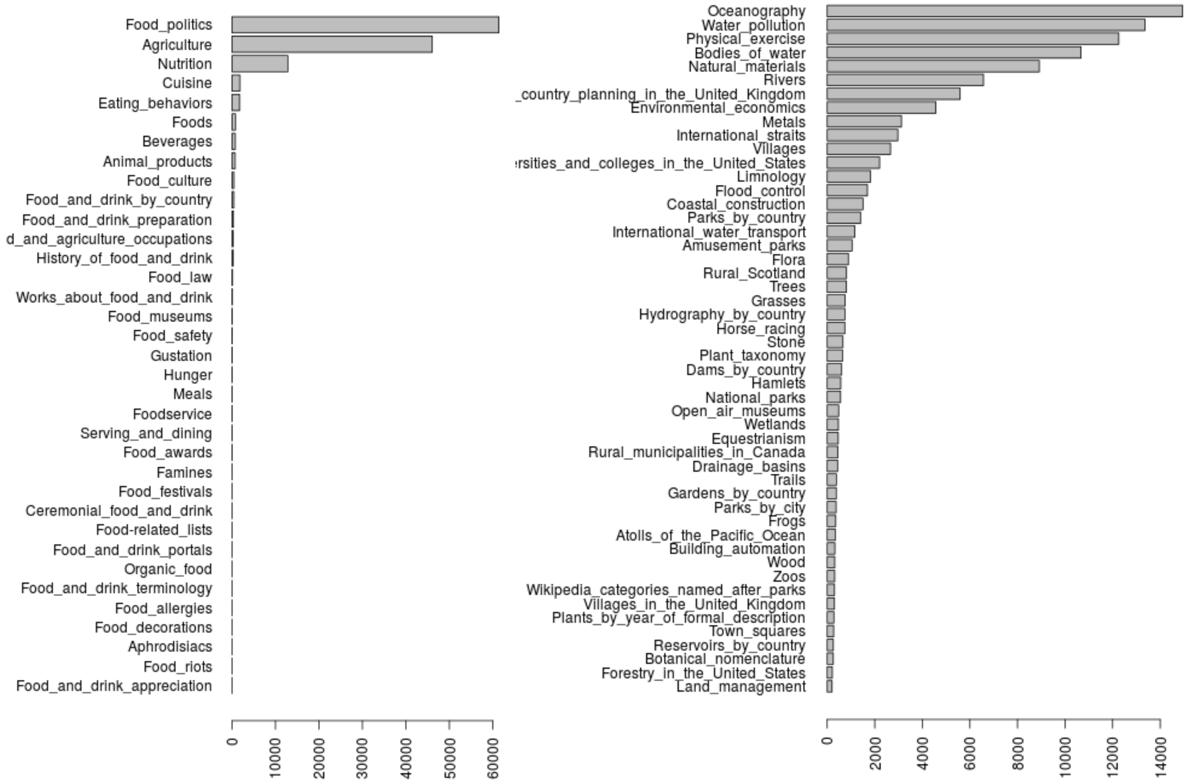


***Figure 1 Most populous categories reachable from FD: at level 2 (left), at level 5 (right)***

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 (see 2.6)
- A tool for manual curation of the tree (chopping out irrelevant branches)
- Evidence-based feedback (see 2.9)

As a result we were able to reduce the categories by 98%: from 880k to **17.5k FD-relevant categories**. This excellent result was achieved by removing only **314** categories and their connections.

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:

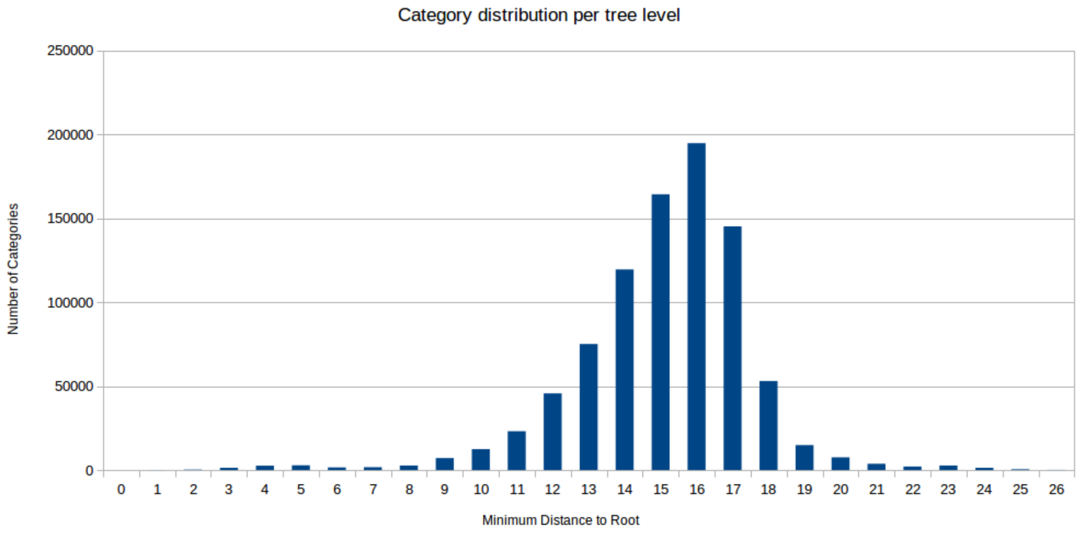


*Figure 2 Category distribution per level, total network (before chopping)*

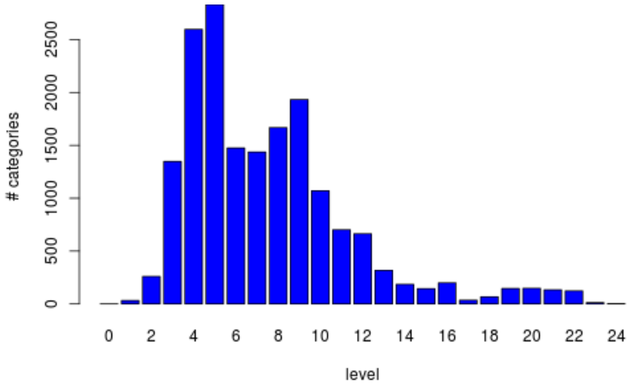


*Figure 3 Category distribution per level, FD-relevant tree (after chopping)*

The 17.5k relevant categories include **221k FD-relevant articles**. This confirms our bet that Wikipedia is the largest dataset with FD-relevant items.

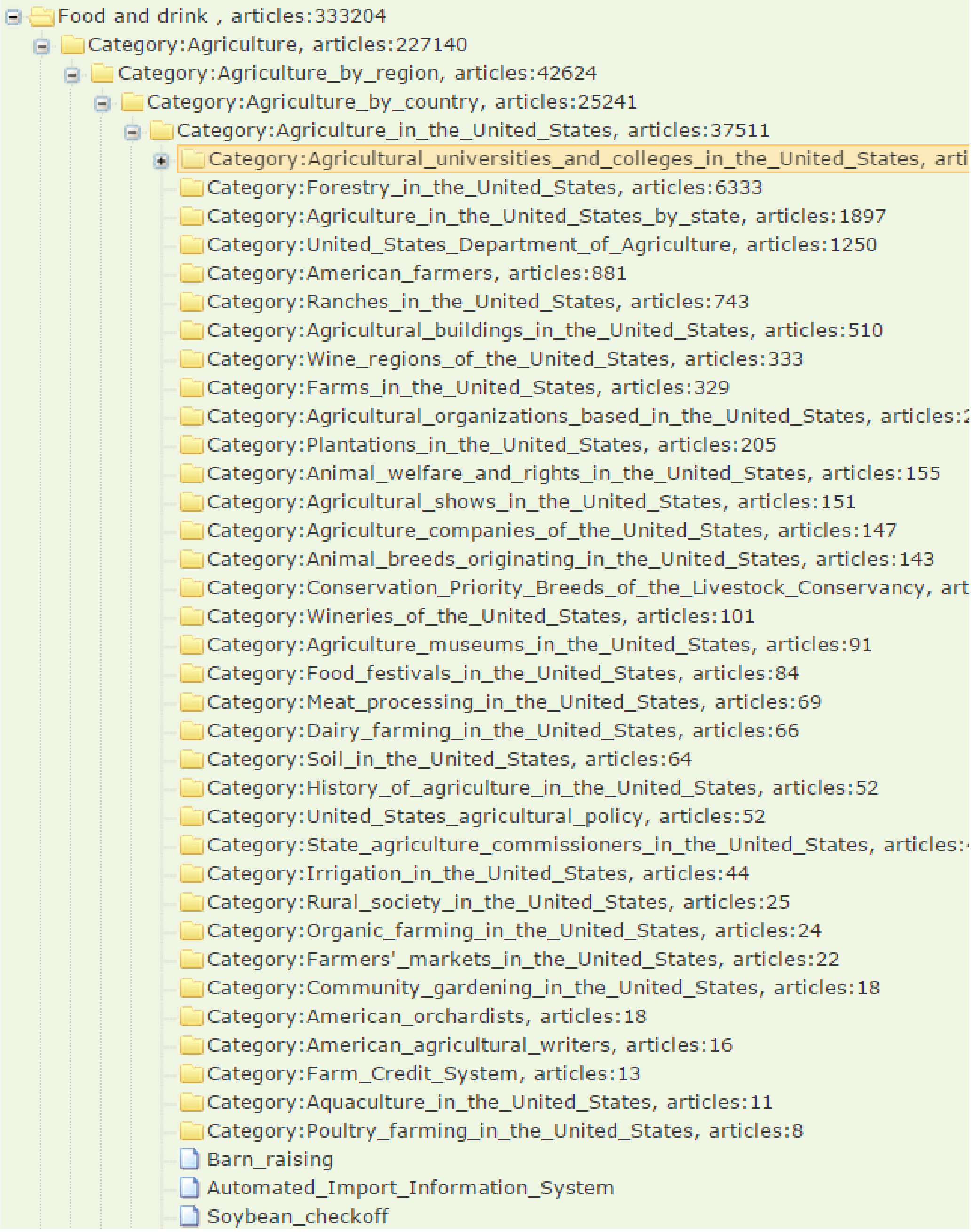
This number is subject to revision due to the following:

- The number can be reduced further (perhaps by half) by fine-grained chopping. E.g. currently the most populous categories are Agriculture and Nutrition, and there are a lot of sub-categories and articles that are barely relevant (e.g. Agricultural_universities_and_colleges_in_the_United_States, see next). We'll also examine categories that are far from the root (level>=10)
- Evidence feedback (processing articles or CHOs that are proven FD-relevant by other means) may enlarge 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 "artificially".

## 2.5 Category Tree UI

We developed a tool to view and manipulate the tree: http://efd.ontotext.com/tree/?category=Food_and_drink (User efd, password efd123). The tool is written in JavaScript and uses the software described in the previous section as backend server, communicating with it in JSON.

**Please don't use the delete** (red X) buttons: they delete whole branches of the tree without confirmation. A screen-shot is shown below.



- Food and drink , articles:333204
  - Category:Agriculture, articles:227140
    - Category:Agriculture_by_region, articles:42624
      - Category:Agriculture_by_country, articles:25241
        - Category:Agriculture_in_the_United_States, articles:37511
          - Category:Agricultural_universities_and_colleges_in_the_United_States, arti
          - Category:Forestry_in_the_United_States, articles:6333
          - Category:Agriculture_in_the_United_States_by_state, articles:1897
          - Category:United_States_Department_of_Agriculture, articles:1250
          - Category:American_farmers, articles:881
          - Category:Ranches_in_the_United_States, articles:743
          - Category:Agricultural_buildings_in_the_United_States, articles:510
          - Category:Wine_regions_of_the_United_States, articles:333
          - Category:Farms_in_the_United_States, articles:329
          - Category:Agricultural_organizations_based_in_the_United_States, articles:2
          - Category:Plantations_in_the_United_States, articles:205
          - Category:Animal_welfare_and_rights_in_the_United_States, articles:155
          - Category:Agricultural_shows_in_the_United_States, articles:151
          - Category:Agriculture_companies_of_the_United_States, articles:147
          - Category:Animal_breeds_originating_in_the_United_States, articles:143
          - Category:Conservation_Priority_Breeds_of_the_Livestock_Conservancy, art
          - Category:Wineries_of_the_United_States, articles:101
          - Category:Agriculture_museums_in_the_United_States, articles:91
          - Category:Food_festivals_in_the_United_States, articles:84
          - Category:Meat_processing_in_the_United_States, articles:69
          - Category:Dairy_farming_in_the_United_States, articles:66
          - Category:Soil_in_the_United_States, articles:64
          - Category:History_of_agriculture_in_the_United_States, articles:52
          - Category:United_States_agricultural_policy, articles:52
          - Category:State_agriculture_commissioners_in_the_United_States, articles:
          - Category:Irrigation_in_the_United_States, articles:44
          - Category:Rural_society_in_the_United_States, articles:25
          - Category:Organic_farming_in_the_United_States, articles:24
          - Category:Farmers'_markets_in_the_United_States, articles:22
          - Category:Community_gardening_in_the_United_States, articles:18
          - Category:American_orchardists, articles:18
          - Category:American_agricultural_writers, articles:16
          - Category:Farm_Credit_System, articles:13
          - Category:Aquaculture_in_the_United_States, articles:11
          - Category:Poultry_farming_in_the_United_States, articles:8
          - Barn_raising
          - Automated_Import_Information_System
          - Soybean_checkoff

***Figure 4 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.

## 2.6 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.

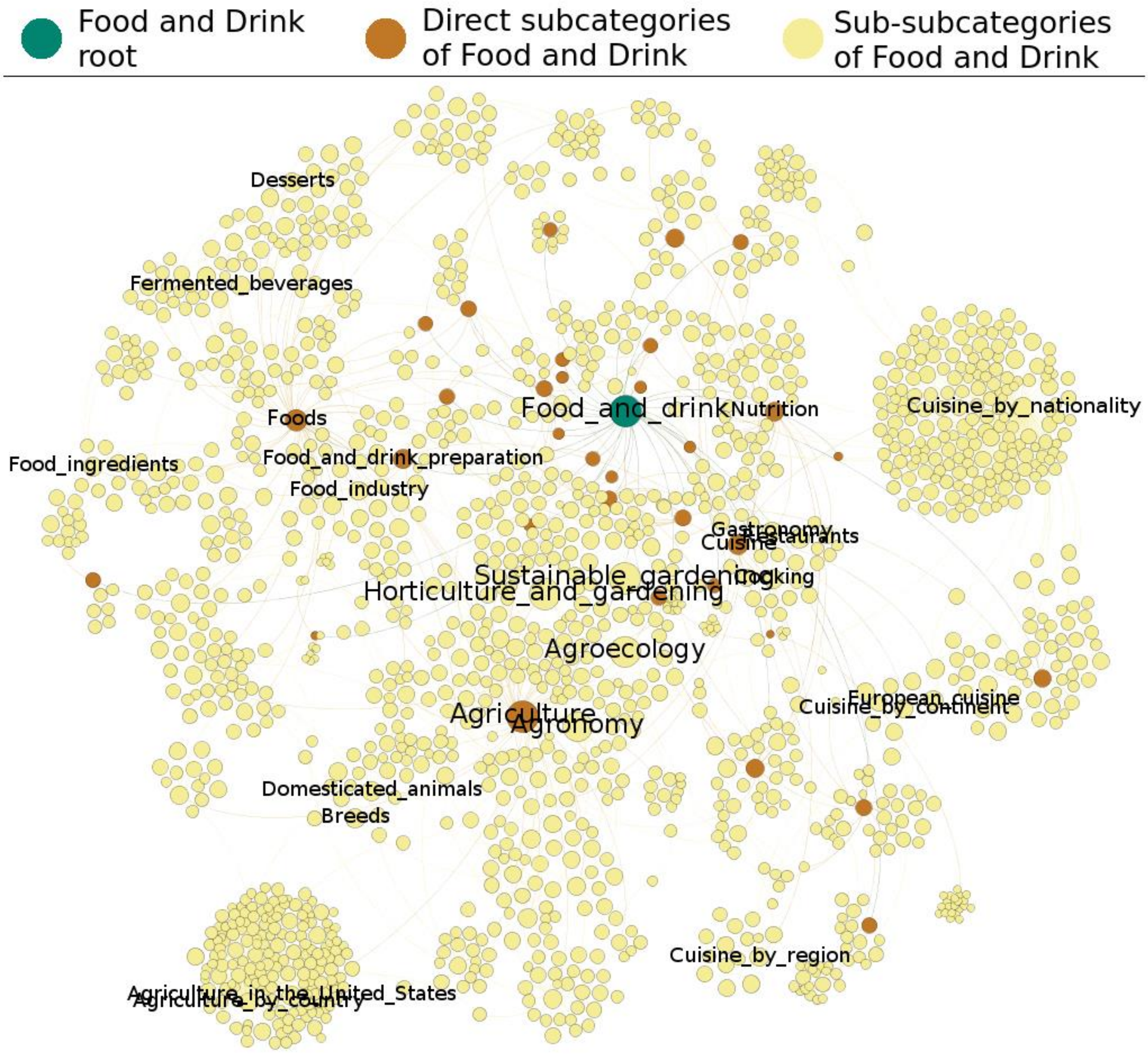


*Figure 5 Cluster graph of the FD categorization developed with Gephi*

## 2.7 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"

- Adding additional roots (related to Hunting and hunting weapons)
- Manual matching of some Horniman terms

## 2.8 Manual Curation (to Wikipedia)

Since we use semantic representations of Wikipedia (DBpedia or Wikidata) as our main source of classification, in many cases the best course of action is to add missing categories and labels to Wikipedia:

- **Adding parent categories**. 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). Since most gourds are used as (primitive) bottles, we added it.
- **Adding redirects (aliases)**. E.g. "Muller" is a vessel for making mulled beer or wine (see Horniman's site[19]). 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.
- **Adding text and redirects**. 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 Horniman[20] and Burke Museum[21]). We added it as a section to article "Fishing Tackle", and as a redirect to that specific section. But our edit was reverted[22] with the comment "not a term commonly used in fishing" and then a suggestion[23] to add to article "History of fishing" (the story is still ongoing)

This emphasizes our conclusion in [Alexiev 2015b] that it's harder to add to Wikipedia than Wikidata, and one needs to learn to work with the editorial community.

Several things are important for the sustainability of this approach:

- Recording all additions, e.g. in evaluation & manual curation sheets of the semantic enrichment process
- Developing scripts/tools for easy addition to Wikipedia (using the Wikipedia API) or Wikidata (using WMF Labs tools)
- Periodically refreshing the semantic KB (in particular DBpedia) from new Wikipedia extracts

## 2.9 Bottom-up Augmentation (Evidence Propagation)

The top-down tree formation & cleaning described in 2.4 is only one of the ways in which we elaborate the EFD classification. The other way is based on evidence (contribution) propagation from two 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 2015c sec.3.11.1] and UMBEL FD super-type [Alexiev 2015c sec.3.19]

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

[20] http://www.horniman.ac.uk/collections/browse-our-collections/object_type/term-504068

[21] http://collections.burkemuseum.org/ethnology/display.php?ID=46106

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

[23] https://en.wikipedia.org/wiki/Talk:Fishing_tackle#Cord_Attacher

We use bottom-up propagation of this evidence to elaborate a scoring scheme for FD-relevance of Wikipedia categories. We implemented two approaches; one is described in [Tagarev 2015 p.6]:

- Decaying propagation of the evidence (contribution) without regard to the FD category structure. The contribution decreases with path length going up

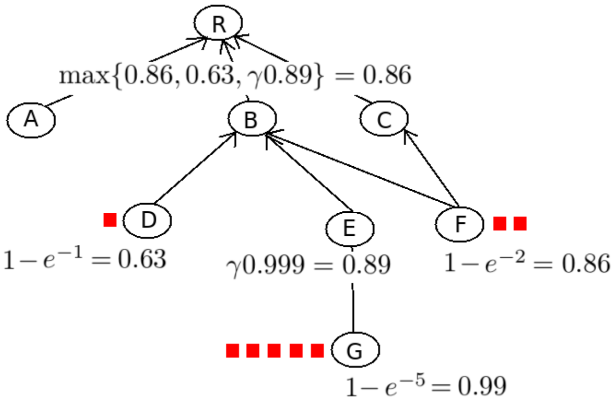


***Figure 6 Decaying propagation***

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

These approaches serve two means:

- Discover new categories that are currently outside the tree but are judged relevant (e.g. Spears as hunting weapons)
- Establish a more refined relevance scoring

## 2.10 Semantic Enrichment

Semantic enrichment (in this case relating CH objects and thesauri to the FD classification) is main purpose of the classification and the mainstay information to be processed by the semapp. The first collection we selected for semantic enrichment was Horniman (see 2.1.13) 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 won't deal with the Materials terms because a material (e.g. "wood") does not indicate relevance to FD.

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

- Adapted Ontotext's Concept Extraction Service 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.

As reported in [Tagarev 2015 sec.6.1], we achieved estimated Precision 0.91 and estimated Recall 0.7, which are raised to 100% by curation. The next figure shows which parts of the whole FD tree are activated by Horniman terms.

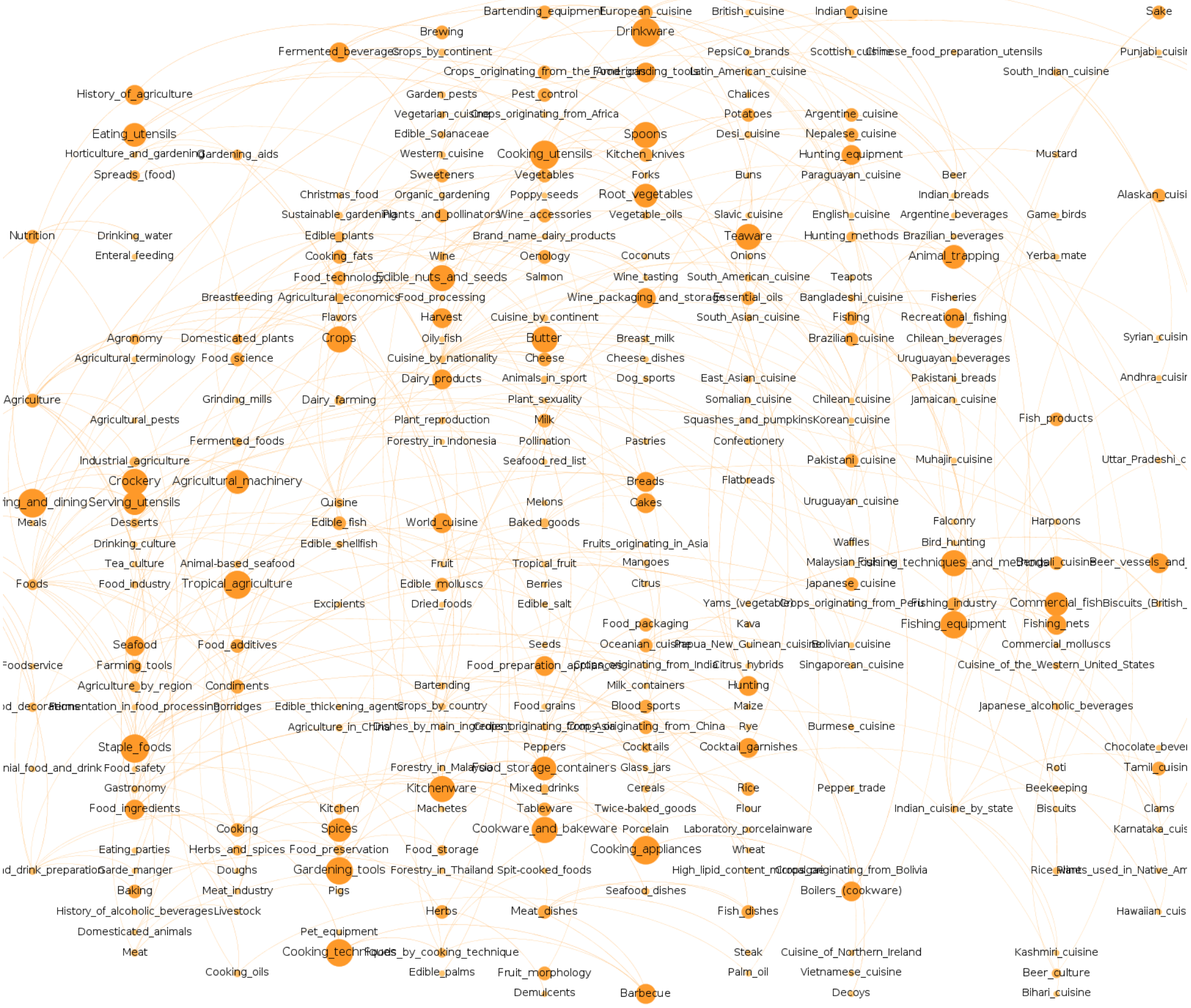


***Figure 7 FD concepts activated by bottom-up propagation from Horniman terms***

## 2.11 AAT-Wikidata Coreferencing

We want to use Getty AAT for its strong Styles and Periods facet[24], which also includes numerous Cultures and Tribes. It has 5487 terms, which can be checked with this query[25]:

```
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

[24] http://www.getty.edu/vow/AATHierarchy?find=periods&logic=AND&note=&subjectid=300015646

[25] At SPARQL endpoint http://vocab.getty.edu/sparql, maintained by ONTO

compare to Wikipedia ethnic groups, because we don't know how the coverage of the two datasets compare.

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 tool[26], as mentioned in [Alexiev 2015d sec.2.5.2]. We started a task for this at the Wikidata project Authority Control[27] (also initiated by us). AAT is already loaded to Mix-n-Match, however the precision of automatically matched concepts is only 50% [28]. So we devised an approach to salvage old AAT-WordNet 2.0 mappings through WordNet 3.0 and BabelNet:

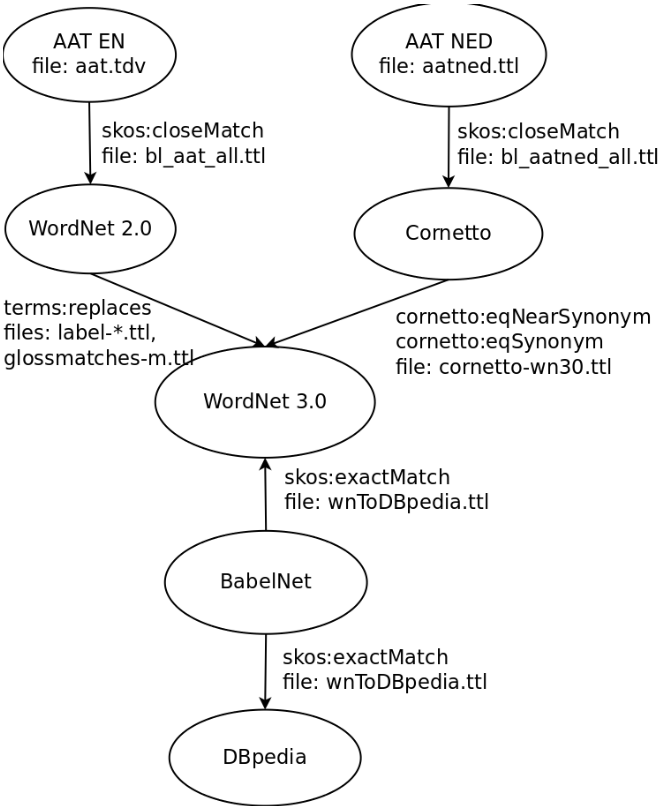


***Figure 8 Data flow for matching AAT to DBpedia through WordNet and BabelNet***

This automatic matching is completed:

***Table 2 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 very high precision (over 95%). Some mismatches include:

[26] https://tools.wmflabs.org/mix-n-match/

[27] https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control#Coreference_AAT

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

- 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"

We still need to curate the matching, put it in Wikidata and press on with the remaining manual matching through Mix-n-match.

## 2.12 Task Forces

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

- Evaluation and Enrichments[29]. 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 will be evaluated and compared against 5 other trial submissions, by projects such as LoCloud[30]. 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[31] 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).

## 2.13 Publications

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

[Alexiev2015a] 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

[Alexiev2015b] 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 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.

[Tagarev 2015] This paper describes our approach to building a domain-specific gazetteer for EFD and includes more scientific details on the approach than this document. It was submitted to a conference on semantic keyword search (Keystone) to be held in Sep 2015.

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

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

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

# 3 Project Management

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

## 3.1 Immediate Next Steps

This section describes tasks that we are addressing now.

### 3.1.1 EN Collection Enrichment

EN is the first language we are addressing, and we'll enrich all EN collections that are submitted before mid-Oct 2015.

- Finish up enrichment of the Horniman collection
  - Finish manual curation to increase precision from 91 to 100% (we've done half of 700 Object Type concepts)
  - Connect from Horniman object thesaurus to objects
  - Run general Concept Extraction for places
  - MAYBE run Concept Extraction for cultures
- Enrich the UK-Wolverhampton collection
- Enrich the UK-TopFoto collection (if we can obtain all objects)
- Enrich the HU-MKVM collection (if we can obtain all objects, and a significantly large number have EN translations in addition to HU text)

### 3.1.2 Select Next Language for Enrichment

Conduct preliminary enrichment experiments with another language. Our current candidates 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 2015c sec.3.7] are promising (lots of articles), but the counts in sec 3.8.1 are not promising (poor category structure).
- **Italian**: we have preliminary NLP experience

An important sub-task will be to leverage inter-language links, i.e. owl:sameAs connections across category networks. A description of such connections for LT can be found in 2.1.11, and considerations in [Alexiev 2015d sec.2.3.2]

### 3.1.3 Simple Semapp

We will develop a simple semapp to showcase the results of semantic enrichment. It will include the following:

- A simple object view: URL, textual descriptions, image thumbnail. Since we'll need to work with various metadata formats initially and will switch to EDM only when the relevant collections are converted to EDM, proper abstraction of the code and storage of the object data is very important
- Left frame: the category tree, limited to branches that contain CH objects. Allow browsing through the tree
- Concept search: auto-completion on articles that include classified CH objects
- MAYBE: simple geospatial and/or temporal browsing

We will first develop wireframes and clear them with the WP3 lead. See [Alexiev 2015d sec.2.13] for more detailed ideas.

### 3.1.4 Discover Europeana Objects

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 2015d sec.2.12].

Focusing on the technical side, this presents significant challenges:

- 2.4 describes that we've identified 221k articles in 17.5k categories relevant to FD. We may still cut this in half by removing Agriculture and Nutrition articles marginally relevant to FD (e.g. US Agricultural Universities)
- Each article has many titles (labels): Wikidata aliases or Wikipedia redirects. Furthermore, Wikidata provides multilingual labels directly. We have seen items/articles with over 40 labels; assume 10 labels on average.
- So this makes 1M labels that need to be queried against Europeana. It makes sense to make 100k queries, each being a disjunction (OR) of all labels of that article.
- It may be better to do this using a local FTS index through SPARQL (at http://europeana.ontotext.com/sparql), rather than the Europeana API.

We are confident that in this way we will discover many FD-related objects in Europeana, perhaps over 1M. We have still not decided how to process them,

### 3.1.5 List Management

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

## 3.2 Scope for Oct 2015

ONTO was invited at the Project Management Board (PMB) meeting on 15 April 2015 in London. At the meeting we shared our concern about lack of collected objects, and that the potential scope of work on the semapp is very large (see [Alexiev 2015d] sec. 2; also Antoine Isaac as reviewer has commented the same).

Therefore the following scope was agreed for the D3.20 deadline 31 Oct 2015:

- Handle 1 language (EN)
- Handle all collections that have EN metadata: UK-Horniman, UK-Wolverhampton, UK-TopFoto, HU-MKVM.
- Build classification based on Wikipedia and AAT, including UI for tree manipulation
- Implement FD-specific semantic extraction based on the classification
- Generic Place extraction
- Correlation from X Cuisine to place/culture X, e.g.:
  - Bulgarian cuisine → Bulgaria
  - Cajun cuisine → Acadia, Louisina
- Semantic search using autocomplete over matched concepts
- Semantic browsing through the category tree, showing actual number of matched objects (faceted search)

- Some simple demo applications, e.g.: geographic search, "lightbox" (i.e. simple gallery), timeline

## 3.3 Extended Scope

After submitting D3.20, ONTO plans to continue work on the semapp in the following directions:

- Extend semantic enrichment to other languages
- Continue elaborating the FD Classification through bottom-up augmentation
- Provide more end-user apps
- Provide semantically enriched content to other application creators, e.g. for creating Cultural Paths (such an application is being considered by the consortium).

ONTO's budget includes 40k EUR for subcontracting for a mobile application. Given that there is a separate task to create a mobile application, the PMB agreed to ask the EC to reallocate this budget for further core development by ONTO. This will get us more than 10p/m of additional effort, so we can develop more.

## 3.4 Future Applications

ONTO has been contacted by the following parties regarding the potential future application of the EFD Classification approach using Wikipedia:

- Steven Stegers (EUROCLIO) invited ONTO to the Europeana for Education task force. We participated in the task force kick-off meeting (21-22 June 2015 in Paris)
- Antoine Isaac (Europeana) enquired about applying the approach to build the Europeana Arts channel.
- Stefano Caneva (WeLand and Wikipedia, Italy/Belgium) enquired about semantic integration of Italian food resources for a cultural path information

# 4 Conclusions

This document has described all work done on the semapp (D3.20) in the first 2.5 months, and the progress. We have made good progress on the fundamental tasks, and work is proceeding apace.

We hope that this document demonstrates the specific approach we are taking towards the EFD Classification. It is still early to show end-user applications, but we hope that the document has hinted sufficiently on the usefulness of the semantic approach.

Our work provides a direct response to the concerns of project reviewers that EFD is not doing enough with Europeana content, both for contributing and for reusing.

We plan to provide a shorter progress report in end-Aug 2015, and a final report end-Oct 2015.

# 5 References

[Alexiev 2015a] 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-europeanas-semantic-strategy-glamwiki-2015

[Alexiev 2015b] 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 2015c] 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

[Alexiev 2015d] 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 modeling: Towards a Food and Drink Gazetteer. First International Keystone Conference, Coimbra, Portugal, Sep 2015 (submitted). http://vladimiralexiev.github.io/pubs/Tagarev2015-DomainSpecificGazetteer.pdf