D2.3.1 Multilingual mapping of schemes and vocabularies

This deliverable document reports on the results of task 2.3. We describe our approach to and the results of semi-automatic mapping between Europeana vocabularies in different languages

co-funded by the European Union

The project is co-funded by the European Union, through the eContentplus programme

http://ec.europa.eu/econtentplus

EuropeanaConnect is coordinated by the Austrian National Library

Distribution

Version	Date of sending	Name	Role in project
0.1	10.10.2011	Victor de Boer, Antoine Isaac, Guus Schreiber, Jacco van Ossenbruggen, Jan Wielemaker (VUA)	
1.0	12.10.2011	Juliane Stiller (UBER)	
1.1	24.10.2011	Victor de Boer (VUA)	

Approval

Version	Date of approval	Name	Role in project

Revisions

Version	Status	Author	Date	Changes
0.1		Victor de Boer, Antoine Isaac, Guus Schreiber, Jacco van Ossenbruggen, Jan Wielemaker (VUA)	10.10.2011	
1.0		Juliane Stiller (UBER)	12.10.2011	Minor changes
1.1	Final	Victor de Boer (VUA)	24.10.2011	Minor changes based on review by Elaheh Momeni

Table of Content

1. Introduction

To achieve the proposed interconnectedness of Europeana resources from the large number of different sources in different languages, the goal of EuropeanaConnect subtask 2.3 is to relate value vocabularies (thesauri, person authority lists, etc) that are relevant in the domains Europeana resources come from. Rather than attempting to produce a single unified ontology, our approach is to find alignments between the local vocabularies used to annotate the original data and more general pivot vocabularies.

As the number of different vocabularies and the number of terms in those vocabularies used by the Europeana content providers will be very large, there is a need for semi-automatic methods that produce mappings with a sufficient reliability.

The work in WP2.3 seeks to adapt and run the matching method and tooling developed in WP1.2 for producing automatic mappings between vocabulary elements. In this deliverable document, we present the results of our work.

1.1 Relation to other Tasks and Milestone documents

There is a strong connection between EuropeanaConnect task 1.3 and 2.3 and therefore we recognize the interdependencies and relations between the two tasks. More specifically, a number of the controlled vocabularies described in this document have been converted to the SKOS standard using the XMLRDF ingestion tool developed as part of task 1.3, and described in Milestone document M1.3.2. That same milestone document also documents the Amalgame alignment tool, used to perform many of the semi-automatic mappings discussed in this document. Section [VIC check this] 3.4 of this document gives a short description of Amalgame and its use for this task.

This deliverable document is partly based on the milestone document (M2.3.1) of the same task preceding it. Some of the text in this document is copied from that milestone document. Also, a number of the alignments that we report on in this deliverable document have already been described in Milestone M1.2.2: Semantics of descriptions aligned (intermediary).

2. Task specification

Multilingual alignments are a special case of vocabulary alignment. From a functional perspective, they enable similar features to the ones enabled by monolingual vocabulary alignments to be integrated in the semantic layer (query reformulation, browsing across concept networks, etc). The main difference is the scope: multilingual alignments allow to bridge collections from different countries, a crucial feature in the Europeana context.

2.1 Languages

The core language set as determined in the project objectives contains six languages (Polish was added later) that are the most common European languages and should be covered for multilingual search and browsing:

French	German	English
Italian	Spanish	Polish

Table 1: Europeana core languages

Additionally, four secondary languages were determined where multilingual capabilities should be offered, however not to the extent of the core languages:

Portuguese	Hungarian
Swedish	Dutch

Table 2: Europeana secondary languages

2.2 Matching value vocabularies vs. matching metadata element sets

The description of task 2.3 mentions the mapping of both controlled vocabularies and metadata schemes. However our focus is truly on matching elements from controlled vocabularies like thesauri, classification schemes, gazetteers and person authority lists, and not metadata element sets and profiles of them (Dublin Core, LIDO, CIDOC-CRM…).

Europeana users seldom perform fielded search, directly using specific metadata fields (e.g. “creator”) to constrain their query. Further, for doing this the Europeana portal offers mediation via its “advanced search” interface[1], which uses general categories abstracted from the fields used in ESE and manually translated in all European languages by Europeana partners. It is much more urgent to provide with alignment data which can benefit to all the searches are made on the values that would appear in those fields. E.g., to be able to retrieve all Polish items for Warsaw in the results of a search using the French word “Varsovie”.

2.3 Evaluation

In EuropeanaConnect two main sources of use cases can be used to validate the alignments, i.e., to determine whether they enable efficient implementation of desired functions:

- the current Europeana ThoughtLab
- the use case and requirements for semantic functions described in M1.4.1

The evaluation of the alignments presented here is out of scope for this document.

[1] http://europeana.eu/portal/advancedsearch.html

3. Approach

3.1 Vocabulary selection

When prioritizing the vocabularies to be aligned together, three main (interrelated) motivations come into play.

Institutional and collection adequacy: the vocabularies to consider first are the ones that are most relevant for the Europeana domain, in terms of scope and uptake. A general, museum-oriented vocabulary commonly exploited in many museums will be a priori more interesting than a specialized administrative vocabulary used at one or two providers. This can be mitigated by a more practical observation of the collections that are currently ingested in Europeana, or that are planned for ingestion in the short term.

Methodological adequacy: the first vocabularies to be chosen must fit well the approach we have chosen for linking the vocabulary within the semantic layer (cf. M1.2.1). In general we will aim at matching smaller and specialized vocabularies to larger and more general ones. Potential "pivot" vocabularies (multi-lingual, wide-coverage and widely adopted vocabularies) are thus of utmost priority. Note that our "pivot" approach is not strict. Slightly more specialized vocabularies may be used as anchoring points in the semantic layer, depending on the characteristics of CH domains and already existing alignments. For example, we initially retained Wordnet[2] as a general pivot for general topics. But the Library of Congress Subject Headings (LCSH)[3] or the Dewey Decimal Classification[4] can be used as a more specialized alignment focus, as these vocabularies are widely used in the library domain and already (partially) mapped to other vocabularies of that sector. It may also be that a "pivot" may be made of several vocabularies that have comparable importance and complementary coverage (e.g., from a lexical perspective). If these vocabularies are perfectly aligned together, it is possible to map one more specialized vocabulary to one or the other.

Usage adequacy: to match the current foreseeable priorities of Europeana users, different types of vocabularies must be considered, that provide values for the who/what/where/when queries:

- places
- persons
- date
- events
- types of objects
- general topics

Surveys especially indicate that places, topics and persons are currently at the core of users' concerns (Dobreva et al, 2010).

Also, features such as the number of concepts (as an indicator for the coverage and the grain of the vocabulary), the lexical coverage of the vocabulary's concepts (possibly in different languages), the completeness and correctness of the semantic relationships linking concepts

[2] Wordnet is a reference database for English, cf. http://wordnet.princeton.edu/

[3] http://id.loc.gov/authorities/

[4] DDC ist he most widely used classification system in libraries, coming in over 30 languages. See http://www.oclc.org/dewey/ and http://dewey.info/ for a (partial) linked data version.

together are considered when prioritizing the vocabulary to align. A large, well-structured vocabulary may enable more valuable semantic functions to be built on top of the object metadata.

Of course the licensing approach—namely whether Europeana may easily be allowed to get access and exploit the data from one vocabulary—plays an important role as well.

3.2 Mapping to multilingual pivot vocabularies

1. For persons, VIAF[5] and to a lesser extent ULAN[6] have been used. Getty ULAN in SKOS RDF is only available in a research only version. Note however that VIAF includes ULAN.
2. For places Geonames[7] and to a lesser extent Getty TGN[8] have been used. Geonames has more multilingual coverage, and is entirely free. Geonames is also being used for the first internal experiments at the Europeana Office. Like ULAN, TGN is only available in a research-only version.
3. For general topics, we did not have a single (or double) pivot vocabulary, rather, we use individual wordnets in specific languages (notably English, French and Dutch) connected to the Princeton English Wordnet. The latter was available in two versions (2.0 and 3.0), which are also aligned. Next to these, we use large library subject heading lists that have been manually aligned in the MACS project[9]: LCSH, RAMEAU[10], SWD[11]. Getty AAT, although only available in a research-only version has been used as an art-specific thesaurus. DBPedia and DBPedia categories[12] have also been used, as they form a de-facto pivot vocabulary for the Linked Data Cloud[13] also an option, which will be investigated in relation with WP1.2.4 (connection of the semantic layer to external knowledge sources).

As the scope of these vocabularies overlap, specific alignments have been created between them, e.g. between AAT and Wordnet and between LCSH and Geonames (Giunchiglia et al., 2010). “Associative” alignments should also be made between the pivots to allow for associative semantic search—e.g., matching the person pivot to the place pivot, to connect a person to the

[5] Virtual International Authority File. See http://viaf.org.

[6] Union List of Artist Names, by Getty. See http://www.getty.edu/research/tools/vocabularies/ulan/.

[7] See http://geonames.org.

[8] Thesaurus of Geographic Names, by Getty. See http://www.getty.edu/research/tools/vocabularies/tgn/

[9] See (Landry, 2010) and http://macs.cenl.org

[10] RAMEAU is the subject heading list used at the French National library. See http://rameau.bnf.fr/, http://stitch.cs.vu.nl/rameau for a linked data version.

[11] SWD ist he subject heading list used at the German National Library. See http://www.d-nb.de/standardisierung/normdateien/swd.htm and https://wiki.d-nb.de/display/LDS/ for a linked data version.

[12] These are the Wikipedia categories (http://en.wikipedia.org/wiki/Category) as represented through the DBpedia linked data project. See http://dbpedia.org.

[13] A visualization of the LOD cloud can be found at http://richard.cyganiak.de/2007/10/lod/

most relevant places for him. This will however not be investigated in the foreseeable future, as it is unsure whether Europeana will implement functionality based on such links soon.

3.2.1 Specialized vocabularies to be matched to pivots

Specialized vocabularies are the basic starting point of our vocabulary matching effort: we focus on trying to “anchor” these smaller, more focused vocabularies to larger-scope vocabularies.

Language and institution-specific vocabularies can either be mapped directly to the set of pivots (ideal solution), or to other vocabularies, which are mapped to the set of pivots. For instance RAMEAU can be aligned to Wordnet via LCSH, using the existing MACS mappings between RAMEAU and LCSH and an alignment we would create between LCSH and Wordnet.

The second solution, though less optimal for application scenarios (mapping links would have to be combined, potentially leading to lower precision) may turn cheaper if we can re-use existing alignments.

At the time of writing the current “local” vocabularies have been identified as target of matching efforts, both for core and secondary language sets, in Table 1 and 2. In WP2, we focus on the vocabularies that WP1 has received after the survey sent to Europeana partners. Other sources were also investigated, as reported in a previous WP2 memo (Gäde et al, 2010).

3.3 Interactive alignment and the Amalgame platform

As was identified in Milestone document M 1.2.1, we employed a semi-automated approach to the alignment of the large vocabularies. Rather than using one automatic method for all alignment tasks, each alignment is done in an iterative manner, through the combination of various mapping techniques. At each step in the iteration, one is able to do a quick evaluation of the intermediate results so that new decisions can be made. We also identified the need for the procedure to be transparent with respect to the verifiability, provenance and quality of the produced mappings.

This methodology was implemented in the Amalgame alignment tool, an alignment platform that focuses on the predictability and transparency of the alignment process by drastically reducing the complexity of the technology. It integrates various basic alignment methods such as methods based on (partial) label matching and hierarchical information. The tool allows the user to combine these different alignment steps into a transparent workflow. In each step, the intermediate results can be assessed through an integrated evaluation interface. Amalgame has been described in further detail EuropeanaConnect Deliverable document D 1.3.1 and in Hildebrand et al. (2011). A number of the alignments described here have been constructed with prototype versions of Amalgame.

3.4 Other sources of mappings

Before the prototype versions of Amalgame were available for alignment, a similar methodology was used to produce alignments for a number of mapping. In this methodology, multiple automatic methods were used to produce mapping sets of correspondences. For the individual mapping sets, samples were evaluated, leading to an assessment of the quality of the mapping sets. The overlap of the various sets was calculated, resulting in high-quality mappings. The methodology was described in full detail in Milestone document M 1.2.1 and [Tordai et al, 2009]. The mappings resulting from this methodology, the precursor of Amalgame, are also presented in the next section under the heading ‘automatic alignment (pre-Amalgame).

Furthermore, we import pre-existing links between the loaded source and target vocabularies. In some cases, these had to be converted to SKOS format.

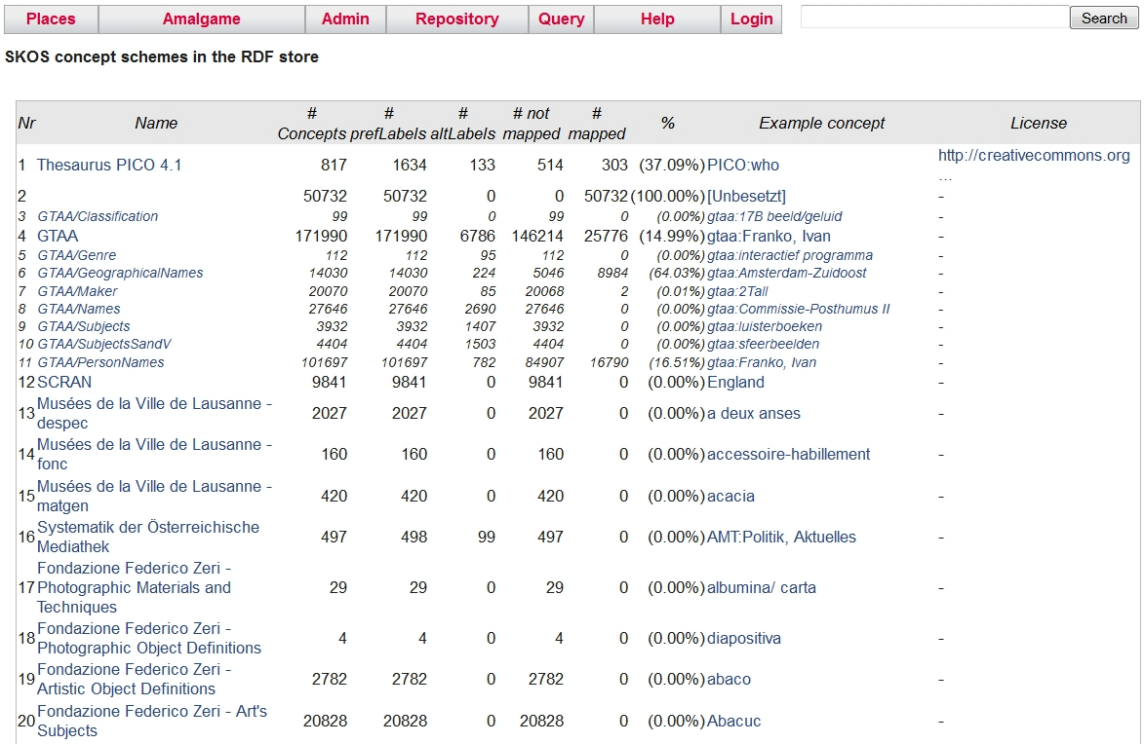
Lastly, we re-use mappings produced by the MACS project as well as those produced within the context of WP5.

3.5 Availability

The vocabularies as well as their mappings are available from two locations. The Europeana Semantic Layer provides web access through the ClioPatria semantic Web Server at http://semanticweb.cs.vu.nl/europeana/home. Here, the user is provided with a list of loaded vocabularies as well as basic statistics. Through this interface, the vocabularies can be further examined. The semantic layer also shows the loaded data and the alignments in the form of a data cloud graph (see figure 2).

The semantic layer provides a SPARQL endpoint[14] allowing for either manual or automated querying of the RDF database.

The vocabularies and the mappings are also available through the EuropeanaConnect SVN repository at http://sandbox08.isti.cnr.it/econnwp1svn/econnectwp1/trunk/vocs/.



Places | Amalgame | Admin | Repository | Query | Help | Login | Search

SKOS concept schemes in the RDF store

Nr	Name	# Concepts	# prefLabels	# altLabels	# not mapped	# mapped	%	Example concept	License
1	Thesaurus PICO 4.1	817	1634	133	514	303	(37.09%)	PICO:who	http://creativecommons.org …
2		50732	50732	0	0	50732	(100.00%)	[Unbesetzt]	-
3	GTAA/Classification	99	99	0	99	0	(0.00%)	gtaa:17B beeld/geluid	-
4	GTAA	171990	171990	6786	146214	25776	(14.99%)	gtaa:Franko, Ivan	-
5	GTAA/Genre	112	112	95	112	0	(0.00%)	gtaa:interactief programma	-
6	GTAA/GeographicalNames	14030	14030	224	5046	8984	(64.03%)	gtaa:Amsterdam-Zuidoost	-
7	GTAA/Maker	20070	20070	85	20068	2	(0.01%)	gtaa:2Tall	-
8	GTAA/Names	27646	27646	2690	27646	0	(0.00%)	gtaa:Commissie-Posthumus II	-
9	GTAA/Subjects	3932	3932	1407	3932	0	(0.00%)	gtaa:luisterboeken	-
10	GTAA/SubjectsSandV	4404	4404	1503	4404	0	(0.00%)	gtaa:sfeerbeelden	-
11	GTAA/PersonNames	101697	101697	782	84907	16790	(16.51%)	gtaa:Franko, Ivan	-
12	SCRAN	9841	9841	0	9841	0	(0.00%)	England	-
13	Musées de la Ville de Lausanne - despec	2027	2027	0	2027	0	(0.00%)	a deux anses	-
14	Musées de la Ville de Lausanne - fonc	160	160	0	160	0	(0.00%)	accessoire-habillement	-
15	Musées de la Ville de Lausanne - matgen	420	420	0	420	0	(0.00%)	acacia	-
16	Systematik der Österreichische Mediathek	497	498	99	497	0	(0.00%)	AMT:Politik, Aktuelles	-
17	Fondazione Federico Zeri - Photographic Materials and Techniques	29	29	0	29	0	(0.00%)	albumina/ carta	-
18	Fondazione Federico Zeri - Photographic Object Definitions	4	4	0	4	0	(0.00%)	diapositiva	-
19	Fondazione Federico Zeri - Artistic Object Definitions	2782	2782	0	2782	0	(0.00%)	abaco	-
20	Fondazione Federico Zeri - Art's Subjects	20828	20828	0	20828	0	(0.00%)	Abacuc	-

Figure 1: Part of the online table of loaded vocabularies in the Europeana Semantic Layer. The statistics shown are generated by the ClioPatria platform. This table can be found at http://semanticweb.cs.vu.nl/europeana/amalgame/list_skos_vocs

3.6 Licensing issues

There are a number of licensing issues that affect the availability of the vocabularies. Most prominent is the status of the Getty institute vocabularies: AAT, ULAN and TGN. These vocabularies are very much suited as pivot vocabularies for the cultural heritage domain.

[14] http://semanticweb.cs.vu.nl/europeana/sparql

However, we are only permitted to use them in a research context. As such it is not allowed to distribute them as Linked Data or import them in the Europeana production environment. The official licensing status of OCLC's VIAF dataset is also still unclear at the time this document is being written, though OCLC has agreed to make it entirely available for Europeana and many other partners.

4. Results

4.1 Data cloud

In this section we list the various vocabularies loaded in the Semantic Layer. In Figure 2, we show the graphical representation of the loaded vocabularies and their alignments in the form of a data cloud. The data cloud is automatically generated by ClioPatria. Larger nodes represent larger datasets.

An obvious issue that is apparent from the data cloud visualisation is that not all vocabularies that have been loaded have been mapped. As much of the effort of WP2 went into the conversion of the vocabularies as well as the development of the alignment platform, a limited number of alignment efforts could be made. The prototypical status of Amalgame did not allow us to do large-scale experiments with cultural heritage data experts to produce mappings for their own vocabularies.

Another issue is of computational nature. The largest node in the data cloud is that of VIAF, which we identified as a pivot vocabulary. The size of VIAF made it very difficult for any semi-automatic alignment using Amalgame, as it was too large to load in memory. We are working on resolving this issue. There are two mappings from vocabularies to VIAF: from Libris, which appears in the VIAF source and the relation to Getty ULAN which is actually a part of VIAF. Although these links did not show up in the data cloud, they are loaded in the Semantic Layer and we present them in Section 4.3. To indicate this, we augmented the data cloud graph with these links, represented by dashed lines. This also holds for a non-appearing link between the Euscreen thesaurus and LCSH. The most up-to-date and fully scalable version of the data cloud can be found at http://semanticweb.cs.vu.nl/europeana/data cloud.

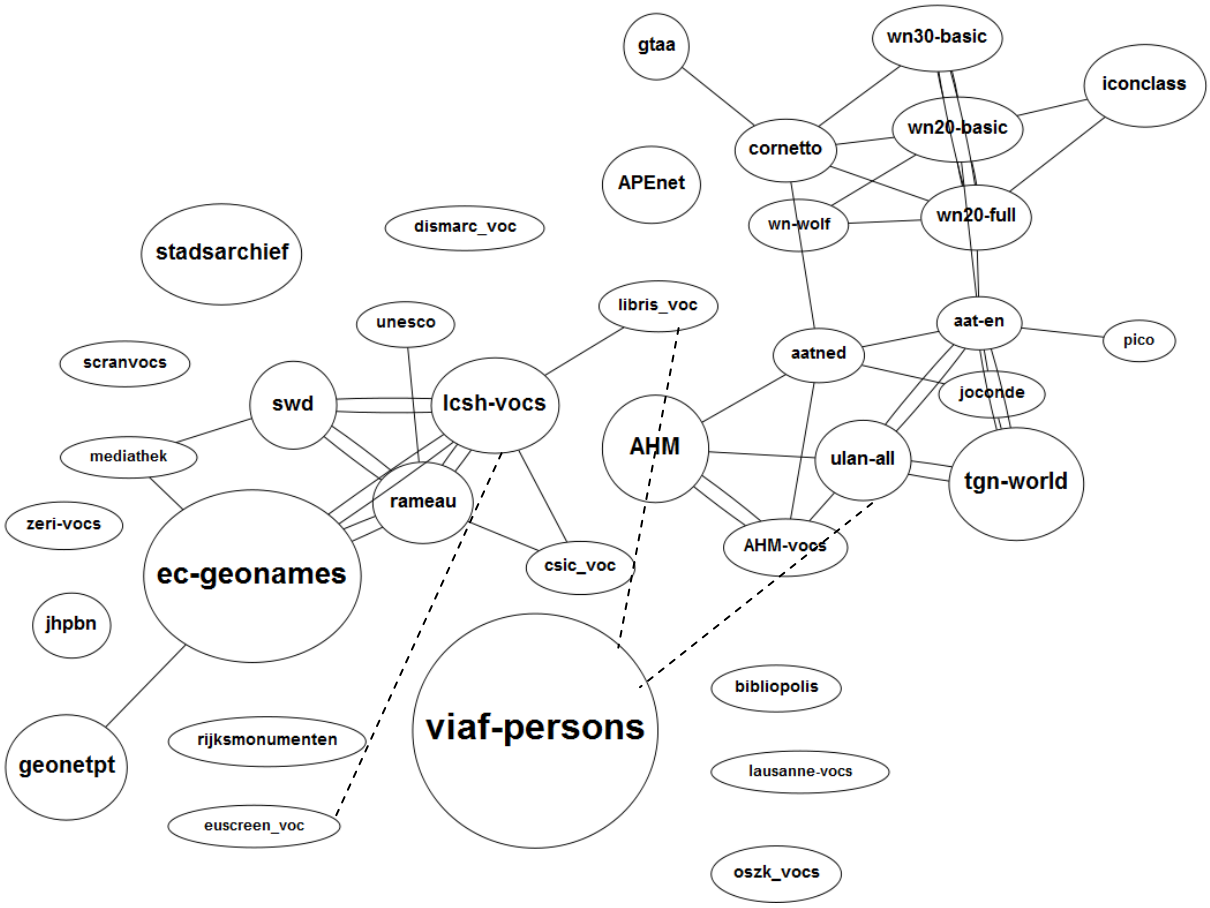


Figure 2: Data cloud visualisation as generated dynamically by the ClioPatria Semantic Layer. The three dashed lines represent links between nodes not yet showing up in the generated cloud.

4.2 List of converted vocabularies

In Table 2, we list the vocabularies now loaded in the Semantic Layer. Listed are the name of the vocabulary, the languages of its labels, the content of the vocabulary as well as any notes. In Section 5, we discuss the number of covered languages.

Vocabulary name	Lang.	Description/Type	Comments/status
Pico	It	Concepts	
GTAA	nl	Concepts, Geo	
SCRAN	en	Concepts	
Lausanne musea thesaurus	fr	Concepts	
Fondazione Zerri vocabularies	It	Concepts	
SWD	de	Concepts	
Cornetto	nl	Concepts	
CSIC (Spanish subject headings)	es	Concepts	only for use in Europeana
DISMARC	en	Concepts	
Getty AAT	en	Concepts	only for research purposes, in Europeana
Iconclass	de,en,fr	Concepts	
LCSH	en	Concepts	Pivot
AATNed	nl	Concepts	
Unesco	en,es,fr	Concepts	
VIAF	mul	Persons	Pivot, for exploitation and sharing within the Europeana network
Wordnet 2.0	en	Concepts	
Geonames	mul	Geo	Pivot
WOLF Wordnet	fr	Concepts	
Wordnet 3.0	en	Concepts	Pivot
Rameau	fr	Concepts, Geo	
Austrian Mediathek thesaurus	de	Concepts	
Amsterdam Museum Vocabularies	nl	Concepts, Geo, Persons	
EUScreen	ca, da, de, en, el, hu, it, nl, sv	Concepts	
Bibliopolis thesaurus	nl	Concepts	
OSZK Thesaurus	hu	Concepts, Geo	
Getty ULAN	mul (but labels are not language-tagged)	Persons	only for research purposes, in Europeana

DBPedia	mul	Concepts, Geo, Persons	Pivot, Not in SemLayer
GeonetPT	pt	Geo	
Polish Subject headings (JHP BN)	pl	Concepts	
Getty TGN	mul ((but labels are not language-tagged)		only for research purposes, in Europeana
Joconde	fr	Concepts, Geo, Persons	only for research purposes, in Europeana

Table 2: List of vocabularies loaded in the Europeana Semantic Layer.

4.3 Table of loaded mappings

Below, in tables 3, 4 and 5, we list the alignments between the various vocabularies that are currently loaded in the Europeana Semantic Layer. For each alignment we list the language of the source vocabulary, the source and target vocabularies, the number of source concepts matched and the origin of the mappings.

Eleven mappings have been produced through interactive alignment using the Amalgame alignment platform. These mappings are shown in Table 3. Other mappings have been done both manually and automatically mappings that have been done by WP2 researchers during or before the development of Amalgame, which are shown in Table 4. Finally, in Table 5 we present reused mappings produced by the WP5 Geoparser, the MACS project or mappings that are included in the original vocabulary distribution.

The mappings annotated with (*) between Libris and VIAF and between EUScreen and LCSH that appear in the tables below do not appear in the data cloud, this is a cloud-generation issue that will be resolved in the near future. A live result is available from the Semantic Layer at http://semanticweb.cs.vu.nl/europeana/amalgame/list_alignments.

Lang	Source vocabulary	Target Pivot(s)	Source concepts matched	Method
nl/mul	Amsterdam Museum Thesaurus	AATNed	3,753	Amalgame
nl/mul	Amsterdam Museum Thesaurus	Geonames	143	Amalgame
nl/mul	Amsterdam Museum Persons	ULAN	1,078	Amalgame

nl/mul	Amsterdam Museum Persons	DBPedia Persons	34	Amalgame
de/mul	Austrian Mediathek Thesaurus	SWD	156	Amalgame
de/mul	Austrian Mediathek Thesaurus	Geonames (Europe only)	47	Amalgame
it/en	PICO	AAT	232	Amalgame
pt/mul	Geo-Net PT	Geonames	3,140	Amalgame
en,es,fr	Unesco	RAMEAU	702	Amalgame
mul	Euscreen	LCSH	338	Amalgame
nl	GTAA	Cornetto	2,347	Amalgame

Table 3: Amalgame-produced alignments loaded in the Europeana Semantic Layer

Lang	Source vocabulary	Target Pivot(s)	Source concepts matched	Method
en	Getty AAT	Wordnet 2.0	10,608	Automatic match (pre-Amalgame)
nl	Dutch AAT	Cornetto	14,535	Automatic match (pre-Amalgame)
nl/en	Cornetto	Wordnet 2.0	47,490	Automatic match (pre-Amalgame)
nl/en	Cornetto	Wordnet 3.0	47,451	Automatic match (pre-Amalgame)
en	Wordnet 3.0	Wordnet 2.0	112,901	Automatic match (pre-Amalgame)
nl/en	AATNed	Getty AAT	27,050	Manual match (pre-Amalgame)
fr/en	WOLF	Wordnet 2.0	195,113	Automatic match (pre-Amalgame)
fr	Joconde	AATNed	15	Automatic match (pre-Amalgame)
fr	Joconde	Getty AAT	662	Automatic match (pre-Amalgame)
fr	Joconde	TGN	639	Automatic match (pre-Amalgame)

en, fr, de	Iconclass	Wordnet 2.0	7,372	Automatic match (pre-Amalgame)
mul	Getty ULAN	Getty TGN	73,211	Enriched from original source

Table 4: Pre-Amalgame mappings made by WP2 loaded in the Europeana Semantic Layer

Lang	Source vocabulary	Target Pivot(s)	Source concepts matched	Method
es/en	CSIC	RAMEAU	27,630	Available in CSIC source
es/en	CSIC	LCSH	28,923	Available in CSIC source
mul	Libris	VIAF	294,866	Available in VIAF source
fr/de	Rameau	SWD	20,538	MACS project (manual)
fr/en	Rameau	LCSH	55,964	MACS project (manual)
fr/mul	Rameau	Geonames	30,006	WP5 Geoparser (automatic)
en/mul	LCSH	Geonames	12,479	WP5 Geoparser (automatic)
en/de	LCSH	SWD	28,395	MACS project (manual)

Table 5: Imported mappings loaded in the Europeana Semantic Layer

4.4 Alignment strategy example

As an example of an interactive alignment, in Figure 2 we show the Amalgame provenance graph corresponding to the strategy used to align the Unesco thesaurus with RAMEAU. To deduce the different steps in the alignment strategy, the graph is to be read bottom up (against the direction of the arrows)[15].

In the first step in this alignment strategy, the SKOS preferred labels of the source and target vocabularies are matched. In the next step, the results are split by whether they are unambiguous (one source has one matching target) or ambiguous. The ambiguous results are evaluated by hand. Both sets of (presumed correct) matches are then subtracted from the original vocabularies, leaving the unmatched source and target concepts. Of these concepts, the SKOS alternative labels are matched to generate a new mapping.

In total, the unambiguous preferred label matches, the evaluated ambiguous preferred label matches and the alternative label matches on the ‘leftovers’ are selected by the user doing the alignment to be ‘finalized’ and added to the semantic layer.

[15] For a more thorough explanation of Amalgame provenance graphs, we refer the reader to Deliverable document D 1.3.1.

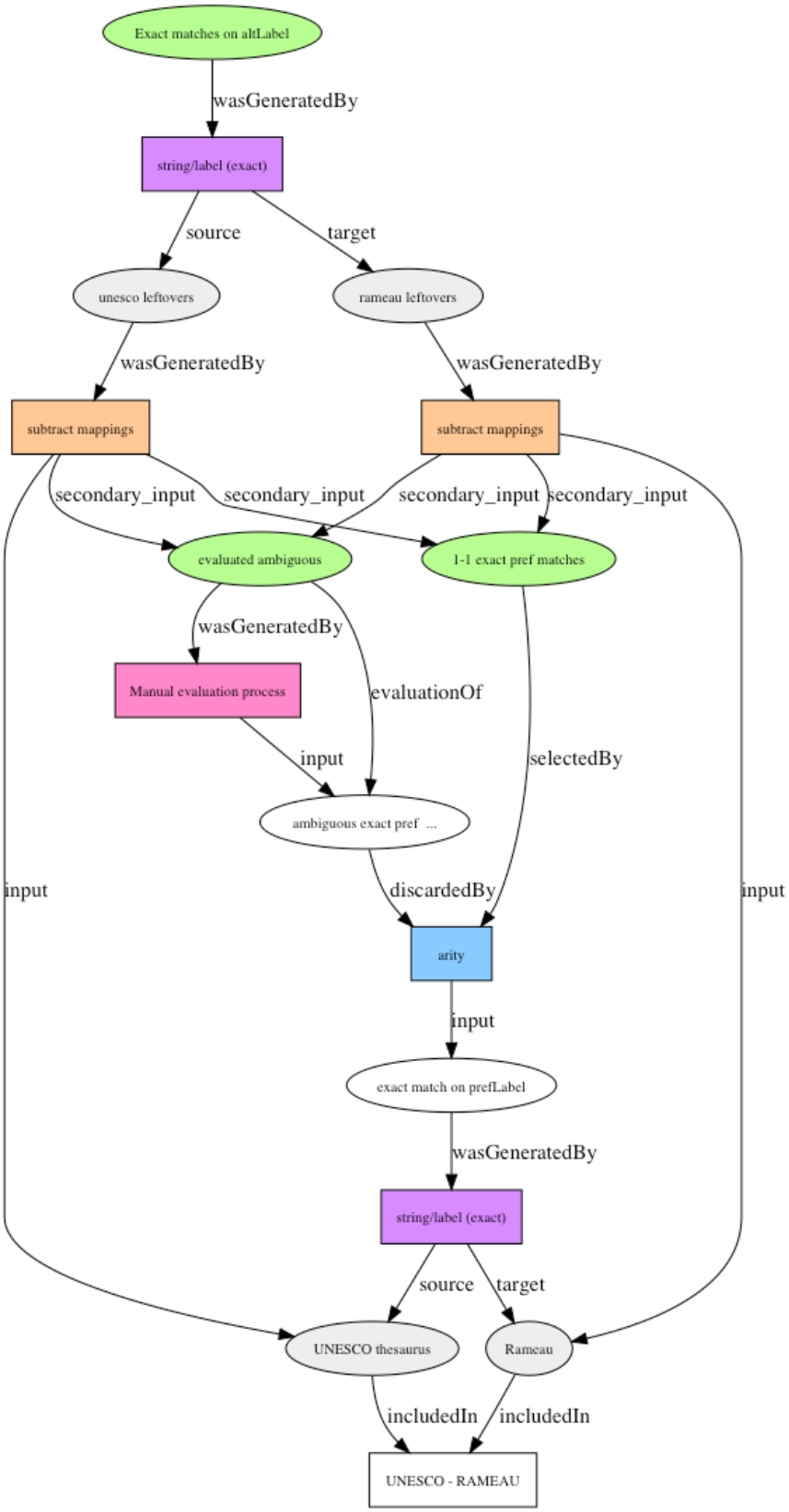


Figure 2: Amalgame provenance for aligning Unesco thesaurus with RAMEAU

5. Discussion and Conclusions

5.1 Languages covered

Of the six Europeana core languages (French, German, English, Italian, Spanish, Polish), five of the languages are represented by an aligned vocabulary. For the Polish language, we did not receive a vocabulary other than the Polish Subject Headings. We could convert it to SKOS format, but it consumed much time. The lack of availability of appropriate source and target vocabularies was identified as the major risk for the WP2 task in Milestone document M 2.3.1. Also, the monolingual aspect has not helped us to align it quickly with another source.

Of the secondary languages (Portuguese, Hungarian, Swedish and Dutch), we miss a mapped vocabulary exclusively Hungarian. The multilingual EUScreen vocabulary has labels in Hungarian, but is not mapped to any vocabulary. For Hungarian, the OSZK thesaurus is available but not mapped to any pivot of our vocabularies. Its geographic elements are however aligned to DBpedia. Had we loaded DBpedia in the semantic layer, this link would show in our cloud. But this would be a resource-intensive task, and DBPedia is freely available, so we postponed it. Readers should be aware that the most recently published version of VIAF and LIBRIS are interconnected, which is not yet shown in the semantic layer. For the Portuguese language, we obtained one language-specific thesaurus, Geo-Net PT, which was mapped to GeoNames. The other core languages are represented by multiple vocabularies.

We have a clear surplus of aligned vocabularies in English and Dutch. One reason for these languages appearing on the top of the list is that these languages are mastered by the Amalgame users. The interactive alignment strategy requires that the user is able to assess the quality of the intermediate alignment. As these alignments were mainly done by Dutch, English, and French speaking researchers, there is a clear bias towards these languages.

5.2 Total number of correspondences produced

In total, 1,047,818 correspondences are loaded in the semantic layer. 498,801 of these mappings were present in the original data sources. 537,047 mappings were produced by WP2 using enrichments of data sources, manual or automatic alignment techniques before the use of Amalgame.

Interactive semi-automatic alignment using Amalgame has resulted in eleven mappings of in total 11,970 correspondences. This is a relatively small percentage of the total number of correspondences. One reason is that they have been produced by the researchers during the Amalgame development process using prototype versions of the alignment platform. The focus here lies not in the sheer number of mappings but in the quality of the mappings as well as its provenance. At the same time, the prototype versions used did not allow for large-scale experiments to be conducted with actual cultural heritage data managers (Amalgame's target audience). Such experiments will result in larger amounts of new correspondences.

References

Milena Dobreva, Emma McCulloch, Duncan Birrell, Pierluigi Feliciati, Ian Ruthven, Jonathan Sykes, Yurdagul Unal . Europeana v1.0. User and Functional Testing. Final report (2010). http://version1.europeana.eu/c/document_library/get_file?uuid=1c25ae28-9457-4b0f-be62-654a7cf6c5b7&groupId=10602

Fausto Giunchiglia, Vincenzo Maltese, Feroz Farazi and Biswanath Dutta, GeoWordNet: a resource for geo-spatial applications. ESWC 2010. http://eprints.biblio.unitn.it/archive/00001777

Maria Gäde, Vivien Petras, Juliane Stiller, Antoine Isaac and Victor de Boer. WP 2.3 Multilingual Mapping of Controlled Vocabularies – Languages and Vocabularies for Selection. EuropanaConnect Memo, 29 April 2010. https://version1.europeana.eu/c/document_library/get_file?p_l_id=16989&folderId=26118&name=DLFE-13522.doc

Michiel Hildebrand, Jacco van Ossenbruggen and Victor de Boer. Vocabulary alignment as an interactive and replicable workflow. PrestoPrime Deliverable D4.2.1 (FP7-ICT-2007-3-231161), 18 February 2011http://semanticweb.cs.vu.nl/lod/prestoprimeD421/paper.pdf

Jacco van Ossenbruggen, Michiel Hildebrand and Victor de Boer. Interactive vocabulary alignment. Proceedings of the International Conference on Theory and Practice of Digital Libraries 2011. Berlin, September 26-28, 2011

Patrice Landry.Multilingualism and subject heading languages: how the MACS project is providing multilingual subject access in Europe. Catalogue & Index: Periodical of the Char- tered Institute of Library & Information Professionals (CILIP) Cataloguing & Indexing Group, 157, 2009.

A. Tordai, J. van Ossenbruggen and G. Schreiber, Combining Vocabulary Alignment Techniques. In K-CAP ’09: Proceedings of the 5th international conference on Knowledge capture, pages 25-32, 2009. ACM