



europæana
food and drink

Grant Agreement 621023

Europeana Food and Drink

D3.19 Semantic Demonstrator Specification

Deliverable number	3.19
Dissemination level	PU
Delivery date	March 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

The Europeana Food and Drink Semantic Demonstrator (Semantic Application or simply "EFD semapp") will use innovative semantic technologies to automate the identification, classification and exploration of Food and Drink (FD) related objects. The result will be a body of semantically-enriched metadata that can support a wider range of multi-lingual applications such as search, discovery and browsing.

It will leverage the EFD Classification to identify and classify objects, and at the same time augment the Classification using Machine Learning (ML) techniques. It will use feedback loops and Human-Computer interaction, intermixing incremental Machine Learning and Crowd-sourcing of classification judgements (e.g. positive & negative examples).

It will draw on both EFD CHOs (content & metadata) and CHOs already existing in Europeana. In this way it will build up a much wider FD content base that can be used by other EFD application partners and the future Europeana FD Channel.

The EFD Classification and enrichment pipelines will be published for open re-use (not including proprietary Ontotext ML components). We hope that it will be sustained by Europeana for the future evolution of the Europeana FD Channel.

Revision History

Rev	Date	Author	Organisation	Description
V0.1	15/03/2015	Vladimir Alexiev	ONTO	Initial Draft
V0.2	28/03/2015	Elena Lagoudi	PS	First Review; Conclusions
V0.3	30/03/2015	Hugo Manguinhas	EF	Second Review
		Antoine Isaac	EF	
		Laura Tolosi	ONTO	
		Nick Poole	CT	
V1.0	31/03/2015	Vladimir Alexiev	ONTO	Final Version

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise.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

Contents

Abstract.....	2
Revision History.....	2
Contents	3
1 Introduction	5
1.1 EFD Semapp Processes.....	5
1.2 EFD Semapp User Roles.....	5
1.3 Semantic Web Technologies	6
1.4 Semweb Characteristics	7
1.5 Linked Open Data.....	7
1.6 Role of Semweb in CH.....	9
1.7 Semweb in Europeana.....	11
1.8 Semantic Enrichment in Europeana.....	14
1.9 Role of the EFD Classification	15
1.10 News Enrichment.....	16
1.11 FD Enrichment Attempt 1.....	18
1.12 FD Enrichment Attempt 2.....	19
2 EFD Semapp	22
2.1 EFD Semapp Approach	22
2.2 Semantic Knowledge Base	23
2.2.1 Semantic Data Integration	24
2.2.2 Wikidata's Relevance to CH.....	25
2.2.3 Potential Additional Datasets	26
2.3 Category Management	26
2.3.1 Cat Comb.....	27
2.3.2 Multilingual Categories.....	27
2.3.3 Automatic Filtering	28
2.3.4 Manual Pruning.....	29
2.3.5 Fuzzy (Partial) Relevance.....	30
2.3.6 Category Scoring	30
2.3.7 Category Enrichment	31
2.4 List Management	32
2.4.1 List Identification	32
2.4.2 List Extraction	32
2.5 Thesaurus Alignment.....	33
2.5.1 Automatic Alignment.....	34
2.5.2 Manual Alignment	36
2.6 Automatic Enrichment.....	37
2.6.1 Multilingual Processing	38
2.7 Manual Curation.....	39
2.8 Thematic Classification.....	41
2.9 Semantic Search and Faceting.....	42
2.9.1 Auto-completion.....	43
2.9.2 Semantic Faceting	44

D3.19 Semantic Demonstrator Specification

2.10	Visualisation.....	45
2.10.1	Radial Tree	46
2.10.2	Tree Map	47
2.10.3	Sunburst	48
2.11	EFD Data Flow.....	49
2.12	Discovering Europeana CHOs.....	50
2.12.1	Filtering CHOs by Learning.....	50
2.12.2	Filtering CHOs by Technical Metadata	51
2.13	Sample Apps	51
2.13.1	Topical Discovery.....	52
2.13.2	Timelines	52
2.13.3	Geographic Maps.....	53
2.13.4	Similar Objects.....	53
3	Conclusion.....	54
4	References.....	55

1 Introduction

The Europeana Food and Drink Semantic Demonstrator, or Semantic Application (here simply called "EFD semapp") will enable identification (discovery), classification and exploration of Food and Drink (FD) cultural heritage objects (CHOs) using semantic technologies. It will also facilitate making creative applications by making use of in-depth semantic search, thus enabling the discovery and repurposing of Europeana thematic content in ways not possible previously.

We hope that the EFD semapp will form the basis for a future FD Channel on Europeana, a goal that Europeana has set.

1.1 EFD Semapp Processes

The EFD semapp should achieve these goals by enabling the following inter-related processes (the separation of processes into modules will be part of technical design):

- Semi-automatic classification (semantic enrichment) of thematic content (CHOs related to FD) to be contributed to Europeana by the EFD project and potentially by future projects or partnerships
- Augmentation and elaboration of the EFD Classification, in a positive feedback loop with CHO classification
- Alignment (co-referencing) of local (curatorial) classification schemes used by particular EFD content providers with established global classifications, thus potentially opening their content for global cross-collection search.
- Identification (discovery) and classification (semantic enrichment) of thematic content that already exists in Europeana, and its subsequent reuse in creative applications
- Various types of semantic search and faceting
- A few sample creative applications based on semantic search

1.2 EFD Semapp User Roles

The EFD semapp will have three main kinds of users:

- **Provider:** content providers and other project partners. Includes classification of provider collections or sub-collections, classification of already existing Europeana content.
- **Metadata Specialist:** Ontotext and other technical partners, supported by content providers. Perform semantic text analysis, thesaurus alignment, elaboration of the classification.
- **Consumer:** anonymous users or the general public. Browses the classification, explores objects, uses semantic search

The separation of the semapp into modules is not decided yet.

1.3 Semantic Web Technologies

This report presupposes that the reader has some familiarity with the semantic web (also dubbed Web 3.0) and its role in Cultural Heritage. This movement started 25 years ago, with the original designs of Tim Berners-Lee, creator of the web.

Semantic technologies have the potential to open up or improve many aspects of cultural heritage management and use. By using richer contexts, we can improve the knowledge and understanding of heritage collections. By developing semantically-rich approaches to cataloguing and description, we can take advantage of next-generation tools for search and discovery. See sec.1.6 for some more details.

Semantic technologies involve many technical standards, such as:

- RDF as data model
- XML RDF, Turtle, NTriples, JSON-LD for data representation
- RDFa, Microdata, Microformats for embedding RDF data in HTML
- RDFS for basic schema information and schema-based reasoning
- OWL and its profiles Lite, DL, EL, QL, RL, Full for increasingly complex class- and property-based reasoning
- SWRL and RIF for sophisticated rule-based reasoning
- SPARQL for querying semantic databases (repositories)
- SPARQL Graph Protocol and SPARQL Update for storing data in repositories
- Ontologies for defining data schemas

Some well-known ontologies include:

- SKOS and SKOS-XL for defining thesauri (classification schemes)
- DC and DCT for basic resource/bibliographic information
- ORE for interlinking data about the same resource
- CIDOC CRM for historic events and CH content
- EDM for transmitting and storing Europeana objects. EDM reuses DC, DCT, SKOS, FOAF, ORE and is inspired by CRM
- FOAF, ORG, BIO, SIOC, VCARD for personal and social information
- Schema.org and GoodRelations for all kinds of things described on the web: people, organizations, products, offers, locations, creative works, restaurants, recipes, books, bibliographic references, etc.
- PROV for recording provenance (the way a piece of data was generated or changed)
- BIBO, RDA, FRBR, FRBRoo for bibliographic information
- ADMS, DCAT, VOID, VANN, VOAF, VOAG for describing datasets and vocabularies (e.g. location, size, access modes, etc)

Ontology representation syntaxes include:

- RDF (e.g. Turtle)
- OWL XML
- Manchester notation

1.4 Semweb Characteristics

Some key characteristics of the semantic web include:

- Every piece of data is globally addressable through a URI. Resolvable URLs are preferred and should stay permanent
- Content negotiation is used to obtain human-readable or machine-readable information about a URL (e.g. HTML and RDF XML respectively). Or the two can be combined (e.g. using RDFa).
- Anyone is free to make statements about any resource, and/or mint URIs for that resource
- There is a way to relate 2 different URIs that refer to the same resource (e.g. http://dbpedia.org/resource/Leonardo_da_Vinci and <http://viaf.org/viaf/24604287/>)
- Semantic integration can happen "simply" by putting statements made by different authorities in the same repository
- Schemas (ontologies) are represented in the same way as data: using RDF and global URLs

This last point has important consequences:

- The traditional dichotomy between schema and data (which is present in relational databases and XML) is blurred
- RDF data is self-describing since each fact (statement) refers to the ontology it came from
- RDF data can use as little or as much schema as needed, and even different applications can use different levels of reasoning over the same data
- Ontology engineers are stimulated to reuse and recombine already developed ontologies. For example, the GVP Ontology [Alexiev2015b] reuses about 12 other ontologies

1.5 Linked Open Data

One does not need to be conversant with and deploy all the standards and technologies described in sec.1.3 to reap some benefits from semtech. An important development in the last 8 years is Linked Open Data (LOD). The main idea is to represent data as RDF, provide stable (permanent) URLs, and interlink with already published datasets to facilitate reuse and discoverability. Following Berners-Lee,¹ people often talk of 5-star Open Data²:

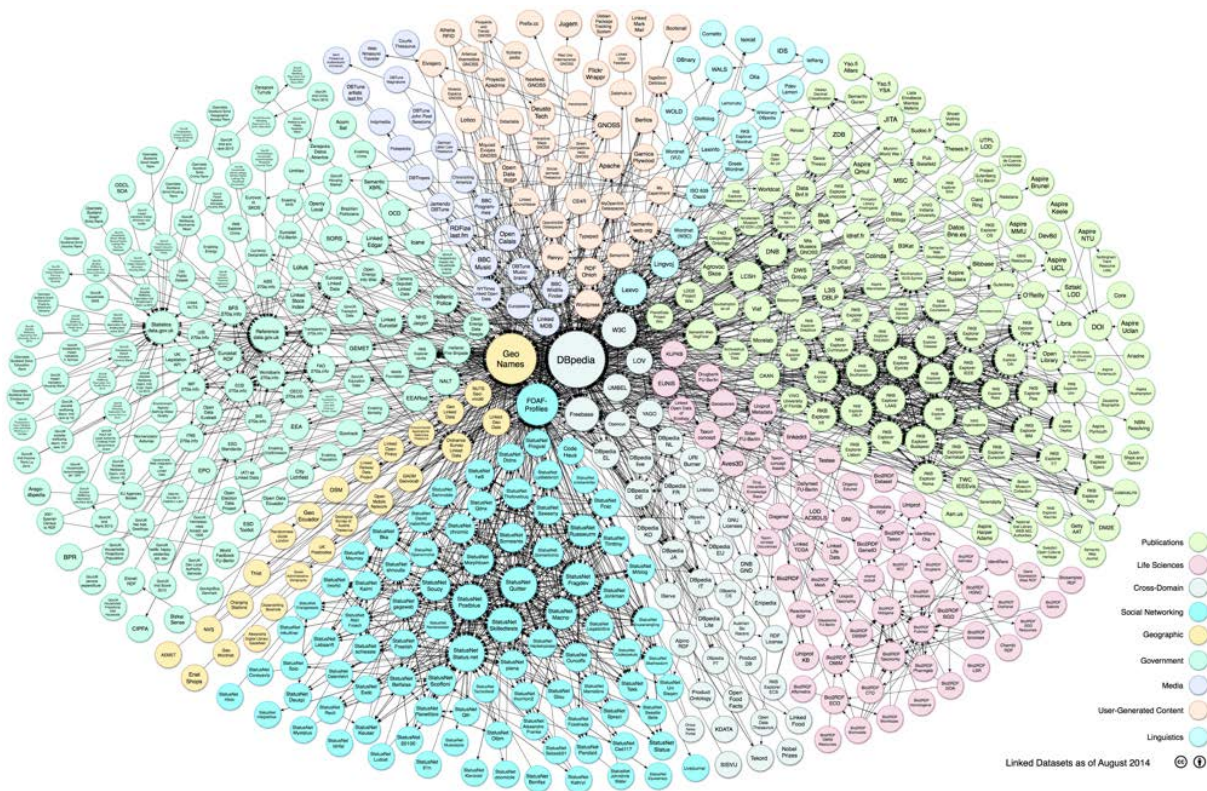
¹ <http://www.w3.org/DesignIssues/LinkedData.html>

² <http://5stardata.info/>



★	make your stuff available on the Web (whatever format) under an open license
★★	make it available as structured data (e.g., Excel instead of image scan of a table)
★★★	use non-proprietary formats (e.g., CSV instead of Excel)
★★★★	use URIs to denote things, so that people can point at your stuff
★★★★★	link your data to other data to provide context

Starting from a single dataset 2007 (DBpedia [Lehmann 2015], a semweb rendition of Wikipedia), a "LOD cloud" [Schmachtenberg 2014] has grown impressively to 570 datasets³



³ <http://lod-cloud.net/>

D3.19 Semantic Demonstrator Specification

By now semweb has become the de-facto standard for large-scale cross-industry data integration efforts in Life Sciences, e-Government and many other domains.

For example, Linguistic LOD:⁴



Nevertheless, much work is still needed for publishing and interlinking datasets as LOD. The DataHub⁵ lists over 9k datasets; of them only 909 (less than 10%) are available in RDF.

1.6 Role of Semweb in CH

Semweb plays an increasingly important role in CH due to the following characteristics:

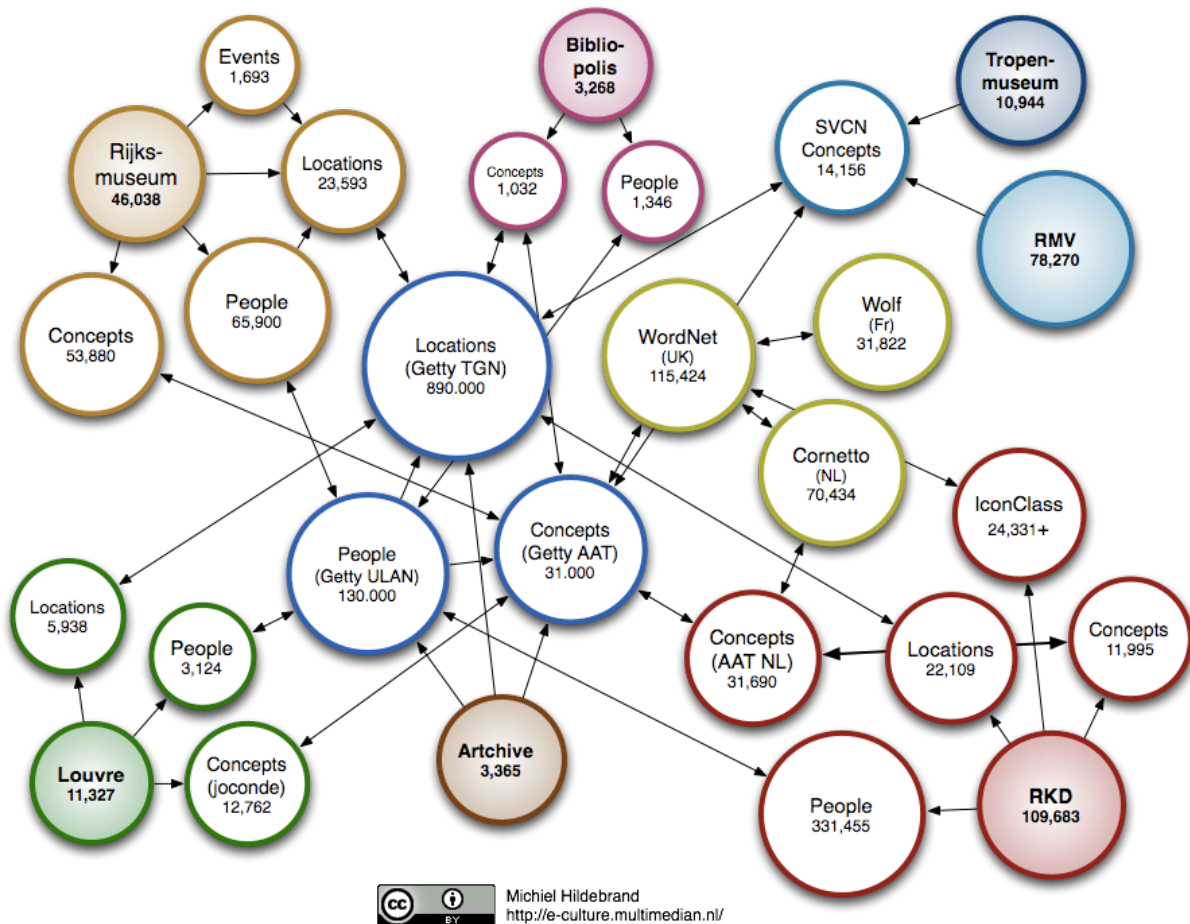
- Historic, CH and Digital Humanities data very often has complex shape and numerous exceptions. E.g. you may be inclined to say that a person has one father, but sooner than later you'll face the need to represent several opinions about several fathers, with attendant provenance and justification information. Therefore CH data does not fit well in traditional relational schemas
- CH data often outlives the lifetime of single systems. E.g. the British Museum collection data (numbering 2.5M CHOs) has gone through several migrations of IT systems.
- Even in a single institution, CH data is often held in a variety of systems, which makes it hard to query efficiently
- Since culture is globally related, the value of interlinking CH datasets is very high

⁴ <http://linguistic-lod.org/lod-cloud>

⁵ <http://datahub.io/dataset>

D3.19 Semantic Demonstrator Specification

A LOD cloud in Cultural Heritage is emerging:



This lists only thesauri. However many museums and other GLAM institutions have started publishing LOD, e.g.:

- Amsterdam Museum
- British Museum *
- Yale Center for British Art *
- Polish Digital Library *
- Europeana (see next section) *
- Cooper-Hewitt
- Smithsonian American Art Museum
- Powerhouse Museum
- The European Library

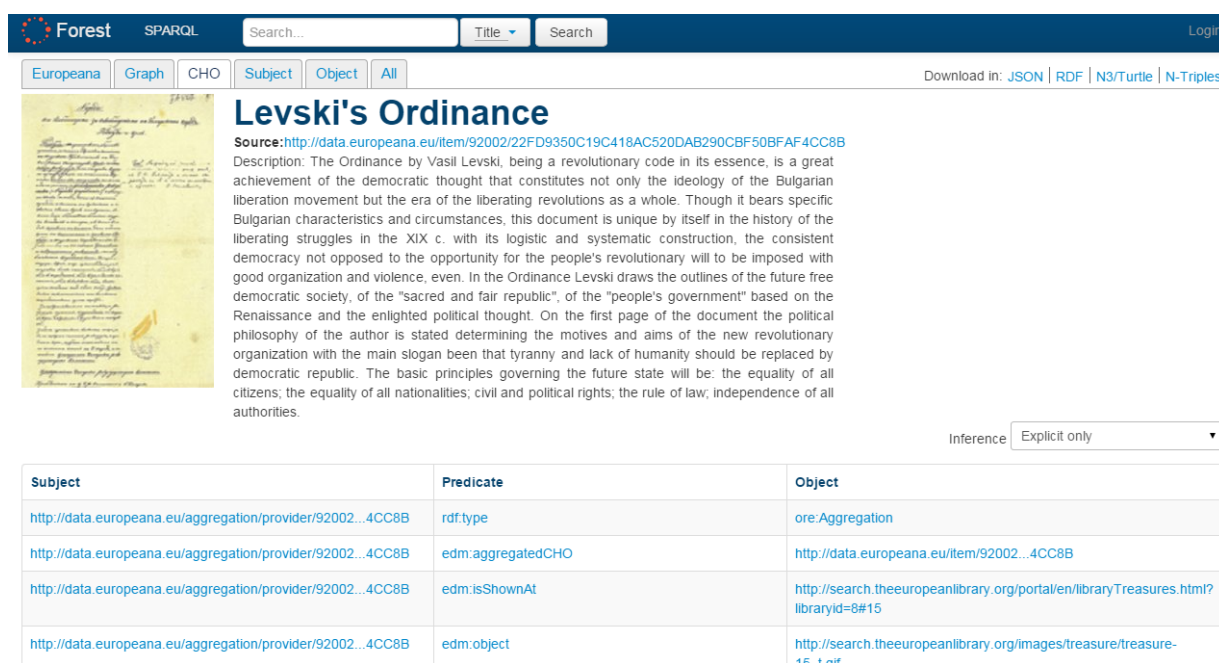
* Ontotext has helped create the LOD for these institutions

1.7 Semweb in Europeana

[Gradmann 2010] describes the importance of interlinking and semantics for Europeana. The Europeana Data Model (EDM)^{6, 7} is based on RDF, which provides a flexible data model for the various providers and allows both providers and Europeana to add links to other semantic resources. Europeana is providing LOD⁸ on an experimental basis

- The Europeana LOD pilot⁹ [Haslhofer 2011] started in Feb 2012 and provided 2.4M objects
- Ontotext provides a semantic repository with SPARQL and full-text search for 20M Europeana objects,¹⁰ last updated in Sep 2012

For example, here are the RDF triples for an important Bulgarian artifact, Levski's Ordinance:



The screenshot shows the Europeana interface for 'Levski's Ordinance'. The page includes a search bar, navigation tabs (Europeana, Graph, CHO, Subject, Object, All), and a description of the document. Below the description is an RDF table with columns for Subject, Predicate, and Object.

Subject	Predicate	Object
http://data.europeana.eu/agggregation/provider/92002...4CC8B	rdf:type	ore:Aggregation
http://data.europeana.eu/agggregation/provider/92002...4CC8B	edm:aggregatedCHO	http://data.europeana.eu/item/92002...4CC8B
http://data.europeana.eu/agggregation/provider/92002...4CC8B	edm:isShownAt	http://search.theeuropeanlibrary.org/portal/en/library/Treasures.html?libraryid=8#15
http://data.europeana.eu/agggregation/provider/92002...4CC8B	edm:object	http://search.theeuropeanlibrary.org/images/treasure/treasure-15_t.tif

Here is the RDF graph representation of the same object:¹¹

⁶ <http://labs.europeana.eu/api/linked-open-data/data-structure/>

⁷ <http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation>

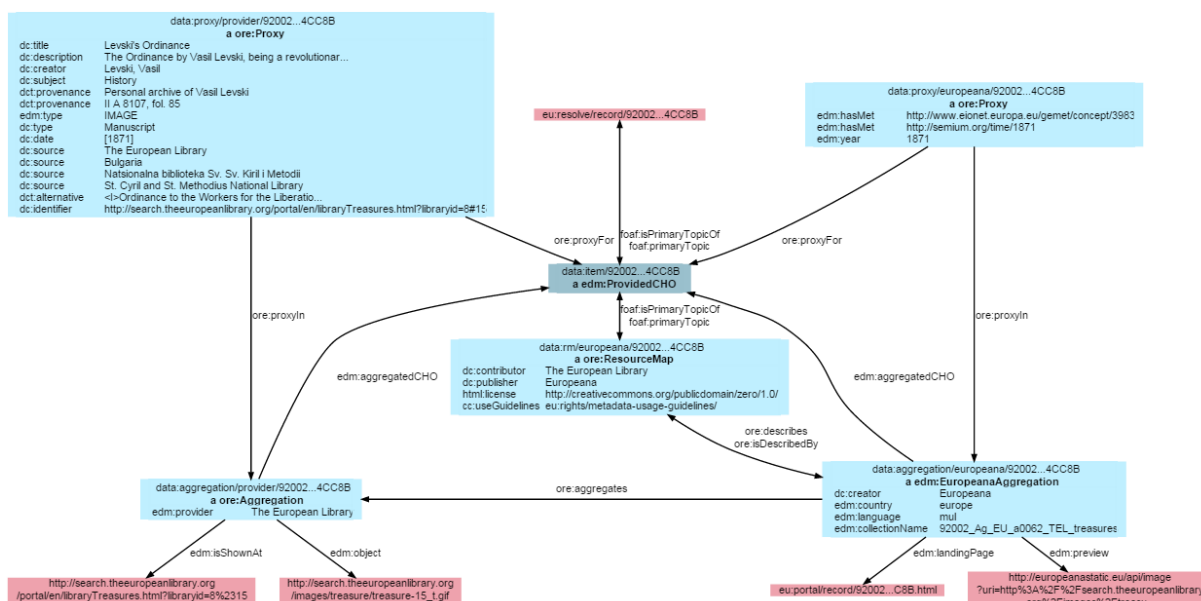
⁸ <http://labs.europeana.eu/api/linked-open-data/introduction/>

⁹ <http://datahub.io/dataset/europeana-lod-v1>

¹⁰ <http://europeana.ontotext.com/>

¹¹ <http://europeana.ontotext.com/europeana/tab?uri=http://data.europeana.eu/item/92002/22FD9350C19C418AC520DAB290CBF50BFAF4CC8B&role=Graph>

D3.19 Semantic Demonstrator Specification



A query to find 100 audio recordings:

```
SELECT ?CHO ?title ?mediaURL ?creator ?source WHERE {
  ?resource edm:type "SOUND" ; ore:proxyIn ?proxy ;
    dc:title ?title ; dc:creator ?creator ; dc:source ?source .
  ?proxy edm:isShownBy ?mediaURL .
  ?resource ore:proxyFor ?CHO}
OFFSET 600 LIMIT 100
```

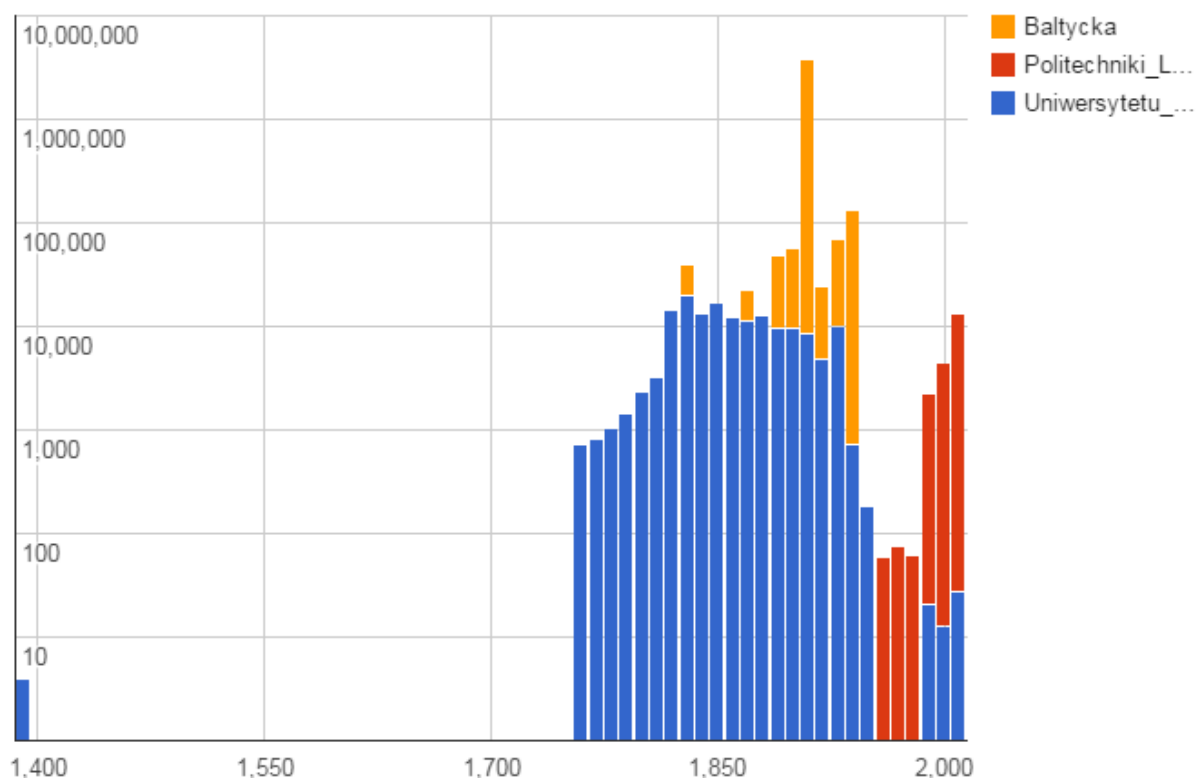
A query to chart Polish newspaper articles from 3 digital libraries by decade¹². Because different providers use different fields and values to indicate "newspaper" (in this case the string 'periodical'@en), we cannot easily fetch all papers from all countries. This finds 154k newspapers.

```
select
  ?date
  (sum(?n1) as ?Uniwerytetu_Warszawskiego)
  (sum(?n2) as ?Politechniki_Lubelskiej)
  (sum(?n3) as ?Baltycka)
{
  ?x dc:type 'periodical'@en.
  ?x ore:proxyIn/edm:dataProvider ?dataProvider.
  ?x dc:date ?date2.
  bind (xsd:integer(concat(substr(?date2,1,3),'0')) as ?date)
  bind (if(?dataProvider='e-biblioteka Uniwerytetu Warszawskiego',1,0) as ?n1)
  bind (if(?dataProvider='Biblioteka Cyfrowa Politechniki Lubelskiej',1,0) as ?n2)
  bind (if(?dataProvider='Bałtycka Biblioteka Cyfrowa',1,0) as ?n3)
} group by ?date order by ?date
```

¹² <http://jsfiddle.net/valexiev/t4aX9/>

D3.19 Semantic Demonstrator Specification

The X axis is the decade and the Y axis is the number of articles (in logarithmic scale)



- As part of the Europeana Creative project, Ontotext developed an OAI PMH server for Europeana, to allow bulk download of EDM CHO's and keep the semantic repository up to date. The test SPARQL endpoint¹³ provides 33M objects: it is still missing 6M objects, does not have full-text search, and has some other defects. Production deployment of OAI and SPARQL in Europeana Labs is expected shortly, which will add 2 more access methods to the existing Europeana API¹⁴

The following query¹⁵ charts papers provided by the Europeana Newspapers project by decade. Because a proper semantic URL from the AAT has been used (aat:300026656¹⁶ meaning "newspapers"¹⁷), we can query confidently. This finds 1.86M newspapers.

```
select ?date (count(*) as ?c) {
  ?x edm:hasType aat:300026656; dc:date ?dat.
  bind(concat(substr(?dat,1,3),'0') as ?date)
} group by ?date having (?c>1) order by ?date
```

¹³ <http://europeana-test.ontotext.com/>

¹⁴ <http://labs.europeana.eu/api/>

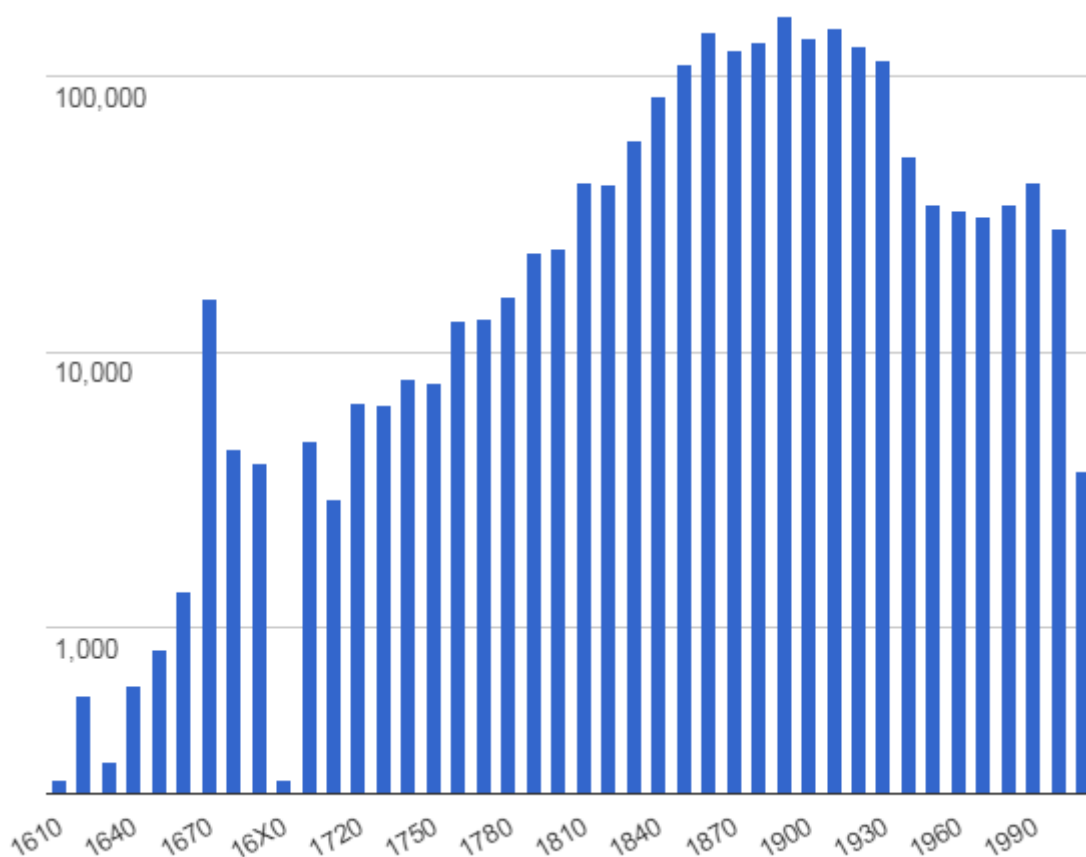
¹⁵ <http://jsfiddle.net/valexiev/ovyL9m42/>

¹⁶ <http://vocab.getty.edu/aat/300026656>

¹⁷ <http://www.getty.edu/vow/AATFullDisplay?find=&logic=AND¬e=&subjectid=300026656>

D3.19 Semantic Demonstrator Specification

The X axis is the decade and the Y axis is the number of articles (in logarithmic scale):



1.8 Semantic Enrichment in Europeana

In addition to expressing CHOs, EDM permits the use of semantic "contextual entities" such as: persons, places, concepts (e.g. subjects), events. There are some good examples of the use of semantic entities by providers, e.g.:

- Consistent type "newspapers" used by Europeana Newspapers, as shown in the previous section
- Collections enriched with AAT [Charles 2014]
- In particular, the Partage Plus (Art Nouveau) project took care to cross-check the concepts gathered by the project partners against the AAT, rather than making an isolated thesaurus. Even though AAT does not have a particular hierarchy for Art Nouveau, 97% of the required concepts were already present in AAT.

Europeana itself does semantic enrichment using datasets such as GeoNames, GEMET (environmental concepts), DBpedia, and Semium (a straightforward conversion of years to named periods, e.g. "15th century"). See [Manguinhas 2014] for details.

However, many Europeana objects still do not include many semantic references, but mostly text (that is one reason why the Europeana **portal** does not currently use

D3.19 Semantic Demonstrator Specification

semantic technologies for its core operations, but SOLR indexing). Consider that the first 15M objects were collected in ESE (the predecessor of EDM) that did not allow semantic URLs; then retro-converted to EDM. And even after the introduction of EDM, few content providers use semantic URLs in their CHOs.

Semantic enrichment (or semantic annotation) is the process of extracting some meaning from free text. Usually that is limited to recognizing entities: concepts, persons, organizations, places. Sometimes more advanced techniques can extract relational info: positions (e.g. "person X is CEO of company Y"), relations (e.g. "painting X is created by Y"), quotes and attributions (e.g. "X said that Y"), events (e.g. "X sold Y to Z for the amount of T"), etc.

Semantic enrichment can enable an important strategic goal of Europeana in the next couple of years: improve data quality, interlinking and discoverability. Europeana has formed two Task Forces on this subject (Ontotext participates in both):

- The Task Force on [Evaluation and Enrichments](#) (2015, ongoing) has the following goals:
 - collect enrichment processes, workflows and efforts in the Europeana network including correcting of enrichments through crowdsourcing, assess what they have in common and how they differ,
 - enhance the interoperability of enrichment services/modules, for example by identifying problems which hinder interoperability,
 - determine a set of methods (incl. metrics) to evaluate the impact of enrichments,
 - help participating projects enhance the enrichment services they are creating, by collecting appropriate vocabularies for enrichment and enrichment rules, and
 - pinpoint the most promising ways to include human feedback in the workflow.
- The Task Force on [Multilingual and Semantic Enrichment Strategy](#) (2014) was motivated by Europeana's goal to ensure that enrichments unfold their whole potential and act as facilitators of access. The main work of the task force consists of:
 - Developing a multilingual and semantic enrichment strategy.
 - Suitable collections are identified as use cases and their metadata fields are analyzed to find matching controlled vocabularies which would be good candidates for enrichment
 - Assembling a final report [Stiller 2014]

1.9 Role of the EFD Classification

The EFD Classification is a multi-dimensional scheme for discovering and classifying CHOs related to FD. It lies at the core of the EFD semapp, both providing the basis for classification and being augmented through its application to CHOs.

To support the broadest possible range of re-use models, we are building upon existing datasets and terminologies to develop the Classification. The EFD

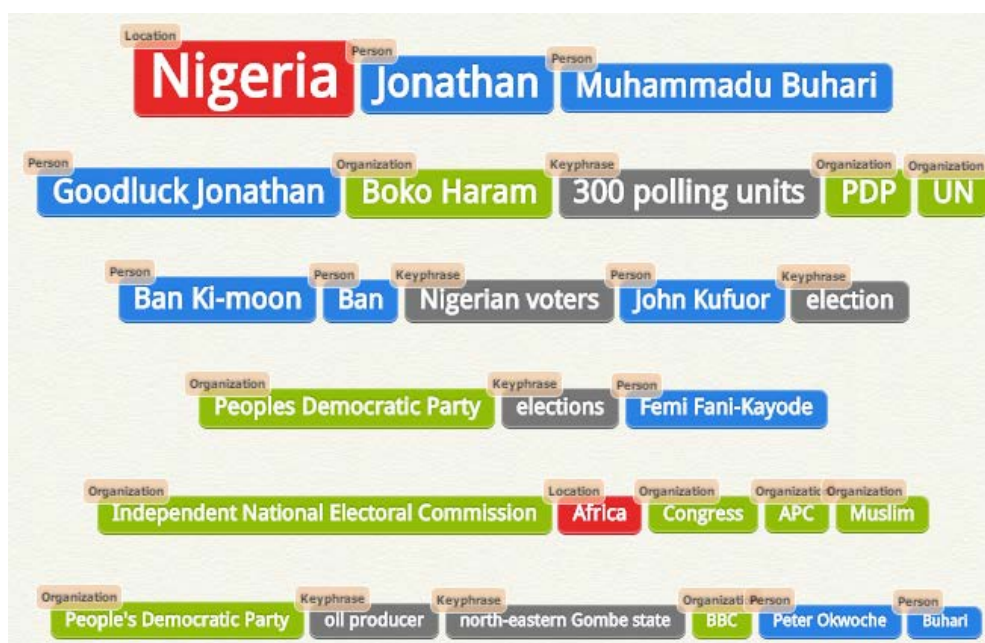
D3.19 Semantic Demonstrator Specification

Classification report [Alexiev2015a] researches some 20 potential datasets and describes the selected scheme in a lot of detail (91 pages). A few of the considered datasets are:

- The Getty vocabularies AAT (concepts), TGN (places) and ULAN (agents).
- GeoNames (places)
- VIAF (agents)
- Wikipedia in its 2 semantic renditions (Wikidata and DBpedia): concepts, places, agents

1.10 News Enrichment

Ontotext creates various enrichment services for a number of commercial clients, in media, publishing, life sciences, etc. Let's try an example: take a BBC news piece¹⁸ on Nigeria's elections. Paste the URL in Ontotext's demo tagging service¹⁹. The results are:



It has recognized a number of entities that point to semantic entities in DBpedia: persons (blue), orgs (green), locations (red). Also, it recognized "Key phrases", i.e. free text that's important in the text. Very importantly, coreferences are discovered and resolved:

- "Jonathan" is resolved to Goodluck Jonathan, even though "Jonathan" is an unusual last name. (Come think of it, Goodluck is an unusual first name as well). Resolved to http://dbpedia.org/resource/Goodluck_Jonathan

¹⁸ <http://www.bbc.com/news/world-africa-32103919>

¹⁹ <http://tag.ontotext.com/>

D3.19 Semantic Demonstrator Specification

- "Ban" (from Mr Ban) is resolved to Ban Ki-moon, the Secretary-General of the United Nations. (The surname in Korean is written first.) Resolved to http://dbpedia.org/resource/Ban_Ki-moon

If you click View Source, you can see the JSON data returned by the tagging service. E.g. for the first occurrence of "Jonathan", there is a plethora of information, including:

- matched instance (http://dbpedia.org/resource/Goodluck_Jonathan)
- class (ptop:Person)
- subclasses (dbo:OfficeHolder and fb:government.politician)
- flags like isTrusted (global instance, rather than local to the document), matchedWithLonger (coreferenced to a longer occurrence in the same document)
- probability and ambiguity assessment (relevanceScore, confidence, overallScore, ambiguityRank, ambiguityRankWithinClass). The tagging service can return several candidates per occurrence, and these fields can help to sort them in the best way.

```
{
  "name": "Jonathan",
  "startOffset": 1857,
  "endOffset": 1865,
  "type": "Person",
  "features": {
    "overallScore": 0.9471591570720387,
    "relevanceScore": 0.934757505773672,
    "end": 1865,
    "start": 1857,
    "matches": [ 3495, 3488, 3500, 3489, 3481, 3478 ],
    "matchedWithLonger": true,
    "ambiguityRankWithinClass": 1,
    "ambiguityRank": 2,
    "confidence": 0.9595608083704054,
    "isTrusted": "true",
    "fnMention": "true",
    "fullName": "false",
    "uiRepresentation": [ "Person" ],
    "subclass": [
      "http://dbpedia.org/ontology/OfficeHolder",
      "http://rdf.freebase.com/ns/government.politician" ],
    "tokenFeature": "string",
    "id": 1322230,
    "string": "Jonathan",
    "class": "http://www.ontotext.com/proton/protontop#Person",
    "inst": "http://dbpedia.org/resource/Goodluck_Jonathan"
  }
}
```

The precision (accuracy of matching) is excellent. The only false hit is:

- "Muslim northerner" is mis-recognized as [dbpedia:Muslim_Brotherhood](http://dbpedia.org/resource/Muslim_Brotherhood)

The recall (number of matches) is fairly good. Some omissions include:

D3.19 Semantic Demonstrator Specification

- "north-eastern Gombe state" is found as a Keyphrase and a local instance is made for it: http://data.ontotext.com/publishing/topic/North-eastern_Gombe_state (isGenerated=true). Instead, it should have been matched to http://dbpedia.org/resource/Gombe_State, while the prefix "north-eastern" is a qualifier, not part of the name. Maybe the fact that "state" is written in lowercase is to blame
- "Independent National Electoral Commission" instantiated as a local entity of type Organization. The reason is that [dbpedia:Independent National Electoral - Commission](http://dbpedia.org/resource/Independent_National_Electoral_Commission) does not have type dbo:Organization (even though it has types yago:Committee108324514 yago:Group100031264 yago:Organization108008335 yago:ElectionCommission108325124 yago:ElectionCommissions) that mark it clearly enough as an organization. Ontotext has become active in the DBpedia community to help with corrections to the DBpedia ontology and mappings, in order to improve the quality and completeness of DBpedia data.

1.11 FD Enrichment Attempt 1

Now let's try the same tagging service with the Coral Pestle from Horniman²⁰ [Alexiev 2015a] sec 5.1:

```
pestles (food processing & storage).  
Collection: Anthropology.  
Place: Oceania - Oceanic Islands - Tubuai Islands - Austral Islands.  
Materials: coralline limestone.  
Part of these projects: Collections People Stories.  
About this project: Oceania Collection Review.  
Description: Conical coral food pounder for mashing taro roots to make poi.  
Commentary: Food Pounder Cut From Coral, Penu, Austral Islands, Central Polynesia.  
Penu food pounders of this horned, concavely conical form are found with several variations in style throughout Central and Eastern Polynesia. The design is highly ergonomic - adapted over centuries to fit the hand perfectly and allow exactly the right kind of mechanical action to be applied to the food in the wooden bowl.  
The selection of a heavy slab of coral from the fringing reef created a working surface of regular pits and ridges that mashed the cooked root vegetables quickly and easily. In general, such pounders were used to make poi, a pudding of mashed taro, yams or breadfruit, moistened and sweetened with coconut milk, and steamed on hot rocks in an earth oven.  
Coralline limestone. Early 19th Century. Purchased at Stevens' Auction Rooms in 1910.
```

²⁰ <http://www.horniman.ac.uk/object/10.278>



The result is not very good:

- Foods and implements (e.g. pestle) are not recognized.
- Only one place is recognized as a global entity: Oceania <http://sws.geonames.org/6255151/>. The rest are made out as local resources, even though Polynesia, Tubuai Islands, Austral Islands are present in global sources. This means that geographic search (by place hierarchy or coordinates) won't work
- Stevens is recognized as an organization, and correctly is not matched to any global source

Note: the date "1910" will be recognized as part of standard Europeana enrichment. This enables the Year facet and constructing timelines.

1.12 FD Enrichment Attempt 2

We have implemented a test version of the tagging service, only available internally.²¹ It integrates more data sources (including Wikidata), implements better type classification (through voting between several typing systems and linguistic features) and connects to an internal integrated knowledge base, not only to DBpedia and GeoNames. It finds the following references:

²¹ <http://192.168.130.143:18080/tag/>

D3.19 Semantic Demonstrator Specification



- It has more alternative labels for the same concept. E.g. it resolves "Coralline limestone" to Limestone
- It finds more places in global sources, e.g. Polynesia is resolved to an entity ([wikidata:Q35942](https://www.wikidata.org/wiki/Q35942), same as [dbpedia:Polynesia](https://en.wikipedia.org/wiki/Polynesia))

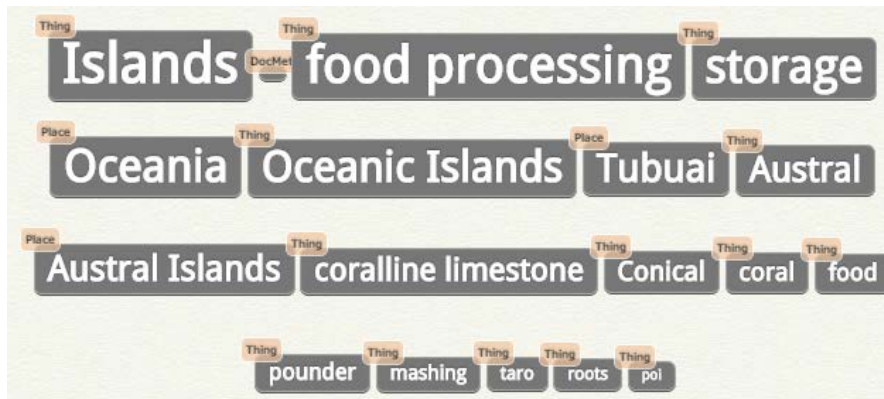
But a lot remains to be done. Most importantly, this extractor is overly eager, e.g. it extracts irrelevant words such as "Description" (one of four rhetorical modes), "Part" (a part or voice is a strand or melody of music played by an individual instrument), "selection" (biological selection). It should adapt better to context

If we limit the text to the key metadata fields:

```
pestles (food processing & storage).  
Oceania - Oceanic Islands - Tubuai Islands - Austral Islands.  
coralline limestone.  
Conical coral food pounder for mashing taro roots to make poi.
```


D3.19 Semantic Demonstrator Specification

The results are better:



But there are still problems (this is work in progress):

- "Austral" is resolved to Argentine austral, a currency
- "Pounder" is resolved to a surname
- "Conical" is resolved to cone, three-dimensional geometric shape
- "Mashing" is resolved to a process in brewing and distilling
- The object type "pestle" is not resolved

2 EFD Semapp

This section provides a specification for the different modules (functionalities) of the semapp. This will form the basis of follow-on work:

- Technical design, including: enrichment experiments, final selection of datasets, linguistic experiments, acquisition of metadata samples from all content providers
- User Interface (UI) design, including wireframes
- Test scenarios and user-acceptance test cases
- Loading and interlinking of all selected datasets to Ontotext GraphDB
- Semapp development, including text analysis pipelines and specialized user interfaces

Important note: the functionality below represents a very ambitious scope indeed. Since the various functions are not designed yet, we don't have reliable estimates either (neither effort, nor duration). Therefore it is possible that the final EFD semapp won't include all the functions described: this is a "programme maximum". Depending on the help we get from project partners and their desired prioritization, we will select the most appropriate functions to implement.

2.1 EFD Semapp Approach

The general approach of the semapp involves:

- incremental machine learning (fine modification of probabilistic models based on changing knowledge base scores or user judgements)
- crowd-sourcing (collaboration on manual tasks),
- human-computer interaction (leveraging user judgements)
- feedback loops and incremental refinement (the more the classification is used, the better it becomes)

In some more detail:

- Whenever a concept (article) is used to tag a CHO, we mark it as appropriate to the domain.
- We also trace upward toward the root (category "Food and drink") and mark all categories along the way as appropriate
- The user can cut out branches from the category hierarchy as appropriate. This is done in a crowd-sourced fashion.

For the case of object discovery from Europeana:

- We use the labels of confirmed concepts (articles) to find existing CHOs that mention the same label
- We show these candidate CHOs to the user and ask for feedback, i.e. to point out some positive and some negative examples

D3.19 Semantic Demonstrator Specification

- We learn from the negative examples, e.g. if many of the rejected CHO have the word "fragment" or "shard" (see [Alexiev 2015a] sec. 2.12.1), we put it on a blacklist.

The feedback loops work like this:

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

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

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

This enables a bootstrapping mechanism and a positive feedback loop: discover entities → human feedback → update model → discover entities

2.2 Semantic Knowledge Base

The first important task is to build a multilingual semantic Knowledge Base (KB) of the FD domain. This KB will form the basis of linguistic information (for building dynamic gazetteers), the semantic URLs to link to, and the additional information that can be used by semantic applications (hierarchy, geographic coordinates, etc)

The EFD Classification report outlines a number of relevant datasets, and we'll start with the following. DBpedia/Wikidata provides universal knowledge including concepts, agents, places, events, etc.

- DBpedia in the 11 EFD languages, at least to provide category assignments (article<category and category<category)
- Wikidata, which provides integrated access to all languages and all labels
- Getty AAT, which provides 40k concepts (in particular a Cultures/Periods/Styles hierarchy with 6k concepts)

We might also need to turn to these datasets that provide more entities than DBpedia/Wikidata in their respective domains:

- GeoNames: a lot more places (9.5M vs 850k)
- VIAF: a lot more agents (31M vs 3M)

The building of the KB will be based on the following steps:

- Acquisition of metadata samples from all content providers, covering the variety of objects they will provide. We have initiated this process in March 2015.

D3.19 Semantic Demonstrator Specification

- Enrichment experiments and evaluation to test the coverage on various kinds of objects
- Selection of datasets
- Interlinking the datasets (semantic data integration)
- Loading the selected datasets to Ontotext GraphDB

Note: the same semantic repository will have the EDM CHO data (likely in different Named Graphs), in order to enable faster queries.

2.2.1 Semantic Data Integration

Semantic data integration is one of the tasks mentioned in the previous section. There are two aspects to it:

- Interlinking the selected datasets by using predicates such as `skos:exactMatch` and `owl:sameAs`
- Loading the datasets into the same semantic repository, so the links can take effect.

This is required in order to have a unified space of concepts. It would be very bad to treat `aat:300024668`²² "knife" as a different concept from `enwiki:Knife`, because that would present the user with two disconnected concepts that are in fact the same.

Only 1% of AAT is coreferenced to Wikipedia, which is a shame since AAT is such key thesaurus in CH. We looked at ways to coreference more of AAT (we expect that 25-30k of AAT's 40k concepts will be present in Wikipedia), and will use 2 approaches:

- We found existing basic coreferences AAT→WordNet created by Anna Tordai as part of Europeana Connect, covering 15k concepts (38% of AAT). BabelNet on the other hand includes WordNet-Wikipedia correspondences. We need to transform (bring forward) the old Europeana Connect coreferences into the latest versions of AAT and WordNet. Then convince the BabelNet authors to provide us a more liberal access key, so we can fetch the relevant WordNet→Wikipedia links that we need. This is described at Wikidata's WikiProject Authority control²³

As a fallback strategy, we could use a 2-step prioritized semantic processing pipeline:

- The first step matches against Wikidata
- Only if there is no match, the second step matches against AAT

However such approach is less flexible and harder to add more datasets to.

²² <http://www.getty.edu/vow/AATFullDisplay?find=&logic=AND¬e=&subjectid=300024668>

²³ https://www.wikidata.org/wiki/Wikidata:WikiProject_Authority_control#Coreference_AAT_through_BabelNet

2.2.2 Wikidata's Relevance to CH

Wikidata is becoming increasingly relevant to CH LOD and Europeana in particular:

- Wikidata will be a major topic of the upcoming GLAM-WIKI 2015²⁴ conference. In particular, the introduction "Wikidata for GLAMs"²⁵ will cover Wikidata projects of special importance for GLAMs. Amongst them are Authority Control (initiated by Ontotext), Visual Arts, and Sum of All Paintings.
- Our presentation proposal with Europeana "Wikidata, a target for Europeana's semantic strategy"²⁶ was accepted for the conference
- Starting in Apr 2015,²⁷ VIAF will transition from English Wikipedia coreferencing to Wikidata coreferencing. As a result it will pick up a lot more multilingual labels, 700k persons and 300k organizations that don't occur in English Wikipedia. In [Alexiev 2015c] sec 3.2 we argued that VIAF and Wikidata have few names in common: we are glad that this development will quickly bridge the gap.
- Google has reconfirmed its commitment to migrate FreeBase data to Wikidata,²⁸ something that was doubted by many.

Now consider the missing Object Type "pestle" in sec. 1.12. In [Alexiev 2015a] sec 5.1 we remarked that English Wikipedia (enwiki) doesn't have an entry for "pestle" but only for the pair [enwiki:Mortar and pestle](#). However, Wikipedia has not one but two entries²⁹ for "pestle" (probably coming from dewiki):

- pestle ([wd:Q907209](#), [dewiki:Pistill](#)): usually less heavy and massive than Q1316130, the curvature suits to that of the working surface (left on the figure below)
- pestle ([wd:Q1316130](#), [dewiki:Stößel](#)): usually heavier and more massive than Q907209, the curvature does not necessarily suit to that of the working surface (right on the figure below)



²⁴ https://nl.wikimedia.org/wiki/GLAM-WIKI_2015

²⁵ https://nl.wikimedia.org/wiki/GLAM-WIKI_2015/Programme/Introductions/Wikidata

²⁶ [https://nl.wikimedia.org/wiki/GLAM-WIKI_2015/Proposals/Wikidata,](https://nl.wikimedia.org/wiki/GLAM-WIKI_2015/Proposals/Wikidata,_a_target_for_Europeana%E2%80%99s_semantic_strategy%3F)

[a_target_for_Europeana%E2%80%99s_semantic_strategy%3F](https://nl.wikimedia.org/wiki/GLAM-WIKI_2015/Proposals/Wikidata,_a_target_for_Europeana%E2%80%99s_semantic_strategy%3F)

²⁷ <http://outgoing.typepad.com/outgoing/2015/03/moving-to-wikidata.html>

²⁸ <https://github.com/google/primarysources>

²⁹ <https://www.wikidata.org/w/index.php?title=Special:Search&search=pestle>

D3.19 Semantic Demonstrator Specification

Unlike creating a Wikipedia article, adding a Wikidata item is easy, and adding an extra label to an existing item is trivial (that's what's happened above). So the latest dump of Wikipedia should be able to provide this object type.

Also importantly, [wd:Q45778](https://www.wikidata.org/wiki/Q45778) Mortar and [wd:Q907209](https://www.wikidata.org/wiki/Q907209) Pestle are now coreferenced to AAT, and Pestle is declared "part" of Mortar (which corresponds to [enwiki:Mortar and pestle](https://en.wikipedia.org/wiki/Mortar_and_pestle)).

As you see, Wikidata development is very dynamic.

2.2.3 Potential Additional Datasets

In addition to the datasets described in EFD Classification, we might also want to consider the following datasets that we discovered during the last month:

- OpenFoodFacts³⁰
 - 40k food products sold in 115 countries and territories.
 - The majority of the data is not interesting, since it is about nutritional values, containers, labels, territories, brands
 - But may have a useful number of food names and varieties. Also has a good classification, though it's not truly multilingual (dominated by English and French branches)
- Kasabi Food & Foodista³¹. Kasabi has closed down, but dumps of these datasets are available at the Internet Archive³²
 - Foodista: 32k recipes
 - Food includes 66k recipes, 22k persons, a couple of thousand foods classified into seafood, spices, etc, etc.
 - The recipes are not interesting for us, but the food names are very interesting. The persons could be interesting if they
 - Food uses the LinkedRecipes ontology that is also dead, but fortunately archived³³ in Feb 2014. We could use it to represent

2.3 Category Management

As explained in [Alexiev 2015a] sec 3.8, Wikipedia Categories are a key element of the EFD classification since they serve to find and organize the concepts (Articles) that we use to classify CHOs.

- Organizing: we restructure the category network into a hierarchy (see the subsections). The providers or consumers can browse the hierarchy, visualize it, and explore objects by the hierarchy. For example "Category:Seafood" to find all articles related to seafood, and thereon all objects classified with those articles.

³⁰ <http://openfoodfacts.org/>

³¹ <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/dataset/foodista>

³² <https://archive.org/details/kasabi>

³³ <https://web.archive.org/web/20140207204602/http://linkedrecipes.org/schema>

D3.19 Semantic Demonstrator Specification

- Finding: we assume that articles in the FD category tree are relevant for enrichment of FD content. This helps with disambiguation: if the same word has two Wikipedia articles, we pick the one that's in the FD hierarchy.

Category assignments (article<category and category<category) are not in Wikidata, so we need to load them from DBpedia. We'll also consider adding this data to Wikidata, so it's easier to access from Wikidata.

2.3.1 Cat Comb

As explained in 3.8.4, categories have various problems:

- they don't form a thesaurus hierarchy, but a general network with loops
- there are some inappropriate categories under the root (category:Food_and_drink)
- there is no assurance on the meaning of a subcategory (doesn't mean it's logically a subset of the parent category)

To deal with the first problem, we need to develop a "category combing" algorithm that eliminates loops by starting from the root and removing edges that disagree with the direction going downward (defined according to the shortest distance from root)

- edges going "backwards" are easy to detect by breadth-first traversal and marking the category "level" (shortest distance to root)
- edges going "sideways" (loops within the same level) are harder to detect since the levels can get pretty big. Once detected, all edges participating in the loop are eliminated.

2.3.2 Multilingual Categories

[Alexiev 2015a] sec 3.8.1 lists the number of categories per EFD language. But the rest of the discussion talks about the English categories only. In fact we should merge the category networks of different languages for the following reasons:

- To leverage inter-language overlap of categories. We can assume the overlap is the same as for articles (each category is described in 2.1 languages). Merging will create a richer hierarchy than any single language
- To improve categorization for national Wikipedias with poor categorization. E.g. Italian, Dutch and Greek articles have only 1.4...1.8 categories on average, whereas English have 4 and the average across all languages is 4.44. Dutch categorization further has very low specificity: each category is applied to 29 articles, whereas the average is 11.9. We can leverage "richer" Wikipedias to improve the situation for "poorer" Wikipedias

The merging should leverage inter-language links at two levels. E.g. look at the category page [enwiki:Category:Maize](#), and trace the linking between **en** and **de**:

D3.19 Semantic Demonstrator Specification

- Category(en)-Category(de): the category (being a page) has inter-language links to other categories. On DBpedia this is represented as:

```
db:Category:Maize owl:sameAs dedb:Kategorie:Mais
```

- Category(en)-Article(en)-Article(de)-Category(de): many categories have an article that represents them (called "Topical"). E.g. [enwiki:Category:Maize](#) says "*The main article for this category is [enwiki:Maize](#)*". The article itself has an inter-language link to [dewiki:Mais](#), which in turn is the main article of [dewiki:Kategorie:Mais](#). (In this case we didn't gain an extra correspondence, but in some cases we will). On DBpedia this is represented as:

```
db:Category:Maize skos:subject db:Maize.  
db:Maize owl:sameAs dedb:Mais.  
dedb:Kategorie:Mais skos:subject dedb:Mais.
```

With several provisos:

- It's not loaded at the [dbpedia.org](#) site, but is in the Topical Concepts download³⁴
- `skos:subject` is a wrong property, should be `foaf:focus`³⁵

Merged hierarchies present some complications, but the gain is worth the difficulties:

- We may have to deal with multiple roots. This is only a technical complication
- It may happen that in the merged tree, two categories A,C in the same language X are separated by another category B that doesn't have language X. This may present difficulties to a user of language X who knows none of the languages of B. But these difficulties are surmountable
 - We could use Google Translate to provide an imperfect translation of B
 - The user could decide to expand below B even without knowing what it means

We could try to make a category browsing UI that omits B, but that's not so easy. B might have a child D in language X, and where should we show D in the tree?

2.3.3 Automatic Filtering

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

- "* templates", e.g. Cuisine templates, Drink templates, Food and drink templates
- "* portals", e.g. Food and drink portals
- "* Stubs", indicating that an article is too small and should be expanded. Note: we are not eliminating the article, only this category. The article should (and often does) have other, topical, categories.

³⁴ http://data.dws.informatik.uni-mannheim.de/dbpedia/2014/en/topical_concepts_en.ttl.bz2

³⁵ <https://github.com/dbpedia/extraction-framework/issues/301>

2.3.4 Manual Pruning

We will implement a crowd-sourcing tool (facilitated by appropriate Visualizations) that will allow Ontotext and content providers to prune the category tree when inappropriate categories are found. [Alexiev 2015a] sec 3.8.4 lists some examples of "spillage" or the inclusion of irrelevant categories under a relevant category. For example, given the disastrous chain:

- Food_and_drink
- Food_politics
- Water_and_politics *
- Water_and_the_environment
- Water_management
- Water_treatment
- Euthenics **
- Personal_life
- Leisure
- Sports
- Sports_by_type
- Team_sports
- Football
- <1000s of teams>

A user may decide to break this chain at one of two points

- *: while Food_politics has many relevant sub-categories (e.g. Kosher, Halal), Water and Water_management may be considered not relevant
- **: this point was broken on 12 June 2014 by editing Category:Euthenics with comment "Removed Category:Water treatment - Euthenics is not water treatment"

Important considerations:

- Should we allow any registered user to prune, or only some, or seek consensus (technically called Inter-annotator Agreement³⁶)? Since the editors community is small and for the sake of simplicity, we'll allow any registered user to prune. (We are more worried that content providers won't participate in this task.)
- Should we use commercial crowd-sourcing platforms like Amazon Mechanical Turk? Perhaps some project funds can be used for this purpose, but we are not sure the annotators using these platforms would make correct judgements.
- It's way to easy to misjudge a category as irrelevant when in fact it's partially relevant (see next section). Pruning is in essence black-listing (a boolean technique), which does not play well with fuzzy relevance or scoring (a numeric technique). Perhaps initially pruning should be conservative (only remove clearly

³⁶ https://en.wikipedia.org/wiki/Inter-rater_reliability

D3.19 Semantic Demonstrator Specification

irrelevant branches). As the ML model gets more elaborated, pruning can be strengthened. We need to figure out this point during technical design.

- What should happen if articles under pruned categories were already used for classification? Should the scoring of their parents (see sec. 2.3.6) be decreased?

2.3.5 Fuzzy (Partial) Relevance

Some categories are partially relevant to FD. E.g. Food_and_drink has child Animal_products. Half of the children of Animal_products are relevant to FD, e.g.:

- Animal-based_seafood(+)
- Dairy_products(+)
- Eggs_(food)(+)
- Fish_products(+)
- Meat(+)

Some are not relevant to FD:

- Animal_dyes(-)
- Animal_waste_products(-)
- Bird_products(-)
- Coral_islands(-)
- Coral_reefs(-)

Some appear not to be relevant:

- Animal_hair_products(*)
- Bone_products(*)
- Hides(*)

But they may be relevant to hunting, which is relevant to the topic. In fact Horniman's Coral Pestle (sec 1.12) makes it possible that even the Coral categories may be relevant.

Another example are Non-Human Food/Eating, since Foods_and_drink includes animal feeding. E.g.:

- Eating_behaviors(*): partially relevant
- Diets(+): relevant
- Eating_disorders(+): relevant
- Carnivory(-): not relevant
- Detritivores(-): not relevant

2.3.6 Category Scoring

One of the key requirements of the semapp is to improve the FD relevance of categories, i.e. augment the EFD classification incrementally as it is being used. We

D3.19 Semantic Demonstrator Specification

should keep the following numbers for each category: they are useful for visualization, auto-completion and disambiguation during enrichment.

- Level: length of shortest path to the root(s)
- Child categories
- Child articles
- Descendant categories
- Descendant articles
- Suggested matches: number of objects that automatic enrichment classified with a descendant article
- Confirmed matches: number of objects that manual curation classified with a descendant article (or confirmed the auto-classification with such article)
- Relevance Score: some aggregate of the above numbers, and the score of descendants (recursively)

These numbers should support the key idea of the EFD classification:

- When an article is used for classification, its ancestor categories **towards the relevant root** get a higher score (the article may well have other categories that are not relevant).
- Thus applying the Classification to objects augments it at the same time (it gets better with use)

The optimal formula for computing the Score (i.e. distributing the benefit of an article's application to an object; upwards into the ascendants of that article) is not decided yet.

- We have experience with Spreading Activation algorithms
- The idea should be similar to PageRank; GraphDB automatically can compute an RDF adaptation of it called RDFRank

2.3.7 Category Enrichment

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

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

Such lateral links are quite important since they enable discovering the cultural essence of FD CHOs.

This is similar to semantic enrichment and can use similar NLP techniques that we use to enrich CHOs. Such links often already exist as category paths, e.g. (here "~" means "topical article" and ">" means "super-category"):

D3.19 Semantic Demonstrator Specification

- db:Christmas ~ dbcat:Christmas > dbcat:Christmas_traditions > dbcat:Christmas_meals_and_feasts > dbcat:Christmas_food
- db:Bulgaria ~ dbcat:Bulgaria > dbcat:Bulgarian_society > dbcat:Bulgarian_culture > dbcat:Bulgarian_cuisine

However, these paths are quite long and uncertain. They should be made more explicit by using relatively simple NLP (at least in English), based on explicit rules and regular expressions, e.g.:

- For cat "X Cuisine": try to find a Place or Culture/Style corresponding to X (e.g. "Bulgarian" → "Bulgaria" and make a relation with type "cuisine of". Some cuisines are not related to particular Places but to Cultures/Styles, e.g. aat: 300198715³⁷ "Creole"
- For cat "X food": try to find entity X and make a relation with type "food used at"

2.4 List Management

As explained in 3.9 and 3.9.1, there are numerous FD-related lists that provide high-quality lists of food articles on a specific sub-topic, e.g.

- [enwiki:List of Christmas dishes](#)
- [enwiki:List of culinary fruits](#)
- [enwiki:List of culinary herbs and spices](#)

They can be used profitably, in a way very similar to categories

2.4.1 List Identification

The first task is to identify lists related to FD. We'll use the following approaches:

- Based on simple keywords, such as: food, meat, vegetable, grain, drink, beer, alcohol, wine, coffee, culinary, cuisine
- Using FD categories. Lists are pages, so they also have categories. [Alexiev 2015a] sec 3.9.1 provides a list of Categories of relevant Lists. We need to be very certain that the categories selected for this approach are relevant, because picking an irrelevant bunch of lists will have strongly negative effect on precision

2.4.2 List Extraction

The DBpedia info for [db:List of Christmas dishes](#) has some useful info: description, categories, aliases. The corresponding Wikidata item [wd:Q1770135](#) properly classifies the entry as "instance of: Wikimedia list article", but doesn't have much other info. Unfortunately neither provides the actual list of articles.

³⁷ <http://www.getty.edu/vow/AATFullDisplay?find=&logic=AND¬e=&subjectid=300198715>

D3.19 Semantic Demonstrator Specification

So we need to extract the articles from the list. We'll implement a simple extractor, handling only the "bullet" syntax for lists. If you examine a few lists (e.g. [enwiki:List of Christmas dishes](#)), you'll see:

- Heading structure (in this case breakdown by country) that we'll ignore
- Somewhat complex structure of each bullet. We'll extract only the first link from the bullet.
 - E.g. for Australia's bullet "Christmas damper", we want to extract only Damper_(food), but none of wreath (try eating that), butter (it's only a garnish), nor Australian_bush (that's a place of sorts)
 - This means that e.g. for Argentina's bullet "Cider", we'll fail to extract Sparkling_wine. Sorry about that.
- Replace the list node with its topical category. If no such, create a new similarly named category. Ensure we don't use the list for object classification.
 - E.g. [db:List of Christmas dishes](#) has a category [dbc:Christmas food](#) (and an alias [db:Christmas food](#) confirming the meaning is the same), which means we don't need to create a duplicate category. If [dbc:Christmas food](#) happens to have [db:List of Christmas dishes](#) as its topic, we won't create a new category.
 - However, [db:List of culinary fruits](#) has only the categories [dbc:Fruit](#), [dbc:Lists of foods](#), [dbc:Lists of plants](#) and none of them is topical. The description makes it clear that Culinary Fruits are different from (Botanical) Fruits: "many edible plant parts that are true fruits botanically speaking, are not considered culinary fruits. They are classified as vegetables in the culinary sense, (for example: the tomato, cucumber, zucchini)". So we need to create a new category **mycat:Culinary_fruits** to replace [db:List of culinary fruits](#)

2.5 Thesaurus Alignment

In the case when a content provider already has an established thesaurus that is applied to all its objects (e.g. the Horniman and IAPH), it makes more sense to align (cross-reference) this thesaurus to the global dataset, rather than trying to do it object by object.

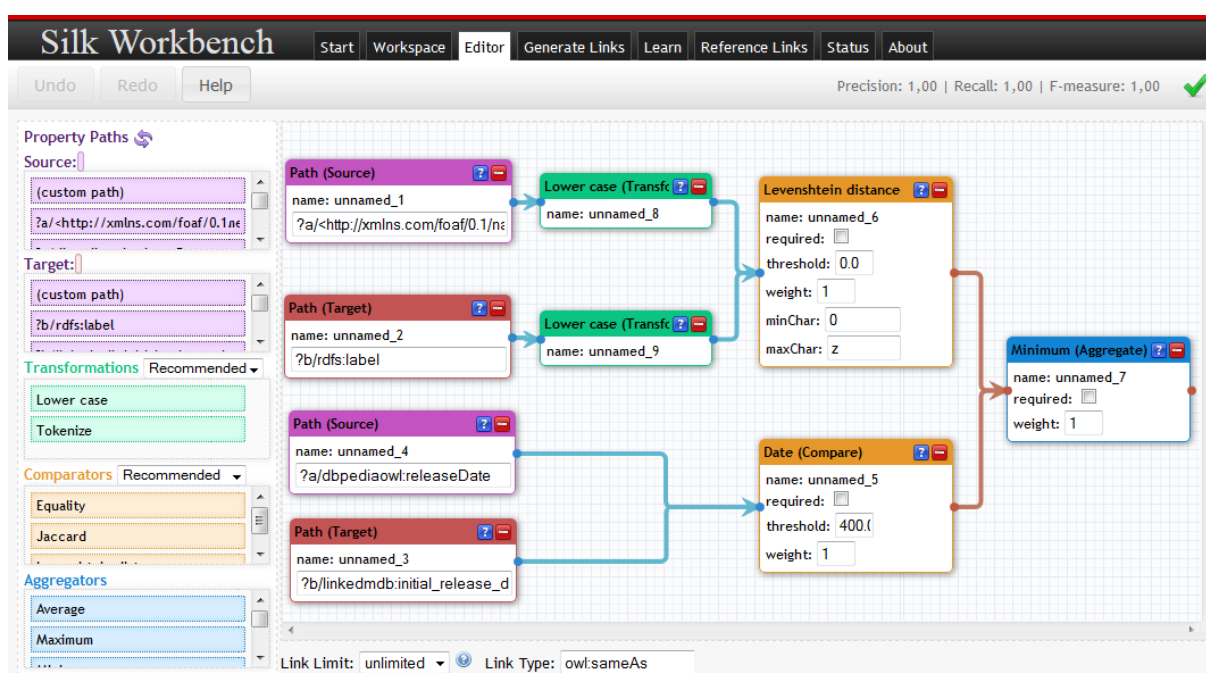
Note: this may apply to part of the metadata, e.g. Object Type; yet other parts of the metadata still merit semantic enrichment. For example, Horniman's Coral Pestle (sec 1.12) has the Type field "pestles (food processing & storage)" assigned from their Object Type thesaurus, but the following extra concepts can be extracted even from the short description:

- "coralline limestone" (material)
- "mashing" (operation)
- "taro roots", "poi" (subject foods)

2.5.1 Automatic Alignment

There are some established tools for thesaurus alignment that are part of the LOD2 Stack³⁸ (see bubble Interlinking/Fusing)

- SILK:^{39, 40} Link Discovery Framework: can be used to generate RDF links between LOD datasets. It enables data flows where one can use more strict or more relaxed similarity measures (parameters), includes a curation (manual approval/correction) tool, and even learns parameter settings using a genetic algorithm approach. The latest version includes Free Text Preprocessor.⁴¹ Defining an alignment data flow:



- LIMES:⁴² implements time-efficient approaches for large-scale link discovery based on the characteristics of metric spaces. Incorporates a number of algorithms and also has a data flow editor

³⁸ <http://stack.lod2.eu/blog/>

³⁹ <http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/>

⁴⁰ <https://www.assembla.com/spaces/silk/wiki>

⁴¹ https://www.assembla.com/spaces/silk/wiki/Silk_Free_Text_Preprocessor

⁴² <http://aksw.org/Projects/LIMES.html>

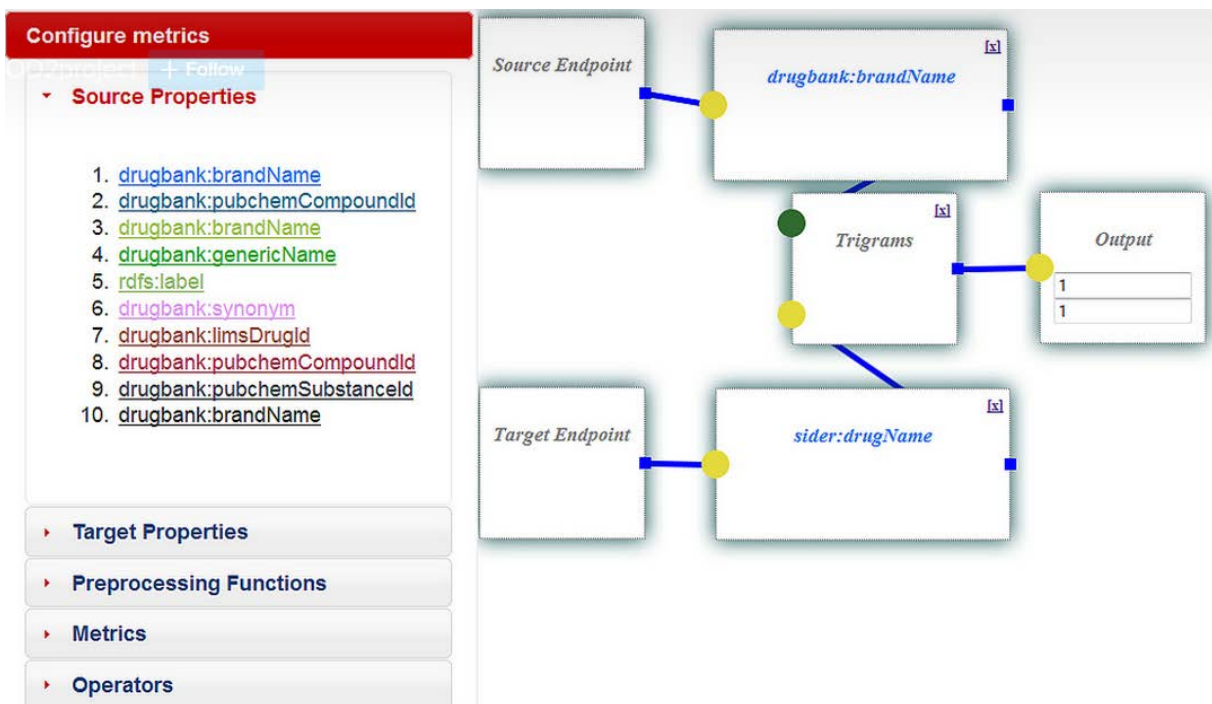
D3.19 Semantic Demonstrator Specification

Examples (toggle)
 Drugbank
 Vacations
 Duplicate Cities



Download
 Manual

Source:	Target:
Endpoint: <input type="text" value="http://dbpedia.org/sparql"/>	Endpoint: <input type="text" value="http://dbpedia.org/sparql"/>
Graph: <input type="text" value="-1"/>	Graph: <input type="text" value="-1"/>
Var: <input type="text" value="?x"/>	Var: <input type="text" value="?y"/>
Pagesize: <input type="text" value="1000"/>	Pagesize: <input type="text" value="1000"/>
Restriction: <input type="text" value="?x rdf:type dbpedia-o:City"/>	Restriction: <input type="text" value="?y rdf:type dbpedia-o:City"/>
Property: <input type="text" value="dbpedia-o:populationTotal (+)"/> <input type="text" value="rdfs:label"/>	Property: <input type="text" value="dbpedia-o:populationTotal (+)"/> <input type="text" value="rdfs:label"/>
Metric: <input))"="" type="text" value="AND(euclidean(x.dbpedia-o:populationTotal,y.dbp:"/>	
Output: <input type="text" value="N3"/>	
Execution: <input type="text" value="Linear"/>	
Acceptance:	Review:
Threshold: <input type="text" value="1"/>	Threshold: <input type="text" value="0.9"/>
Relation: <input type="text" value="owl:sameAs"/>	Relation: <input type="text" value="owl:sameAs"/>
Detected prefixes:	
rdf: <input type="text" value="http://www.w3.org/1999/02/22-rdf-syntax-ns#"/>	(-)
rdfs: <input type="text" value="http://www.w3.org/2000/01/rdf-schema#"/>	(-)
owl: <input type="text" value="http://www.w3.org/2002/07/owl#"/>	(-)
dc-terms: <input type="text" value="http://purl.org/dc/terms"/>	(-)
dbpedia-o: <input type="text" value="http://dbpedia.org/ontology/"/>	(-)
dbpedia-p: <input type="text" value="http://dbpedia.org/property/"/>	(-)



On the other hand, the problem of thesaurus alignment is similar to, but simpler than, semantic enrichment. We're trying to match a thesaurus **entry** to a dataset, which is matching text that's fully delineated, i.e. the entry's label includes only the text to match (plus an optional qualifier in parentheses). Therefore, we may just implement our own matcher.

The tricky part with any of these tools is how to use the concept's context (parent, siblings) for disambiguation. One cannot use the hierarchy directly (no two thesaurus

D3.19 Semantic Demonstrator Specification







hierarchies are structured the same), but can use the context as hints about the general area of interest.

2.5.2 Manual Alignment

Just like with semantic enrichment, automatic matches are suggestions that can be wrong. So a manual crowd-sourced tool for checking and editing matches would be useful. Ontotext can deploy such a tool and help with required data conversions

But the content providers should perform the manual adjustment, since we don't know the local languages, nor thesaurus specifics (especially if the thesaurus does not provide definitions or scope notes). If the content provider does not have the time to do this, his thesaurus will remain isolated from the global dataset that we are setting and won't be searchable with the other CHOs.

A very nice and user-friendly tool is **Mix-n-Match** that implements Wikidata co-referencing. [Alexiev 2015c] sec.4.3 describes it and provides references. One can "manage by exception", e.g. look only at unmatched concepts. Here's a screen shot of ULAN alignment:

Hannes Meyer	Swiss architect, theorist, and designer, 1889-1954	Matched by multichill
Hannes Meyer 	Swiss-german architect and photographer (1889–1954) ♂; Swiss architect and second director of the Bauhaus	Remove match
Francesco Morelli	Italian painter, active ca. 1581-1584	Not matched
Search Wikidata Search en.wikipedia Google-search Wikipedias Google-search Wikidata		Set Q No WD N/A
Alfred Mansfeld	Israeli architect, 1912-2004	Matched by multichill
Alfred Mansfeld 	Architect (1912–2004); Israel Prize and Rechter Prize ♂; Israelischer Architekt	Remove match
Master of the Martyrdom of the Ten Thousand	German engraver, active 15th century	Matched by Vladimir Alexiev
Albrecht Dürer 	Holy Roman Empire painter (1471–1528); child of Albrecht Dürer the Elder; spouse of Agnes Dürer ♂; German painter, printmaker, mathematician, and theorist	Remove match
Giovanni Battista Merano	Italian painter and printmaker, 1632-1698	Matched by multichill
Giovanni Battista Merano 	Painter (*1632) ♂	Remove match
Jan van Orley	Flemish painter and printmaker, 1665-1735	Matched by multichill
Jan van Orley 	Dutch painter (1665–1735) ♂	Remove match
Aleksandr Orlovsky	Russian painter, printmaker, and draftsman, 1777-1832	Matched by Magnus Manske
Aleksander Orłowski 	Polish painter (1777–1832) ♂; polnischer Maler	Remove match

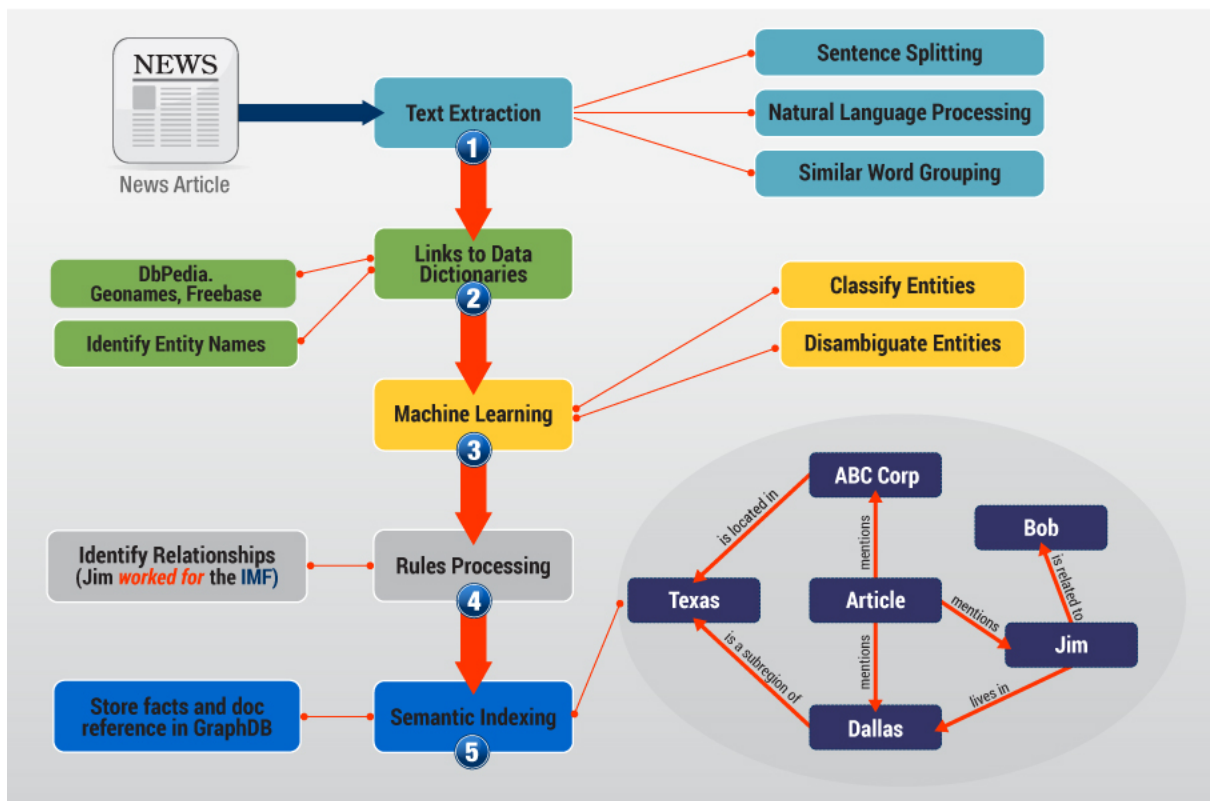
We initiated matching the British Museum person-institution thesaurus (see the same reference). The BM thesauri are currently not co-referenced to anything. The BM is keen on cross-linking, so we believe that this effort will provide high value to the BM.

The only downside is that this tool is not under our control, and its creator (Magnus Manske) is rather busy with Wikidata tooling. But he has been responsive.

Two other options are the SILK Workbench (see previous section) and xTree [digiCULT 2013].

2.6 Automatic Enrichment

Here is a diagram of a typical Ontotext semantic enrichment pipeline/architecture.



Of course, the pipeline needs to be adapted significantly for our purposes.

- Use article names and aliases (or Wikipedia redirect pages) as gazetteer. Ontotext has developed gazetteer components for extremely large-scale lookup, and that can be synchronized with changes in the RDF knowledge base.
- Limit to FD-related categories. Initially this selection will be uncertain and include wrong categories (low precision), but it will improve with usage through human feedback
- Use NLP techniques for the major languages (we have NLP experience with en, nl, it, bg), and simpler gazetteer-based techniques for the other EFD languages (see 2.6.1 for details)
- Enrich object (CHO) free text. For some fields, limit to (or augment) only some kinds of enrichments (e.g. dc:creator→persons/orgs, dct:spatial→places)
- Dealing with ambiguity:
 - Use article & category relevance Scores (2.3.6) to pick more relevant articles
 - Match primary labels with higher priority
 - Use context words and other semantic features. This is especially important for Places because place names are widely ambiguous, e.g.:
 - Use parent place names for disambiguation. E.g. if "Mexico" is in the CHO then prefer Guadalajara, Mexico to Guadalajara, Spain.
 - Prefer places with bigger population (assuming that they hold higher cultural interest as well)

2.6.1 Multilingual Processing

The multilingual requirements of EFD are very demanding. EFD covers content in 11 languages, and additional Europeana content spans over more than 20 languages. Ideally, it should be possible to query for a concept expressed in one language, and find CHOs indexed with another language (semantic or conceptual search). But NLP processing has many specific aspects for each specific language, so the quality of processing depends on the depth of NLP techniques and resources available for that language.

- English (**): taken as a base, both because there are the largest number of NLP resources available for it, English Wikipedia is the largest, and Ontotext has the most experience with it

Content in the following languages has been translated into English. So possibly we may not have to deal with them, but we don't yet have information whether that covers all providers and all content.

- Dutch (**)
- French
- Hungarian
- Lithuanian
- Romanian

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

- Bulgarian (*)
- Italian (*)
- Greek
- Polish
- Spanish

The stars in the bullet points above mean the following:

- (**) languages in which Ontotext has production experience. Sophisticated NLP processing will be employed, such as: stemming/lemmatization, Part of Speech (POS) tagging, recognizing language-specific clues for entity extraction (e.g. "X of Y" is usually an organization, if "Y" is a place).
- (*) languages in which Ontotext has experimental experience. We'll deploy NLP techniques and resources as are available to us. Somewhat lower precision and recall can be expected
- No stars: languages in which Ontotext has little or no experience. Extraction will be based on gazetteer lookup only. Significantly lower precision and recall can be expected. We will have to rely on the respective content partners for feedback and quality evaluation.

Strong preference is given to data sources that have good coverage across a significant number of the required languages. In particular, Wikipedia has 2.11x inter-

language overlap, i.e. each concept is described in that many languages. We will leverage this by building a merged hierarchy of Multilingual Categories (see 2.3.2).

It is crucially important that during MINT conversion, every text metadata field of every CHO is properly marked with the appropriate IANA language tag. E.g.:

- "pasta"@en not "pasta"@eng nor "pasta"
- "Козунак"@bg or "Kozunak"@bg-Latn for languages that allow transliteration (in Bulgarian that is not really appropriate for cultural objects)

If this requirement is not met, it will limit severely the quality of multilingual enrichment

2.7 Manual Curation

The automatic enrichment produces match candidates that may be wrong, since it cannot deal with all ambiguity. Therefore a manual curation tool is needed to allow content providers and other interested parties to select the correct candidate, remove a match, or record a new match. The automatic enrichment should try to minimize the effort by:

- Setting a relevance threshold that is appropriate to the desired number of enrichments. E.g. the example in 1.11 has too few enrichments, but the one in 1.12 has too many.
- Use color and other UI mechanisms to enable "management by exception", i.e. draw the user's attention to occurrence matches of least confidence, documents with too few enrichments, documents with too much uncertainty, etc (see an example of the latter at the end of this section).

This should be a user-friendly tool that museum curators can use with confidence. E.g. below is a screen-shot for a similar tool developed for journalists:

OLYMPICS Athletics 2nd lead

Great Britain's **Jessica Ennis** stands just 800 metres away from Olympic heptathlon gold after she continued to dominate the competition at London 2012.

Ennis will take a massive 188-point lead into the final event this evening and only an unexpected injury setback realistically could prevent her following in the footsteps of fellow Briton **Denise Lewis**, who won the title in Sydney in 2000.

If **Ennis** runs a time of around two minutes 10 seconds in the 800m, nearest challenger **Austra Skujyte** of Lithuania would need to clock close to 1min 57secs to win, a time which would be good enough to get in the Olympic individual final.

Skujyte's personal best of 2:15.92 was set in the **Athens Olympics** in 2004, while **Ennis** set hers of 2:07.81 on the way to silver at the World Championships in Daegu last year.

Ennis could even become only the fourth woman in history to score 7,000 points with a run of 2:05.69 tonight, while equalling her personal best would give her 6,968 and fifth place on the all-time standings.

However, such achievements will be secondary to simply making sure of victory, four years after the 26-year-old was forced to watch the Beijing Games on television after suffering a career-threatening foot injury.

"Obviously I was anxious coming in today, the long jump has been up and down all year so I was quite worried about that," **Ennis** said. "I've done a lot of no jumps this year which I don't normally do."

In the Story

- + Add an event X
- + Add another event
- Athens Olympics (2)** aka Athens Olympics 2004 or Athens Olympics 1896 X
- Austra Skujyte** 33 years old ✓ X
- Denise Lewis** 40 years old ✓ X
- Jessica Ennis** 26 years old ✓ X
- London 2012** relates to Olympics ✓ X
- Mo Farah** 29 years old ✓ X
- Nataliya Dobrynska** 30 years old ✓ X
- Sydney** of country Australia (AU) ✓ X
- Tatyana Chernova** 24 years old ✓ X
- + Add an entity X
- + Add an entity X
- + Add another entity

In addition to document-level curation (CHO by CHO), collection-level processing can also be very useful. Below is a screenshot of such tool, developed for Euromoney (a large financial and investment publisher). It lists documents grouped by enrichment confidence, and below them entities grouped by recognition confidence. E.g. "Meat Loaf" is low-confidence for a Person in the financial domain (it would be high-confidence for a musical artist in the Music domain).

Curation dashboard articles entities 2014-03-27 18:25 Change

ARTICLES

- 18 Correct** view all
 - Diversity wins big at 86th Oscars** By JESSICA HERNDON and JAKE COVLE AP Film Writers LOS ANGELES (AP) -- Diversity was perhaps the biggest winner at the 86th annual Academy Awards, for the first time, a film directed by a black film...
 - Daphne Oz welcomes baby girl** NEW YORK (MYFOXNY) - Cardiac surgeon and TV host Dr. Mehmet Oz is a grandfather for the first time. He Tweeted on Thursday morning that his daughter, TV host Daphne Oz, had given birth. "My daughter...
- 132 Partially correct** view all
 - Peter Dinklage graces cover of Esquire's Style Issue** By Liz Raftery, Esquire has selected Peter Dinklage to appear on the cover of its Style issue, which will be for sale on newsstands Tuesday, according to Women's Wear Daily. Winter Olympics: The mo...
 - Drake slams Rolling Stone for misquoting him, putting Philip Seymour Hoffman on cover** By Liz Raftery, Drake won't be doing any more magazine interviews -- or so he says. The rapper and Degrassi alum slammed Rolling Stone Thursday for misquoting him about mentor Kanye West -- and for ...
- 4 Wrong** view all
 - TUESDAY, Jan. 28, 2014 (HealthDay News)** uri http://www.myfoxny.com/stor
 - TUESDAY, Jan. 28, 2014 (HealthDay News)** -- A 3-D model of the brain of a man who lived for 55 years with almost total amnesia is revealing new clues about what caused his memory loss, and could lead

ENTITIES

- 3363 New** view all
 - Healthday, Facebook, Google, YouTube, Twitter, Apple, Nokia, Samsung
- 1577 Ambiguous** view all
 - John Kennedy, Washington, Washington, University o, University o, University o, University o, University o, University o, University o, University o
- 596 Low confidence** view all
 - Good, The Voice, Meat Loaf, 2010 Nations, Mutual Inter-, Lauren, Sullivan, Steel mean

Cancel

2.8 Thematic Classification

EFD consortium partners have expressed an interest in defining a small thesaurus of common topics of interest (themes). Themes may provide an important feature for applications. It is hardly possible to derive the themes automatically, so they would be applied through manual classification. One of the following approaches is possible, to be selected and confirmed by the content partners:

- Apply themes to whole sub-collections (preferable), as part of the schema mapping process using MINT
- Apply themes to individual objects.

The maintenance of themes must also be undertaken by the content partners (led by PS), while Ontotext will provide the technical infrastructure:

- If the themes are a small number (10-100), we'll create them in a shared spreadsheet and then convert to SKOS.
- If the themes (or other shared classifications) will be bigger (hundreds or thousands of entries), a more serious collaboration environment needs to be deployed. An appropriate tool is VocBench⁴³, an open source SKOS+SKOSXL editor that works directly over semantic repository (Ontotext GraphDB). Although technically more complex, this approach is more scalable.

Themes under consideration may include the following (the list and hierarchy is by no means final). Please note that these are individual fixed values, not lists of values (e.g. lists of fests/events are part of Wikipedia):

- Cultural and Traditions
 - customs and traditions
 - heritage foods and recipes
 - cultures and food
- Industrial and Industrial/craft
 - agriculture
 - traditional food production and
- Time-based themes, e.g.
 - Daily life
 - Traditional holidays, remembrances, feasts
- Socio-cultural phenomena
 - famine
 - immigration
 - emigration
 - economic crisis
 - war-time food and advice
 - nostalgia
- Social use of food and drink

⁴³ <http://aims.fao.org/vest-registry/tools/vocbench-2>

D3.19 Semantic Demonstrator Specification

- food fests
- wine and beer fests
- drinking culture
- healthy eating

2.9 Semantic Search and Faceting

The semantic search should deal with all EFD classification dimensions (facets) as described in [Alexiev 2015a] sec 2. The first few (places, cultures, agents, events) are generic. The last few (concepts, foods, drinks, festivities) are FD-specific and the reason why we need EFD classification-specific processing (most of the previous subsections).

Each concept captured by Semantic Enrichment (Automatic Enrichment and Manual Curation) has additional useful information:

- It is part of a meaningful hierarchy, e.g.:
 - Chicken is Meat
 - Pestle is <grinding and milling equipment>⁴⁴, together with grinders, mortars, grindstones, manos, and (according to Getty) sausage stuffers
 - Austral Islands is in French Polynesia
 - Christmas Bread is a kind of Christmas food
 - Bread Loaf is a type of bread
 - Bread Loaf is associated with Bulgarian Cuisine
- Some have additional information (see sec. 2.3.7), e.g.
 - Bulgarian Cuisine is related to the place Bulgaria
 - Creole cuisine is related to the Creole culture
 - Austral Islands has coordinates S 23° 0' 0", W 150° 0' 0"
 - Bread is mostly made of grains

This enables powerful searches by concept, higher-order concept, geospatial (within rectangle or near a place), etc. It also enables faceting by object type, food or ingredient type, location etc. As an example, here is an integrated semantic search UI developed by Ontotext in the news domain (you can try a similar UI at this News Demo⁴⁵).

⁴⁴ <http://www.getty.edu/vow/AATHierarchy?find=&logic=AND¬e=&subjectid=300024716>

⁴⁵ <http://news.ontotext.com/>

The screenshot displays the SEMANTIC PUBLISHING interface. At the top, there is a search bar with filters for 'Richard Luker', 'Major League Soccer', and 'United States of America'. Below the search bar are 'Facets' and 'Trending' tabs. The main content area shows search results for 'US Society' with the article 'World Cup turns Americans on to football'. The article includes an image of a crowd and a date of 'Wednesday July 2 2014'. To the right of the article is a 'Top Concepts' sidebar listing terms like 'football', 'Richard Luker', and 'Beau Dore'. Below this is a 'Similar Articles' section with a snippet about 'US viewers learn to love the World Cup'. At the bottom left, there is a semantic facet table with columns for Type, Label, Person, Located in, and Has position. On the right side, there is a popularity comparison chart for 'American Football' and 'World Cup Soccer' over time, and a snippet of search results for 'President Barack Obama'.

Type	Label	Person	Located in	Has position
Person	Barack Obama President Obama	President Barack Obama president Obama US President Barack Obama US president Barack Obama Barack Obama, the US president Barack Obama, the US president Barack Obama, US president Barack Obama, US president	United States United States of Washington Washington DC	President president vice-president Japanese prime minister US president German chancellor premier

- Top-right has a conceptual search (in this case by Person, Concept, and Country)
- The main frame shows an article, with the top concepts and their relevance to the article highlighted in orange.
- It also shows Similar Articles, clustered by semantic features (see sec.2.13.4)
- Below it are some semantic facets (Person, Location, Position)
- On the right is a popularity comparison of two concepts (Soccer vs American Football) over time. This is directly applicable only to the current news domain, but a similar idea could be adapted to CH.

2.9.1 Auto-completion

Auto-completion is an efficient way to select a concept (article) from a large set. Use:

- Provider: to add or correct an object enrichment (article)
- Consumer: to select a concept (article) **or category** that has objects

E.g. below is an example from Ontotext's LinkedLifeData,⁴⁶ a large-scale semantic warehouse in the Life Sciences and Pharma domain.

⁴⁶ <http://linkedlifedata.com/search/quick>

asthma	
asthma	free text
Asthma	Disorders
ASTHMA BRONCHIAL, br asthma, asthmatic, Bronchial Asthma, Asthma, Asthma [Disease/Finding], ASTHMA, BRONCHIAL, asthma disorders, Asthmas, asthmatics, Asthma, Bronchial, Bronchial asthma, asthma, Asthma, unspecified, bronchial asthma, ASTHMA, bronchitic asthma, BRONCHIAL ASTHMA	
ASTHMA EDUCATION, NON-PHYSICIAN PROVIDER, PER SESSION	Procedures
ASTHMA EDUCATION, NON-PHYSICIAN PROVIDER, PER SESSION	
Asthma education	Procedures
asthma education health, asthma health education, health education - asthma	
BRACELET,MEDICAL ALERT ASTHMA	Devices
BRACELET,MEDICAL ALERT ASTHMA	
Chronic Obstructive Asthma	Disorders
Chronic obstructive asthma	

Useful auto-completion depends on the following factors (all are used in the above example):

- Use a full-text index to provide fast response. GraphDB Enterprise Connectors enable integration to Lucene, Solr or Elastic Search to be used for this purpose.
- Trigger the search after enough letters are typed and/or a sufficient timeout, to avoid user annoyance.
- Display useful information, e.g.:
 - Name
 - Type (person, place, concept...)
 - Short description (e.g. the first Wikipedia paragraph is saved in the DBpedia field dbo:abstract)
 - Relevance score, number of objects, number of sub-cats, articles, classified objects.
- Order by a combination of FTS (Lucene) rank and relevance, so the concepts that are most relevant to the search query are shown first.
- Optionally, hit highlighting, to quickly show which part of the text was matched.

2.9.2 Semantic Faceting

Semantic faceting shows the entities that are recognized in a collection of CHOs or documents. Entities are grouped by type (the different listboxes below) and ordered by number of occurrences. Selecting some of the entities limits the document set to those that contain all selected entities, and correspondingly refreshes the occurrence counts of all other entities, thus showing the number of co-occurrences. This allows

D3.19 Semantic Demonstrator Specification

the user to quickly narrow-down to some CHO that are of interest, without knowing beforehand what entities occur in the collection.

Below is an Ontotext example from the medical patents domain, where the entities are Drugs, Dosages, patent Applicants, etc. We will adapt the idea to the CH domain.

The screenshot displays the ExoPatent interface for a semantic search. The search term is 'ALUMINUM HYDROXIDE'. The interface is divided into several sections:

- Selected Items:** ALUMINUM HYDROXIDE
- Terms from FDA Orange Book:** This section contains five facets:
 - FDA Drug Name:** 100 of 684 shown below. List includes ACETAMINOPHEN, ACETIC ACID, ACYCLOVIR, ADENOSINE, ALBUTEROL, AMMONIUM CHLORIDE, AMOXICILLIN, AMPICILLIN, ANTIEMETIC, ATROPINE, AZITHROMYCIN, BACITRACIN, BAL, BUPIVACAINE, CALCIUM ACETATE, CEPHALEXIN.
 - Dosage Form:** 46 matching entities. List includes AEROSOL, CAPSULE, CLOTH, CONCENTRATE, CREAM, DISC, DRESSING, ELIXIR, EMULSION, ENEMA, FOR SOLUTION, FOR SUSPENSION, GAS, GEL, GRANULE, GUM, CHEWING.
 - Active Ingredient:** 100 of 972 shown below. List includes ACETAMINOPHEN, ACYCLOVIR, ADENOSINE, ALCOHOL, ALUMINUM HYDROXIDE, AMINO ACIDS, AMOXICILLIN, ASCORBIC ACID, ASPIRIN, AZITHROMYCIN, BETAMETHASONE, BIOTIN, CAFFEINE, CALCIUM, CALCIUM CARBONATE, CALCIUM CHLORIDE.
 - Applicant:** 2 matching entities: Actifas, Actifazneca.
 - Route of administration:** 33 matching entities. List includes BILIARY, BUCCAL, CARDIAC, DENTAL, EPIDURAL, IMPLANTATION, INHALATION, INTRACRANIAL, INTRADERMAL, INTRAMUSCULAR, INTRACULAR, INTRAPERITONEAL, INTRATEHAL, INTRATRACHEAL, INTRAUTERINE.
- Document Keyword Filter:** Patent Documents Containing FDA-related Terms. 1-10 of 360 documents matching the search criteria.
- Table of Patent Documents:**

Publication Date	Document Number	FDA Applicant	Title
26 April, 1994	US-5306485-A	RICHARDSON-VIDUIS INC.	Suncare compositions
19 November, 1996	US-5576014-A	YAMANUCHE PHARMACEUTICAL CO.,...	Intrabuccally dissolving compressed moldings ...
21 December, 1999	US-6004777-A	ARIZONA STATE UNIVERSITY; VIRO...	Vectors having enhanced expression, and metho...
12 August, 1997	US-5656653-A	SMITHKLINE BEECHAM PLC	Compositions containing histamine-H2-receptor...
23 December, 1997	US-5700459-A	HOECHST AKTIENGESELLSCHAFT	Pharmaceutical composition containing polyel...
08 January, 2003	EP-1273291-A1	HONEIL-PPC, INC.	Brittle-coating, soft core dosage form
18 January, 1995	EP-0453397-B1	WARNER-LAMBERT COMPANY	Multiple encapsulated flavor delivery system ...
24 May, 1995	EP-0301992-B1	CENTRO NACIONAL DE BIOPREPARAD...	Vaccine against group B <i>Neisseria meningitid...</i>

Europeana includes a similar faceted search, but the facets are fixed to ones that are explicit EDM fields or are easy to extract: media type, language, year, country, provider, aggregator, license. We will use these facets, and add flexible semantic facets, grouped by types such as Person, Organization, Place, Event, Food, Concept.

2.10 Visualisation

With such large amounts of data it is essential to find effective visualization methods. Visualisation will help:

- providers to navigate the classification, cut off inappropriate branches, and perform other crowd-sourcing actions
- consumers to explore objects in a large dataset

In addition to the visualization tools researched in [Alexiev 2015a], we may want to research these:

- <http://www.gojs.net/latest/samples/>
- <http://visjs.org> ^{47 48 49} , ,
- <http://js.cytoscape.org/>

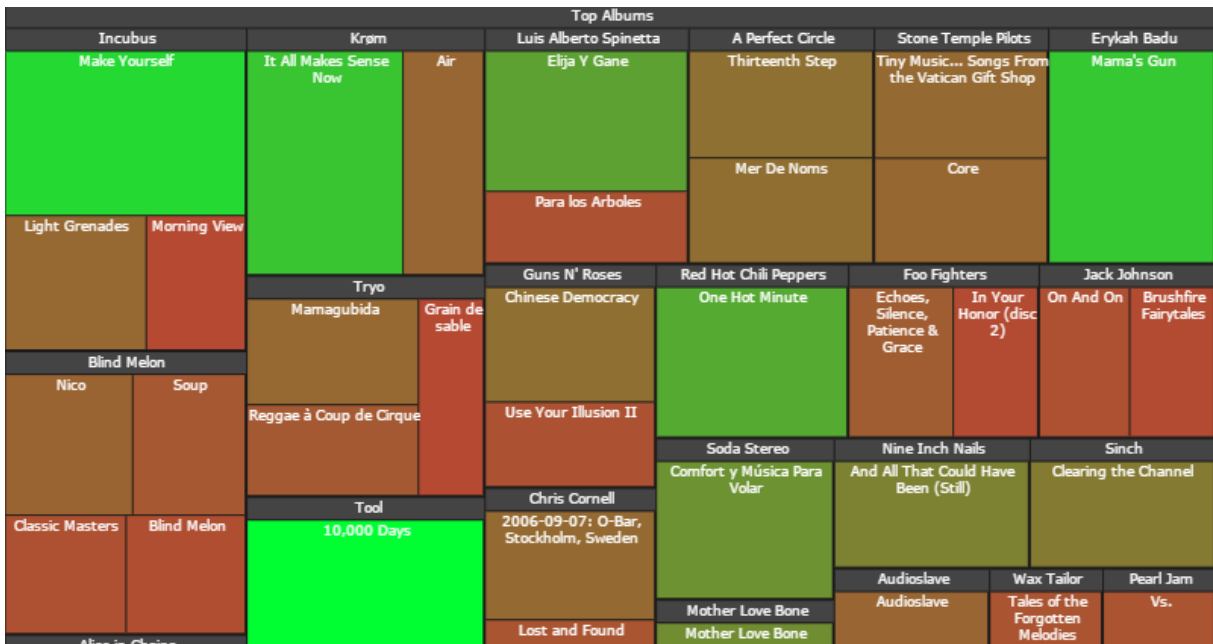
⁴⁷ http://visjs.org/network_examples.html#allExamples

⁴⁸ http://visjs.org/graph2d_examples.html#allExamples

⁴⁹ http://visjs.org/graph3d_examples.html#allExamples

2.10.2 Tree Map

A tree map shows category "size" (number of descendant categories and/or articles) by area. This example⁵⁴ shows some albums. 2-3 levels can fit.



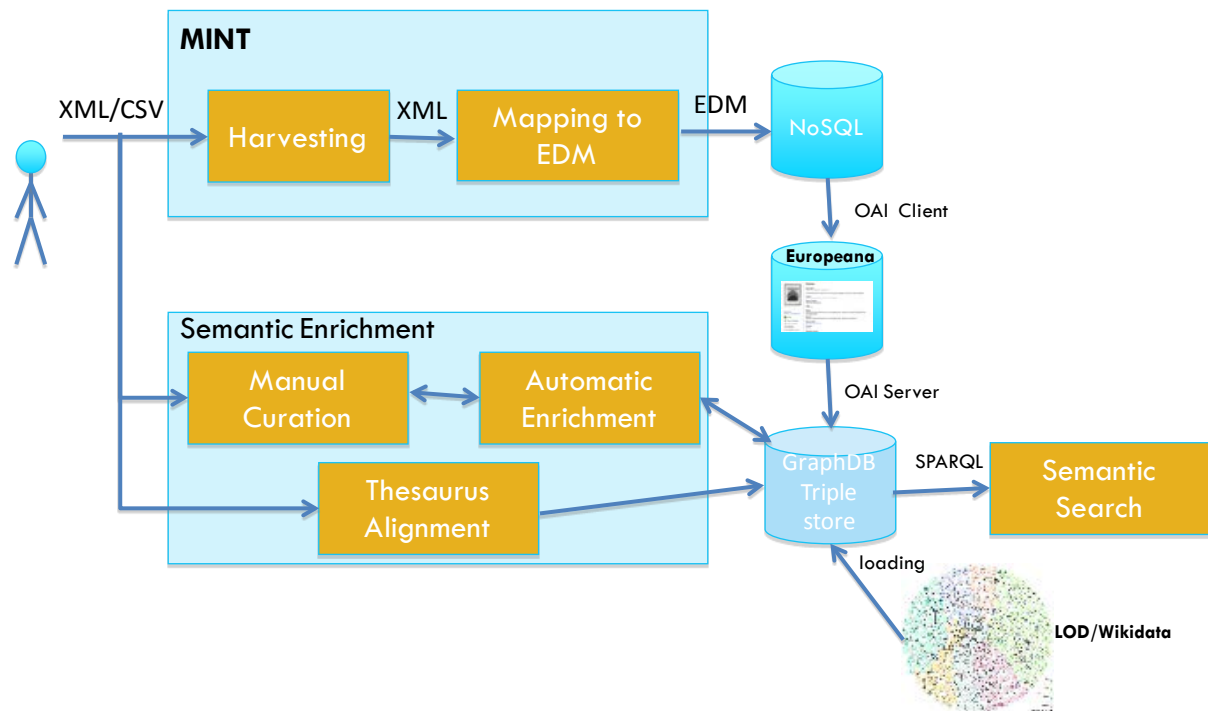
A zoomable treemap⁵⁵ allows you to drill down by clicking on a nested category, and go back up by clicking on the header.

⁵⁴ <http://philogb.github.io/jit/static/v20/Jit/Examples/Treemap/example1.html>

⁵⁵ <http://bost.ocks.org/mike/treemap/>

2.11 EFD Data Flow

An idealized data flow for EFD enrichment (and a draft system architecture diagram) is shown below:



The data flow is described as follows (capitalized phrases refer to previous headings)

- Ontotext provides an RDF repository (triple store) that holds existing Europeana objects in EDM, new EFD objects, and
- Ontotext loads the selected classification datasets and integrates them semantically to a Semantic Knowledge Base. This step doesn't show Category Management, but we'll need help from content providers with Manual Pruning.
- If the content provider has local thesauri, he aligns them to global datasets using Thesaurus Alignment tools provided by Ontotext
- The content partner maps his metadata to EDM with MINT and ingests it to Europeana (with the help of PS and NTUA)
- Ontotext refreshes the metadata from Europeana, using the OAI server developed within Europeana Creative. This can happen by time interval (e.g. weekly) and/or by collection (OAI Set)
- The content provider initiates Semantic Enrichment of his collection
- The EFD semapp performs automatic Semantic Enrichment
- The content provider performs Manual Curation to check validity or to correct match candidates. In response to user judgements, the semapp updates ML text analysis models (Category Scoring)

Unfortunately due to time pressures it's likely we'll need to cater to deviations from this workflow.

D3.19 Semantic Demonstrator Specification

- Our experience in Europeana Creative showed that data had to be processed several times, and in various formats (e.g. ESE, CSV), until the final workflow with EDM was achieved.
- We need to start working on NLP enrichment tasks ASAP, without waiting for providers to be ready with EDM conversion.
- So the semapp will implement **CSV and EDM file input**, in addition to EDM input from the semantic **repository**.

2.12 Discovering Europeana CHOs

A conservative guess is that between 1 and 10% of Europeana CHOs are related to FD, i.e. between 390k and 3.9M. This is a **lot more** than the 50k CHO to be collected in the EFD project. So a major goal of the semapp is to enable the discovery and classification of Europeana objects that relate to EFD.

This is very similar to EFD CHO classification, the main difference is size: Europeana has a total of 39M objects. Therefore it is recommended:

- Europeana classification to be done after EFD classification, leveraging the knowledge learned from that process
- Europeana EDM access to be made local (e.g. on the Ontotext repository) to avoid traffic delays
- Europeana classification will be performed on subsets, e.g. collections, parts of collections (e.g. by date range), or full-text pre-selections (e.g. keyword "jar", see next section). A FTS index is available in Europeana (Solr) as well as GraphDB (builtin Lucene or connected Solr or ElasticSearch)
- User confirmation of Europeana classification (manual curation) should be performed proactively after automatic classification, in order to augment the scoring of articles and categories used in the classification. In this way keywords appearing in Europeana FD objects will quickly extend the EFD classification

Another aim of the proactive confirmation is to avoid wrong learning:

- The user can correct a match if disambiguation could not pick the correct candidate
- The user can filter out inappropriate objects, described in the next section.

2.12.1 Filtering CHOs by Learning

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

- jar fragment⁶⁰ from the Petrie Museum of Egyptian Archaeology (UCL)
- cup fragment⁶¹ from the Fitzwilliam Museum, Cambridge, UK

⁶⁰ <http://europeana.eu/portal/record/2022347/8BA4652040C28D97167F10C8A07FB03747BCB5B8>

⁶¹ <http://europeana.eu/portal/record/2022304/1CDFAE9C1AC3F86DE38CAB40B6764324A1CF634F>



This is not nit-picking, since about half the objects with the keyword "jar" are fragments or shards. Since Europeana does not enforce **object quality** criteria, nor has **notability** criteria (like Wikipedia), all kinds of objects that hold only specialist interest have made their way into Europeana.

We can use machine learning approaches to acquire "blacklists" in a dynamic way. E.g. after a pre-selection for "jar" and automatic enrichment, the user is shown all matches, then

- He provides feedback by pointing a few that are relevant and a few that are irrelevant
- The application discovers keywords used in irrelevant objects (e.g. "fragment", "shard"), and records these as negatives
- The semapp offers manual curation of the relevant objects
- The semapp augments the score of the classification nodes used in the relevant objects, and their ancestors (spreading activation)

2.12.2 Filtering CHOs by Technical Metadata

Another common problem is the availability of good images. The thumbnails above are not useful for an application. The original page for the left object⁶² was not available on 24 Feb 2015 (the right object⁶³ was available). Such concerns are addressed by the Content Reuse Framework developed in Europeana Creative:

- Image Checker tool tries to fetch content objects (edm:WebResource) Technical Metadata tool extracts characteristics like image size and color distribution (e.g. from JPEG headers)

These tools record their results in the CHO metadata (against each WebResource, e.g. image URL). EFD should closely coordinate activities with Europeana Creative to see when metadata results will be available, and how they are expressed in EDM. Then we can formulate some static filters, e.g. "Image width has to be at least 1024 pixels".

2.13 Sample Apps

The main purpose of the EFD semapp is to build a **platform** for semantic enrichment and discovery/classification on a topic of interest. In addition, we describe here some

⁶² <http://www.ucl.ac.uk/museums/objects/LDUCE-UC47371>

⁶³ <http://webapps.fitzmuseum.cam.ac.uk/explorer/index.php?oid=68359>

sample applications that can be built on that platform. These are preliminary ideas, we reserve the right to define different applications together with consortium partners. See 2.9 for inspiration.

2.13.1 Topical Discovery

Technically speaking, the NLP and semantic approaches involved have little to do specifically with FD (apart from selecting root Wikipedia categories). One may as well define a domain of interest consisting of historic events and evidence about them (e.g. newspapers). Such powerful topical discovery tool was discussed several times in the frame of the Europeana Creative project with Steven Stegers, Deputy Director of EUROCLIO (European Association of History Educators). The Historiana Learning App⁶⁴ includes a "Search and selection tool" [Sanders 2013] that is a simplified version of an intelligent (semantic) selection tool. It searches by keyword and the same facets as the Europeana API.

We could not address the development of a semantic selection tool in the frame of Europeana Creative because of effort limits, scope limitations (we did not work on the History Education pilot) and immaturity of the OAI server and EDM dataset. But we hope to implement such tool in the frame of Europeana Food and Drink.

We call the discovery of existing Europeana content that's relevant to a topic of interest "Topical Discovery". To our knowledge, the attempt to use Wikipedia categories in order to delineate concepts in such a broad topic (e.g. FD) is completely innovative.

2.13.2 Timelines

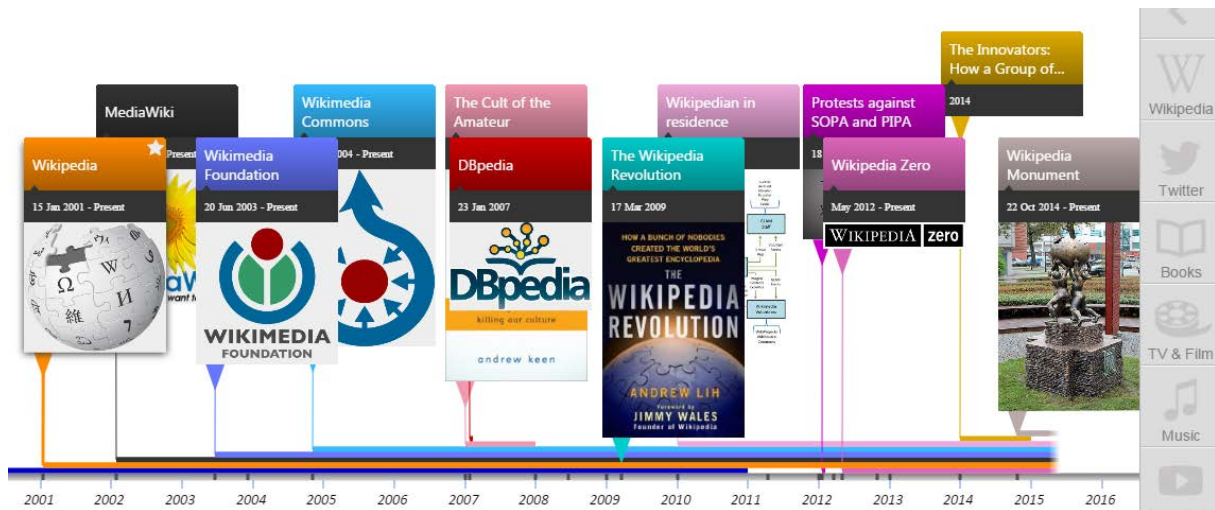
Armed with topical discovery and semantic classification, and assuming that a year is extracted from each CHO (that's standard Europeana enrichment), it should not be hard to make interesting timelines, e.g.

- **History of the Beer Jug:** we just need to find chronologically the first CHO representing each kind of beer jug/glass/stein
- **Food Nostalgia:** assuming manual Thematic Classification is performed (according to the themes outlined in 2.8) and there are a few CHO with this classification, we can just use them all.

There is a number of tools for designing timelines. E.g. here's a beautiful timeline of Wikipedia events created with Histropedia⁶⁵. (It includes some total opposites: from the book "The Cult of the Amateur" that criticizes user-generated content, to the "Wikipedia Monument" in Słubice, Poland.) Histropedia uses data from Wikipedia/Wikidata. But in our case, we want to make a timeline of **cultural objects** (CHOs) classified with Wikidata articles, not of **Wikipedia articles**.

⁶⁴ <http://la.historiana.eu/la/>

⁶⁵ <http://histropedia.com/timeline?4v4rtpg9bg0t>



2.13.3 Geographic Maps

Assume that we achieve:

- good classification of hunting/fishing objects
- decent geographic enrichment. That should be relatively easy at least for the major locations, especially if provided in a separate field (dct:spatial). It would be nice to distinguish between place created/used/found, but we are not sure there is such distinction in providers' metadata.

Then we can make an interactive map:

- **Hunting Round the World:** Show counts of hunting/fishing objects on a map of the world. When zooming, switch to individual pins. On mouse over, show a picture of the artifact and short metadata.
- **Journey of the Samovar.** Same, but for samovars (probably limited to Europe). Object type should be easy to recognize because of the unique designation. We could even use the Europeana 4D⁶⁶ explorer to make an animation.

2.13.4 Similar Objects

Europeana's standard "Similar objects" feature can be made much more precise if it's based on enriched concepts rather than simple keywords. E.g. it could find objects of similar type but submitted in different languages.

⁶⁶ <http://labs.europeana.eu/apps/Europeana4D/>

3 Conclusion

The Europeana Food and Drink Semantic Demonstrator will enable discovery, classification and exploration of Food and Drink cultural heritage objects, using advanced semantic technologies.

- For content providers, it will significantly add to their semantic skills and enhance the documentation of their collections.
- For end-users, it will offer an interesting tool for exploring concepts relating to food and drink. Using a rich array of thesauri and vocabularies that enrich digital objects with thematic, spatial, temporal, social, agricultural calendar and customs, festivities and cultural attributes, it will enable a fascinating journey into semantics. It will allow users to explore relationships between human history, society, living conditions, migration, agriculture and commerce and food and drink, highlighting how diverse content under a universal theme can be explored multi-dimensionally
- For Europeana, it will provide the beginnings of a Food and Drink Channel, and develop new technology for Topical Discovery that can be used in other domains as well, e.g. History (see sec. 2.13.1)

Here are some examples of the value created for various niche audiences:

- A food historian can find relevant objects to research historic commerce routes and how local foods and recipes travelled, affected by specific socio-economical and political developments.
- A fabric designer can use research evolution of patterns on china (porcelain) and draw inspiration from various eras and styles.
- A member of the GLAM community can use it to find interesting objects to use in virtual exhibitions
- A teacher can draw useful primary sources to assist him/her in developing teaching resources for the classroom.
- A chef or gastronomical start-up can source recipes from a particular era to develop a menu for their niche restaurant or cook book or foodie app

4 References

- [Alexiev 2015a] Vladimir Alexiev. Europeana Food and Drink Classification Scheme. Deliverable D2.2, Europeana Food and Drink project, February 2015. [Report](#), [Presentation](#)
- [Alexiev 2015b] Vladimir Alexiev. [Getty Vocabulary Program \(GVP\) Ontology 3.0](#). Namespace document, Getty Research Institute, March 2015.
- [Alexiev 2015c] Vladimir Alexiev, [Name Data Sources for Semantic Enrichment](#), part of Europeana Creative deliverable D2.4, Jan 2015
- [Gradmann 2010] Stefan Gradmann. Knowledge = Information in Context: on the Importance of Semantic Contextualisation in Europeana. [Europeana White Paper 1](#), April 2010
- [Charles 2014] Valentine Charles and Cécile Devarenne. [Europeana enriches its data with the AAT](#). Blog post, Europeana Pro, Nov 2014.
- [digiCULT 2013] digiCULT - Verbund, [xTree Product Flier](#) (in German). 2013
- [Haslhofer 2011] Bernhard Haslhofer, Antoine Isaac. [data.europeana.eu - The Europeana Linked Open Data Pilot](#). International Conference on Dublin Core and Metadata Applications (DC-2011), The Hague, 2011
- [Manguinhas 2014] Hugo Manguinhas, Antoine Isaac, Valentine Charles, Yorgos Mamakis, Juliane Stiller. [Europeana Semantic Enrichment Framework](#). Technical Documentation, 5 November 2014.
- [Lehmann 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer. [DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia](#). Semantic Web Journal, Vol. 6 No. 2, pp 167–195, 2015.
- [Sanders 2013] Nique Sanders, [Software development overview - History Education Pilot](#), Europeana Creative, Sep 2013
- [Schmachtenberg 2014] Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. Linking Open Data cloud diagram 2014, <http://lod-cloud.net/>
- [Stiller 2014] Juliane Stiller, Antoine Isaac, Vivien Petras et al. [Europeana task force on a multilingual and semantic enrichment strategy](#). Final report, April 2014