Selecting target datasets for semantic enrichment

29/10/2015

Editors
Antoine Isaac, Hugo Manguinhas, Valentine Charles (Europeana Foundation R&D) and Juliane Stiller

Co-financed by the European Union
Connecting Europe Facility

1. Introduction

As explained in the main report one of the components of the enrichment process is the target vocabulary. The selection of the target against which the enrichment will be performed, being a dataset of cultural objects or a specific knowledge organization system, needs to be carefully done. In many cases the selection of the appropriate target for the source data will determine the quality of the enrichments.

The Task Force recommends users to re-use existing targets when performing enrichments as it increases interoperability between datasets and reduces redundancies between the different targets[1]. Candidates datasets will be mentioned in this section and used as examples. We refer to the section "Find your datasets" for an exhaustive overview of these datasets.

However, in some cases it might not be possible to re-use an existing target and an enricher might decide to create its own target.

The Task Force explores both approaches using examples from the Europeana Network in the sections below.

2. Methods for analyzing and selecting datasets for semantic enrichment

The first approach in selecting targets for enrichment is to look at existing vocabularies available on the Web.

The Task Force identified a series of steps which provides a methodology to analyse datasets. The evaluation of the targets for semantic enrichment can be decomposed into 5 steps:

- Analyse your source data
- Identify your requirements
- Find your targets
- Select your targets
- Test the selected target.

The selection of a target in itself is supported by a series of more specific criteria.

These criteria will mostly help an enricher identifying the properties of a specific target which will help her to decide whether or not it should be selected for performing enrichment.

[1] http://www.w3.org/TR/2015/WD-dwbp-20150625/#dataVocabularies

2.1. Analyse your source data

A good knowledge of the source data is required before starting to select any target. The Task Force therefore recommends to perform an analysis of the source data prior to the selection of the target.
This analysis should look into several aspects of the data that will help identify the requirements enrichment should support.

- Define the scope or the domain of the source data.

Before looking into targets it is important to identify the different dimensions of the data that will need to be enriched or contextualised. Are you interested in a specific type of entity (places, time spans, agents…), from a specific domain?, in specific time range? For instance, in the case of place names enrichment, Geonames[2] might be more suitable for enriching contemporary place names as opposed to Pleiades[3] which focuses on historical place names.
The identification of the relevant dimensions can be done in selecting a list of keywords or categories that are representative of the source dataset. Identifying these keywords will also make the search for targets easier. Europeana has for instance selected a list of keywords[4] related to Art such as "Architecture, Baroque, Cubism" in order to perform enrichment for the Europeana 280 project[5].

- Identify the needs for enrichment.

Before starting data enrichment, one should have already identified the type of information requiring enrichment. This can be done by identifying the gaps in the dataset such as missing information that would add context to the data, quality issues that would require normalisation of the data against an authority or missing translations in a monolingual dataset. For instance, Europeana relies on multilingual labels to address the diversity of its data sources and for this reason Europeana tends to re-use generic but multilingual datasets such as DBpedia[6] rather than monolingual ones such as the Library of Congress Subject Headings[7] (LSCH) (only available in English, but aligned to French RAMEAU and German SWD by the MACS project [Landry, P. (2009)])

[2] http://www.geonames.org/
[3] http://pleiades.stoa.org/vocabularies
[4] https://docs.google.com/spreadsheets/d/1qjyyneg6aMoPC2v5hwC8YinmHKNyJtvTJp1HJdnnPc8/edit#gid=0
[5] http://pro.europeana.eu/pressrelease/europeana-280-art-from-the-28-countries-of-europe
[6] http://wiki.dbpedia.org/
[7] http://id.loc.gov/authorities/subjects.html

- Analyse the structure of your data.

It can be interesting to get an overview of the properties used in the source data containing the literals that will be enriched. For instance an enrichment based on concepts could be focused initially on the literals in properties such as dc:title, dc:alternative, dc:subject, dc:coverage.
OCLC[8] for instance provides statistical data on the use of MARC tags across its datasets [OCLC, 2015]. The Digital Manuscripts to Europeana[9] (DM2E) project has also worked on a methodology to evaluate the usage of classes and properties from the DM2E data model [Baierer, K., Dröge, E., Petras, V., Trkulja, V. (2014)]. Tools such as http://vocab.at/info can be also used to generate documentation about linked datasets.

- Identify the size of the dataset to be enriched as it might influence the choice towards a specific target and also the tool that will be used to run the enrichment.

2.2. Identify your requirements

The analysis of the source data should provide the enricher with a list of requirements the selected target should help addressing. In addition to the requirements coming from the analysis of the data, an enricher might identify additional requirements supporting specific services or applications.
The most common use case for enrichment is to support better search and browsing functionalities. In this perspective the following requirements will need to be formulated:

- If the enrichment is performed to improve search and browse across languages, which languages should be covered? For example, an enricher looking at the Art and Architecture Thesaurus (AAT)[10] would need to acknowledge the fact that the vocabulary focuses on English, Dutch, Spanish and Chinese languages and has a minor coverage of the Italian and French languages.

- Is the objective to make a domain specific dataset more generic? (which would make it more discoverable) or on the contrary is the objective to make a general dataset more specific by choosing a domain specific target? In the first case, vernacular datasets such as DBpedia or Wikidata[11] are relevant when in the second case more domain specific datasets such as Pleiades for historical place names or Iconclass[12] for Iconography might be more appropriate.

[8] http://www.oclc.org/
[9] http://dm2e.eu/
[10] http://vocab.getty.edu/
[11] https://www.wikidata.org/wiki/Wikidata:Main_Page
[12] http://iconclass.org/

- Is the objective to link resources to similar resources? For instance aligning works of art in Europeana such as "The Night Watch[13]" with the same work of art defined in Wikidata[14].

- Is the objective to link a resource to a controlled identifier? For instance aligning a person's name with its VIAF[15] or GND[16] identifier.

These questions refer to the different types of enrichment defined in **section 2 Concepts used in this report** in the main report. They should be asked prior to the selection of the target as it will determine what will make a target more relevant than another.

2.3. Find your targets

Once the requirements are formulated, the enricher will start looking for potential targets. Search engines like Linked Open Vocabularies[17] or repositories like the Open Metadata Registry[18] can be used to find targets. Community specific repositories also exist, such as the research vocabularies service from the Australian National Data service (ANDS)[19].
Several overview documents are also available on the Web and are beneficial. This report refers already to several datasets and we provide below more exhaustive list of pointers.

- The AthenaPlus[20] project provides overview of the Linked Open Data sources that can be used for linking cultural content [Köhler, W., Stein, R. (2013)]
- The EuropeanaConnect[21] project worked on the conversion and the alignments of several vocabularies [De Boer, V., Isaac, A. et al (2011)
- Linked Data vocabularies already used by Europeana data providers[22]
- List of datasets on the SKOS wiki [SKOS, 2015]. See also the list of SKOS vocabularies in Datahub[23] and the list of SKOS datasets with SPARQL endpoints[24].

[13] http://www.europeana.eu/portal/record/90402/SK_C_5.html
[14] https://www.wikidata.org/wiki/Q219831
[15] https://viaf.org/
[16] http://www.dnb.de/EN/gnd
[17] http://labs.mondeca.com/dataset/lov/
[18] http://metadataregistry.org/
[19] https://vocabs.ands.org.au
[20] http://www.athenaplus.eu/
[21] http://www.europeanaconnect.eu/
[22] https://www.assembla.com/spaces/europeana-r-d/wiki/Vocabularies_used_by_Europeana_data_providers
[23] http://datahub.io/dataset?q=&tags=format-skos

- Reference value vocabularies in the library domain gathered from the LLD XG use cases [Isaac, A., Waites, W., Young, W. & Zeng, M. (2011).]
- List of datasets relevant for the Digital Humanities [Ridge.M, 2012]

2.4. Select your targets

The Task Force proposes below a list of criteria that can be used to compare and evaluate targets. These criteria are organized around 7 dimensions: availability, access, granularity and coverage, quality, connectivity and size.

2.4.1. Availability

The selected targets should be technically available on the Web and according to the Linked Data best practices and recipes[25] and properly documented. The most common standards used to represent Linked Data are RDFS[26], OWL[27], or related W3C standards.
It is also interesting to evaluate the "use" of a dataset on the Web. The extensive re-use of a target is a good indication of its degree of maintenance (and the persistence of the URI's). The presence of metadata about the vocabulary indicating the data of the last update, the degree of connectivity of the dataset (see Linking criteria below) can be used to measure the degree of re-use of the dataset.

2.4.2. Access

The selected target should be available under an open licence that will make the enriched data easily shareable with other datasets. If the target is not open the licence should be clearly stated in the metadata describing the dataset.

These last criteria can be easily analysed if metadata about the vocabulary is available. The presence of a VOID file[28] is an indication that the vocabulary is properly maintained. PeriodO for instance provides information about the size of its dataset, the licence, the date of updates in its VOID file: https://test.perio.do/.well-known/void

2.4.3. Granularity and Coverage

The coverage identifies the target's domain and scope**.**
The granularity of descriptions will need to be evaluated for each of the dimensions: is the target specific to a domain or more generic? Depending on the granularity of the source data, selecting vocabulary that would too generic could introduce ambiguities or semantic flows. Similar issues could happen if the chosen target is from the wrong

[24] http://datahub.io/dataset?q=&res_format=api%2Fsparql&tags=format-skos
[25] http://www.w3.org/DesignIssues/LinkedData.html
[26] http://www.w3.org/TR/rdf-schema/
[27] http://www.w3.org/ TR/owl2-primer
[28] http://www.w3.org/TR/void/

domain. Europeana has been confronted to these issues while refining its enrichment process [Marlies Olensky, Juliane Stiller, Evelyn Dröge (2012)]. For instance the enrichment of the term "Drawing" in the context of Cultural Heritage was enriched with "(blood) Drawing" when selecting the GEMET[29] vocabulary as the enrichment target. An enricher might also want to pay a particular attention to the degree of expressivity of the dataset such as the completeness and granularity of the information it describes.

The dimensions answering 'Who?', 'What?', 'When?', 'Where?' questions are the most relevant to the cultural heritage domain as they help contextualising CH objects.

- **places:** When looking at place name datasets, time and spatial coverage need to be considered in details. Geonames refers to contemporary places when Pleiades is about historical places names. Figure 1 provides an assessment of a few Gazetteers according to the criteria defined in this section.

- **person:** VIAF, ULAN[30] , GND and Wikidata are potential datasets that can be used for person name enrichment. The Europeana Creative[31] project performed a comparison between several datasets and concluded that VIAF and Wikidata were the best targets available as Linked Open Data to perform person name enrichment [Alexiev, V (2015a)].

- **time spans:** The UK government vocabulary[32], PeriodO[33], English Heritage Periods[34] and soon ChronOntology[35] are examples of datasets specific to time spans. Time spans and years: UK has a vocabulary of years, months, quarters etc for example.

- **events**

- **concepts such as subject headings:**

    - Library of Congress Subject Headings (LCSH)[36], RAMEAU Répertoire d'autorité-matière encyclopedique et alphabetique unifié[37], Dewey Decimal

[29] http://www.eionet.europa.eu/gemet
[30] http://vocab.getty.edu/
[31] http://pro.europeana.eu/structure/europeana-creative
[32] http://reference.data.gov.uk/doc/year/1948
[33] http://perio.do/
[34] http://heritagedata.org/live/schemes/eh_period.html
[35] http://linkedgeodesy.org/chronontology-kick-off/
[36] http://www.loc.gov/aba/cataloging/subject/weeklylists/

Classification (DDC)[38] . These three subjects classifications have been the object of alignment work in the MACS[39] and CRISSCROSS[40] projects.

- Universal Decimal Classification scheme (UDC)[41] is also available as Linked Open Data and in SKOS.

- **types of objects**
- **more general vs domain specific vocabularies**

  - Getty Thesauri[42]: Currently available are the Art & Architecture Thesaurus (AAT), Thesaurus of Geographic Names (TGN), Union List of Artist Names (ULAN) and in the future the Cultural Objects Name Authority (CONA).
  - Iconclass classification system for art and iconography
  - Dismarc vocabulary for Sound Genres[43]
  - WW1LOD[44] unified terminology, events, places and actors related to the First World War
  - AGROVOC[45], which also links to the following resources: US NALT[46], FAO Biotechnology Glossary[47], EUROVOC[48], GEMET.
  - Tilde Terminology[49] provides access to terminological data in 24 EU official languages in a wide range of domains.

As part of the coverage criteria, the Task Force wants to bring a particular attention to the language coverage. Depending on the requirements, the level of multilinguality in a specific target is a crucial criteria of selection. For instance Geonames seems the only massively multilingual thesauri available for places. Geonames has the chance to be a large dataset but it is not true for all the multilingual vocabularies. Eurovoc for instance is rather generic and small.

A small coverage in language doesn't mean that the target needs to be excluded. The Library of Congress Subject Headings (LCSH) for instance is not a multilingual vocabulary but is commonly referenced and used for enrichment [Oshiba, T. and Takehana, K.

[37] http://rameau.bnf.fr/informations/rameauenbref.htm
[38] http://dewey.info/
[39] http://www.dnb.de/EN/Wir/Kooperation/MACS/macs_node.html
[40] http://ixtrieve.fh-koeln.de/crisscross/index_en.html
[41] http://udcdata.info/. It contains a selection of 2,600 top-level classes from the UDC, translated into 49 languages under http://creativecommons.org/licenses/by-sa/3.0/.
[42] http://vocab.getty.edu/
[43] http://www.dismarc.org/vocabulary/dmGenres/
[44] http://www.ldf.fi/dataset/ww1lod/
[45] http://aims.fao.org/standards/agrovoc/linked-open-data
[46] http://fan.sla.org/2011/10/nalt-now-available/
[47] http://www.fao.org/biotech/biotech-glossary/en/
[48] http://eurovoc.europa.eu/
[49] http://term.tilde.com/resources

(2014).]. It has also been aligned with other vocabularies (among others, German, French and Italian subject headings in MACS, Spanish subject headings for datos.bne.es), which makes it a kind of hub in a small multilingual network of subject headings.

The different coverage attributes will help assessing the adequacy of the target with the source data. They also should match the requirements of the services or applications that will use the enriched data.

2.4.4. Quality

The quality criteria refers to different intrinsic aspects of the target that can be manually or automatically assessed.

Structure and representation of the target: one will favour a target based on standard data model that will provide information on the structure the target should respect. For instance if the target is represented in SKOS, it provides information on the structure the target should follow. Quality issues related to the no- respect of the SKOS would then be easy to detect. In general targets should be self-descriptive; each property or term should have a label, a definition and a comment.

Representation of values: the quality of a target can be assessed based on the representation of the values in the target. The values should be normalised and preferably information about the normalisation rules should be provided in the documentation. This is an important aspect that could require the definition of normalisation rules prior the enrichment if this is the selected target. PeriodO[50] for instance provides detailed information on the way the dates in the vocabulary are normalised. Enrichment tools most often use target concept labels to build a gazetteer for entity recognition. Highly pre-coordinated thesauri like LCSH have long labels that combine many concepts into one (eg "Italian love poetry--17th century"). The chance that such a label will occur in free text is smaller compared to a thesaurus where the concepts are smaller and "atomic" (eg like Getty AAT). Therefore prefer non-pre-coordinated thesauri.

Representation of languages. Multilingual vocabularies use preferred and alternatives labels to indicate names variants or designations in different languages. The presence of language tags in the values is also required. One might want to have a look at the lexical variance of labels to verify that a target is truly multilingual. Finally information about character encodings should be explicitly represented.

[50] http://perio.do/data-model/

Richness of the target. The level of richness of the vocabulary can be evaluated by the amount of semantic relationships it contains. The completeness and correctness of the semantic relationships linking concepts together (relationships of type broader, narrower, co referencing links e.g. sameAs…) provide good indicators.

2.4.5. Connectivity

This criteria refers to how a target is linked to others. The level of interconnectivity of targets can be considered as quality criteria which the Task Force decided to separate from the previous section because of its importance.

The Task Force recommends to select targets that are well-connected, e.g., equivalent elements are indicated, or vocabularies already re-use each other, in order to avoid duplication and redundancy. Pivot targets should ideally be made of vocabularies having comparable importance and complementary coverage. They can help smaller and specialised vocabularies to be anchored to the Semantic Web. Linking and integrating datasets can also be a way to increase the language coverage in a given dataset.

This alignment work has been the focus of the Europeana Connect project. The project focused on trying to "anchor" smaller and specialised vocabularies such as Cornetto (in Dutch)[51], the Amsterdam Museum thesaurus[52] to larger and more general vocabularies.

The project converted a series of vocabularies in SKOS and produced alignments [De Boer, V., Isaac, A., et al (2011)] using the Amalgame platform[53].

The evaluation of the interconnectivity of a target can be done by assessing the amount and the quality of incoming and outgoing links.

2.4.6. Size

Depending on the size of the target dataset, the number of concepts is a criterion of selection. A high number of classes, properties and total amount of triples is preferable, if your alignment process can deal with the higher ambiguity. For example, GeoNames has 7.5M place names. The name "Guadalajara" limited to Mexico returns over 15 places, a lot of them are small pueblas with population under 15. Using extra features such as population size can improve precision by giving higher probability to more "important" places.

After the assessment of a target against its coverage, quality, connectivity and size, the evaluation of the availability and the access to the target should help a user to make its final selection.

[51] http://datahub.io/dataset/cornetto
[52] http://semanticweb.cs.vu.nl/lod/am/
[53] http://semanticweb.cs.vu.nl/amalgame/

		TGN[54]	Geonames[55]	Pleiades[56]	HPN [57]
General Information	**Scope**	Places important for the study of art and architecture	Broader	Historical places	Historical Places
	Type	Thesaurus	Large geographical vocabulary	Historical GIS?	Thesaurus
	Source of data	Manually curated	Spine gazetteer (aggregates data from about 100 different sources)	Manually curated (built by the community)	Spine gazetteer (aggregated from several local HPN data sources and generalized gazetteers like Geonames)
	Data model	In-house data model	In-house data model (i.e. Geonames Ontology) based on RDF	In-house data model	data model based on the CARARE metadata schema
	Access	Both dumps and webservices. Dumps are released biweekly. Output for webservices is encoded in a SKOS-extended format.	Both dumps and webservices.	Dumps (KML, CSV, and RDF) and synchronization services (RSS feed)	webservice
	License	Data is published as SKOS-extended format under the	CC-BY	CC-BY	CC-BY?

[54] http://vocab.getty.edu/
[55] http://www.geonames.org/
[56] http://pleiades.stoa.org/vocabularies/
[57] http://support.locloud.eu/LoCloud%20Historical%20Placenames%20Microservice

		ODC-BY 1.0 license.			
Content	**Geo-features**	over 1.4 million unique features of which 0.8 million are from North America	over 9 million unique features	34,827 unique places and 38,687 unique locations. Note that the concept of place together with location corresponds to geo feature in GIS. Locations can be annotated with a time period.	No metrics available.
	Names	over 2 million names Include preferred and alternative names, which can be further annotated with a time period.	over 10 million names. Includes preferred and alternative names, and also ancient names.	30,210 names. Names can have different types as they are considered as classes but makes no distinction between preferred and alternative. Names can be annotated with a time period.	No metrics available. Preferred and alternative names.
	Languages	English, vernacular language and other languages, with most terms being in English.	Terms in a wide range of languages	Vernacular and alternative languages	Very limited language support
	Temporal Coverage	Both historical and contemporary	Contemporary	Historical. In particular, Greek and Roman world, and is expanding into Ancient Near Eastern, Byzantine, Celtic, and Early Medieval geography.	Historical, but no time period information.
	Spatial Coverage	Global	Global	Global	Global or just Europe?
	Classification	Uses place types based on	All features are categorized	Uses a flat type system composed	Only 3 types (country, region and

		the structural vocabulary of AAT. It covers both physical (mountain ranges, oceans, seas, rivers, waterfalls, island groups, and deserts) and political features, which can be further annotated with a time span.	into nine feature classes and further sub-categorized into 645 feature codes.	of around 60 different types corresponding mostly to physical features	subregion) for predefined features
	Footprint	Point base (2-dim coordinate)	Point base (3-dim coordinate)	Polygon base (with 3-dim point coordinate)	Point base (2-dim coordinate)
	Place relations	hierarchical, equivalence, and associative relationships	hierarchical, and associative relations	Untyped associative relations	Fixed hierarchical relations (country, region and subregion) and Untyped associative relations
	Demographics	Population	Population	None	None
	Co-Referencing	None	DBPedia (through wikipedia links)	None	None

Figure 1: Evaluations of gazetteers according to the criteria listed in the report

3. Test the selected target

Once a potential target has been selected, it can be interesting to run an analysis that will simulate the enrichment and its results. This analysis can be done manually or using semi-automatic tools such as CultuurLink (previously known as Amalgame)[58] that allow the evaluation of matchings with a selected target on small sections of a source dataset. This exploratory testing allows the enricher to assess the coverage of the target with the source data, the level of semantics and ambiguities.
The following tests can be interesting to perform:

- Execute test queries on the source data using terms from the target to check that the selected target covers the source data. A low amount of results might bring to the conclusion that the selected target is not good enough.

- Check the semantic of the target terms against some of the source data terms. For instance you might conclude that the enrichment of the term"Ceramic" as a material with the term Ceramic from DBpedia (http://dbpedia.org/resource/Ceramic) might not be good as it describe the technique and not the material.

- Execute test queries to assess the level of ambiguities that an enrichment against the selected target could bring in. For instance a search in Europeana for the term " Bach, Johann Sebastian" shows that there is too much ambiguities between the painter and the musician to decide to enrich this term with the DBpedia term: http://dbpedia.org/resource/Johann_Sebastian_Bach

The different steps and criteria describe in this section provide a framework that will help users identifying targets suitable for their source data. However it is possible that the conclusion of this exercise is that no existing target can be found and that is therefore necessary to build a new target for enrichment.

4. Building a new target for enrichment

The following section looks into two different cases where a new target needs to be build for enrichment:

- a new target is built on top of an existing one
- a new target is built from scratch.

4.1. Refining existing targets to build a new target

[58] CultuurLink (http://cultuurlink.beeldengeluid.nl/app/#/) is based on the Amalgame tool developed during the project EuropeanaConnect project (http://semanticweb.cs.vu.nl/amalgame/)

If a target can't be re-used as such for enrichment, it is possible to create a new target by anchoring newly created terms to existing targets.

4.1.1. Building a classification for Europeana Food and Drink

The Europeana Food and Drink Classification scheme (EFD classification) is a multi-dimensional scheme for discovering and classifying Cultural Heritage Objects (CHO) related to Food and Drink (FD). The topic of Food and Drink is so pervasive in our daily lives and in our culture that assembling a small "specialist" thesaurus is not feasible (such specialist thesauri were successfully used in other Europeana projects, eg ECLAP on performing arts and PartagePlus on Art Nouveau). The Europeana Food and Drink project investigated about 20 datasets for their relevance to FD[59], including the Getty thesauri, Wordnet FD Domain, Wikipedia (in its 2 semantic data representations: DBpedia and Wikidata), etc. Wikipedia has been selected as the basis for the new FD classification and the Wikipedia Categories are used to construct a hierarchy of topics to be used for classification. Out of 800k categories, 15k were selected as relevant to FD. This represents 110k items (en.wikipedia articles or Wikidata entities) with about 300k labels that can be used for enrichment). The project uses innovative semantic technologies to automate the extraction of terms and co-references for other existing targets. The result will be a body of semantically-enriched metadata that can support a wider range of multilingual applications such as search, discovery and browsing [Alexiev, V. (2015b)], [Tagarev, A. et al (2015)].

4.1.2. Building a vocabulary for First World War in Europeana 1914-1918

In the example of Europeana 1914-1918 a set of non-semantic tag had been created to support the functionalities of the Europeana 1914-1918. This set of terms has been aligned with terms from the Library of Congress Subject Headings in order to create a new linked data vocabulary. Converted to Skos, the vocabulary is now maintained into a Europeana instance of the OpenSKOS vocabulary service, at skos.europeana.eu.

Similar methodologies have been used to build faceted classification such as the Artefacts Canada Humanities [Alberts, I., Mas, S., Ménard, E. (2009)].

4.2. Building a new target from scratch

Some users will make the decisions to build a new target from scratch. This case happens quite often when a thesauri exists as a non Linked Data resource. The process of converting a thesauri to Linked Open Data can be seen as a creation of a new target.

[59] Presentation at http://www.slideshare.net/valexiev1/europeana-food-and-drink-classification-scheme

References

Alberts, I., Mas, S., Ménard, E. (2009). Faceted classification for museum artefacts: a methodology to support web site development of large cultural organizations. Retrieved from http://www.iskouk.org/content/faceted-classification-museum-artefacts-methodology-support-web-site-development-large (August 07, 2015)

Alexiev, V. (2015a). Name Data Sources for Semantic Enrichment. Retrieved from http://vladimiralexiev.github.io/CH-names/README.html (August 07, 2015)

Alexiev, V. (2015b). D2.2 Classification Scheme. Europeana Food and Drink. Retrieved from http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Classification-Scheme-(D2.2).pdf (August 07, 2015)

Baierer, K., Dröge, E., Petras, V., Trkulja, V. (2014) Linked Data Mapping Cultures: An Evaluation of Metadata Usage and Distribution in a Linked Data Environment. Retrieved from http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/265/223 (August, 07, 2015)

De Boer, V., Isaac, A., Schreiber, G., Van Ossembruggen, J., Wielemaker, J. (2011). D2.3.1 – Multilingual mapping of schemes and vocabularies. Europeana Connect. Retrieved from http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/EuropeanaConnect/Deliverables/ECONNECT-D2.3.1-Multilingual%20mapping%20of%20schemes%20and%20vocabularies.pdf (August 07, 2015)

Isaac, A., Waites, W., Young, W. & Zeng, M. (2011). Library Linked Data Incubator Group. Datasets, Value Vocabularies, and Metadata Element Sets. Reference value vocabularies. Retrieved from http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025/ and http://www.w3.org/2005/Incubator/lld/wiki/Vocabularies#Reference_value_vocabularies (August 07, 2015)

Köhler, W., Stein, R. (2013). Review on Linked Open Data Sources. Athena Plus. Retrieved from http://www.athenaplus.eu/getFile.php?id=190 (August 04, 2015)

Landry, P.(2009) Multilingualism and subject heading languages: how the MACS project is providing multilingual subject access in Europe. Catalogue & Index: Periodical of the

Chartered Institute of Library & Information Professionals (CILIP) Cataloguing & Indexing Group, 157, 2009.

Marlies Olensky, Juliane Stiller, Evelyn Dröge (2012). Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy. Metadata and Semantics Research. 6th Research Conference, MTSR 2012, Cádiz, Spain, November 28-30, 2012. Proceedings. 2012, pp 252-263.

OCLC(2015). Marc Usage in WorldCat. Retrieved from http://experimental.worldcat.org/marcusage/ (August 07, 2015)

Oshiba, T. and Takehana, K. (2014). No. 197, December 2014, Web NDL Authorities: Authority Data of the National Diet Library, Japan, as Linked Data (Paper presented at IFLA 2014 Satellite Meeting "Linked Data in Libraries: Let's make it happen!"). Retrieved from http://www.ndl.go.jp/en/publication/ndl_newsletter/197/977.html#anchor04 (August 07, 2015)

Ridge, M. (2012). A collection of museum, gallery, library, archive, archaeology and cultural heritage APIs, machine-readable, linked and open data services for open cultural data. Retrieved from http://museum-api.pbworks.com/w/page/21933420/Museum%C2%A0APIs (August 07, 2015)

Simple Knowledge Organization System (SKOS) (2015). List of SKOS dataset. Retrieved from http://www.w3.org/2001/sw/wiki/SKOS/Datasets (August 07, 2015)

Tagarev, A., Tolosi, L., Alexiev, V. (2015) Domain-specific modeling: Towards a Food and Drink Gazetteer. In First International Keystone Conference, Coimbra, Portugal (accepted), September 2015. Retrieved from http://vladimiralexiev.github.io/pubs/Tagarev2015-DomainSpecificGazetteer.pdf (September 01, 2015)