

ECP-2006-DILI-510003

TELplus

Automatic subject alignment experiments

Deliverable number	<i>D3.5</i>
Dissemination level	<i>Public</i>
Delivery date	<i>4 January 2010</i>
Status	<i>V1.0</i>
Editor	<i>Antoine Isaac (VU)</i>
Contributors	<i>Shenghui Wang, Balthasar Schopman, Lourens van der Meij, Stefan Schlobach, Frank van Harmelen (VU)</i>



eContentplus

This project is funded under the *eContentplus* programme¹,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

¹ OJ L 79, 24.3.2005, p. 1.

TABLE OF CONTENTS

1	INTRODUCTION	3
2	AUTOMATIC ALIGNMENT STRATEGIES.....	4
2.1	INTRODUCTION	4
2.2	LEXICAL ALIGNMENT	5
2.3	EXTENSIONAL ALIGNMENT.....	7
3	EVALUATION OF AUTOMATIC ALIGNMENTS	12
3.1	COMPARING WITH MACS MANUAL MAPPINGS	13
3.2	SAMPLE MANUAL EVALUATION.....	16
4	FEASIBILITY ASSESSMENT OF AUTOMATIC ALIGNMENTS.....	19
4.1	LEXICAL MATCHING WITHOUT TRANSLATION.....	20
4.2	LEXICAL MATCHING WITH TRANSLATION	21
4.3	EXTENSIONAL MATCHING USING REAL DUALY ANNOTATED BOOKS.....	22
4.4	EXTENSIONAL MATCHING BASED ON INSTANCE MATCHING AND ENRICHMENT.....	24
4.5	CONCLUDING REMARKS	26
5	USEFULNESS ASSESSMENT OF AUTOMATIC ALIGNMENTS.....	26
5.1	USEFULNESS OF AUTOMATIC ALIGNMENTS FOR MANUAL MATCHING	27
5.2	USEFULNESS OF AUTOMATIC ALIGNMENTS FOR MULTILINGUAL SEARCH.....	28
5.3	CONCLUDING REMARKS	30
	REFERENCES	31
	ANNEX – SYSTEM SPECIFICATIONS	32

1 Introduction

General objective of TELplus WP3.2

The experiments pursued within the MACS project¹ demonstrated how mappings between different vocabularies can improve multilingual access to heterogeneous collections. This approach, even if interesting for portals like TEL, is however heavily hampered by the work manual mapping requires, given the size of the considered vocabularies.

WP 3.2 investigates how automated vocabulary alignment techniques can be applied to the MACS multilingual case, namely the subject headings of RAMEAU, LCSH and SWD.

The main objective is to investigate what can be done to help multi-lingual subject access in a TEL-like framework, from methodological, technical and practical points of view. How can we help vocabulary experts obtain mappings? How can we integrate the obtained mapping knowledge in portal frameworks? And how both mapping computation and integration can be carried out so as to give final users optimal tools to access heterogeneous collections?

Our purpose is to extend the MACS experiment in three ways: (i) applying and evaluating automated alignment procedures to MACS subjects; (ii) extending MACS alignments to subjects that were not investigated by lack of human resources, or achieving alignments with different semantics; (iii) taking into account the need for integrating the produced alignments into the TEL framework.

It is not expected that automation of alignment processes—using techniques such as the semantic-web based ontology matching tools² [E07]—can be fully achieved. State-of-the-art alignment tools still come with significant error rates. Further, in the TELplus context, our means are limited and the case at hand is difficult (cf. Section 2.1). However, these techniques may still enable to heavily speed up a (then semi-automatic) alignment. Mappings could also be sufficiently good to be used in the background of an appropriate search interface.

Contents of this report

This document extends a paper published at the European Conference on Digital Libraries [WISS09]. It first presents the vocabulary alignment techniques that we have implemented for our specific matching problem. It then reports on our evaluation of the obtained alignments. General lessons learned on applying the chosen techniques for the case at hand are given in the third section on “feasibility assessment”. We conclude by investigating the potential use of the mapping produced, exploring both a manual mapping scenario and the multilingual search engine to which TELplus WP3.2 contributed, and which is reported upon in D3.4.

Terminological clarifications

In the following, a *mapping* refers to a semantic link between individual concepts, such as equivalence between *Sprinting* and *Course de vitesse*. An *alignment* is a set of mappings that hold between concepts of two vocabularies, such as LCSH and SWD. A

¹ <http://macs.cenl.org>

² <http://www.ontologymatching.org/>

matcher (or *matching tool*, or *mapper*) is a software tool that delivers alignments between two input vocabularies.

2 Automatic alignment strategies

2.1 Introduction

Scope and difficulties of the alignment case

MACS focuses on three subject heading lists (SHLs) used respectively at the British, French and German national libraries, namely LCSH, RAMEAU and SWD. Those vocabularies are very large, as shown in Table 1.

Vocabulary	Number of concepts
LCSH	339,612
RAMEAU	154,974
SWD	805,017

Tab. 1. Size of SHLs (at the time those vocabularies were sent to us)

Size makes the matching problem very hard to solve. State-of-the-art generic ontology matching tools cannot handle large knowledge structures. The languages of the vocabularies is a second issue, as state-of-the-art tools are mostly focusing on matching English vocabularies, not to mention vocabularies in different languages. Also, the SHLs lack the kind of rich semantic structure that is relied upon by many tools. Here, concepts are related together by hierarchical and associative links (see Tab. 2), but that semantic structure is nor very rich nor consistent across (and within) the vocabularies. Finally, our case lacks auxiliary information that can be useful, such as a multilingual semantic network (to be used as intermediary background knowledge for the matching process (as WordNet¹ is often used for English), or a full-text version of the books described using the SHLs, which could have allowed us to employ corpus alignment techniques.

Vocabulary	Number of concepts	Number of “broader” links	Number of “related” links
LCSH	339,612	247,117	21,380
RAMEAU	154,974	129,347	60,464
SWD	805,017	282,985	30,045

Tab. 2. Semantic structure information in the three SHLs

It is also important to note that the usage of concepts with apparently similar meaning might differ. A preliminary study to the MACS project has indeed shown differences in indexing

¹ <http://wordnet.princeton.edu/>

strategies in the libraries using the three SHLs [L99]. This can hamper the usefulness of the techniques that are based on direct label comparison for finding matches.

As a matter of fact, we have organized over the past three years a Library Track in the context of the Ontology Alignment Evaluation Initiative,¹ a yearly campaign aimed at evaluating state-of-the-art ontology matching tools on specific cases. The first two years, this Library Track focused on two smaller thesauri from the National Library of the Netherlands, and each time only three matching tools participated, with encouraging but imperfect results. In 2009, we offered the participating matching tools to work on our TELplus case.² Only one tool managed to send results, which were rather weak.

Contents of this section

This section reports on how, resulting from the above problems, we had to turn to practical, adhoc solutions, which can cope with both the size and the multilingual aspect of our case. We have implemented four straightforward methods for vocabulary alignment, two based on lexical properties of the concept labels, and two based on the extensions of the concepts, i.e. the objects annotated by them. The simplest approach lexically compares the labels of concepts without translating them; a more sophisticated version uses a simple translation service we could deploy out-of-the-box. A simple extensional method makes use of the fact that all three collections have joint instances (common books), which can be determined by common ISBN numbers. Finally, we adapt work on matching based on instance similarity to the multilingual case.

2.2 Lexical alignment

Mapping tools that exploit the lexical information of concepts in controlled vocabularies, such as their various labels, are numerous. However, those mostly use language-agnostic heuristics, such as string distance, to compute similarity between concepts' lexicalizations. This gives results of relatively good quality for English, but can prove much less beneficial with other languages.

To introduce basic language-aware word comparison, we have adapted to French, English and German languages a lexical mapper that was first developed for Dutch [MIGB07]. This lexical mapper is mostly based on the CELEX³ and DELA⁴ databases, which allow recognising morphological variants of words (via lemmatization) as well as their morphological components (using knowledge on derived or compound words). It can thus produce “semantic equivalence” mappings between concepts, when they have labels which are strictly equivalent or correspond to a same lemma, but also hierarchical (broader) matches, when one concept's label is a morphological component one of the other concept's labels.

¹ <http://oaei.ontologymatching.org/>

² <http://www.few.vu.nl/~aisaac/oaei2009>

³ <http://celex.mpi.nl/>

⁴ <http://infolingu.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html>

Here, we focus however only the equivalence links, as those are the ones that were judged relevant in the MACS project.

The different lexical comparison methods give rise to different confidence levels for the resulting mappings: a mapping based on two concepts' having exactly the same label is assumed to be more reliable than one that is based on equivalence between the labels' lemmas. Also, the mapper uses the "status" in the controlled vocabulary for the lexical features it compares. As part of WP3.2 work, we created SKOS¹ representations of the vocabularies. In this representation, mirroring what can be found in the original vocabulary data, labels of concept can be *preferred* or *alternative*.² The latter ones can in fact be approximate synonyms, thus not fully reflecting the intended meaning of a concept. For two concepts, any comparison based on alternative labels is therefore considered less reliable than a comparison based on preferred labels.

The combination of these two factors (different comparison techniques and different features compared) allows us to grade the produced mappings. We use for this adhoc values³ which are merely aimed at producing a *ranking* of confidence levels.

The basic version of the mapper only uses one language at a time. To apply it in a multilingual case, we translated the vocabularies beforehand. For each vocabulary pair (e.g., RAMEAU and LCSH), we translate each vocabulary by adding new labels (preferred or alternative) that result from translating the original labels, using the Google Translate⁴ service. We then run the mapper twice, once for each language of the pair. In the RAMEAU-LCSH case, for instance, the translation of RAMEAU to English is matched (in English) to the original LCSH version, and the translation of LCSH in French is matched (in French) to the original RAMEAU version. The obtained results are then merged.

2.2.1 Results

For each case, we ran twice the lexical matcher: the first run does not use translations, while the second uses the basic translation strategy presented above. The number of mappings obtained in that process is rendered in the following table:

Lexical technique	LCSH-RAMEAU	RAMEAU-SWD	LCSH-SWD
Not using translation	11,345	11,215	13,119
Using translation	57,595	42,869	44,017

Tab. 3. Number of lexical mappings obtained

¹ <http://www.w3.org/2004/02/skos/>

² <http://www.w3.org/TR/skos-primer/#seclabel>

³ For instance, a mapping based on two preferred labels' being equal has a score of 0.95; a mapping based on the lemma of an alternative label for a concept being equal to the lemma of an alternative label of another concept will have a lower score (0.85).

⁴ <http://translate.google.com/>

2.3 Extensional alignment

Instance-based matching techniques determine the similarity between concepts by examining the extensional information of concepts, that is, the objects (*instances*) they classify. In a library case: those objects will be the books that have these concepts as subject.

The principle of such techniques, already used in a number of works like [V97], is that the similarity between the extensions of two concepts reflects the semantic similarity of these concepts. This is a natural approach, as in most ontology formalisms the semantics of the relations between concepts is defined via their instances. This also fits the notion of literary warrant that is relevant for the design of controlled vocabularies in libraries.¹

2.3.1 Instance-based matching using overlap of common instances

A first and straightforward method is to measure the common extension of the concepts (the set of objects that are simultaneously classified by both concepts) [IMSW07]. This method has a number of benefits. Contrary to lexical techniques, it does not depend on the concept labels, which is can be especially useful when the ontologies or vocabularies come in different languages. Moreover, it does not depend on a rich semantic structure; this is important in the case of SHLs, which are often incompletely structured.

The basic idea is simple: the higher the ratio of co-occurring instances for two concepts, the more related they are. In our application context, the instances of a concept c , noted as $e(c)$, are the set of books related to this concept via a subject annotation property. For each pair of concepts, the overlap of their instance sets is measured and considered as the confidence value for an equivalence relation.

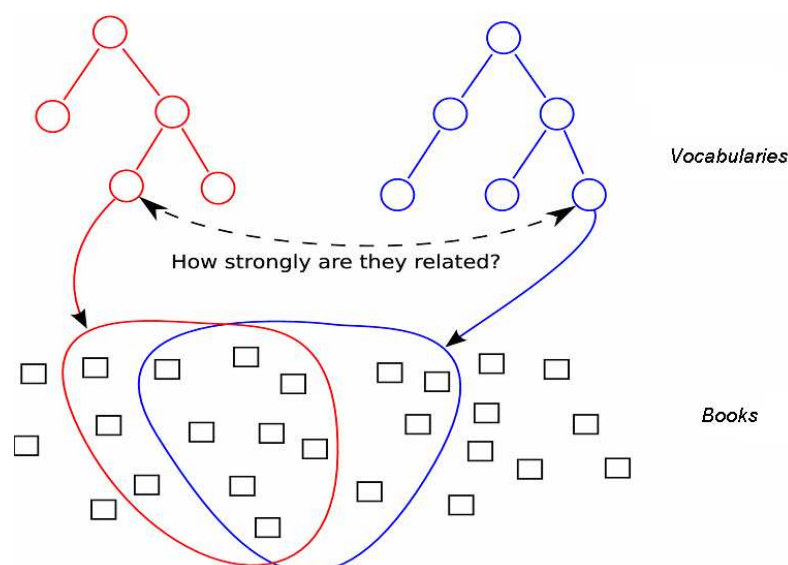


Fig. 1. Simple instance-based method

¹ As Svenonius wrote in [S00] "As a name of a subject, the term Butterflies refers not to actual butterflies but rather to the set of all indexed documents about butterflies. [...] In a subject language the extension of a term is the class of all documents about what the term denotes, such as all documents about butterflies."

Our measure, shown below, is an adaptation of the standard Jaccard similarity, to avoid very high scores in the case of very few instances: the 0.8 parameter was chosen so that concepts with a single (shared) instance obtain the same score as concepts with, in the limit, infinitely many instances, 20% of which co-occur. This choice is relatively arbitrary, but this measure has shown to perform well on previous experiments in the library domain for the STITCH project [IMSW07].

$$\text{overlap}_i(c_1, c_2) = \frac{\sqrt{|e(c_1) \cap e(c_2)| \times (|e(c_1) \cap e(c_2)| - 0.8)}}{|e(c_1) \cup e(c_2)|}$$

Note that one concept can be related to multiple concepts with different confidence values. In our experiments, we consider the concept with the highest confidence value as the mapping to be included in the final alignment.

2.3.2 Instance-based matching using instance matching and enrichment

Measuring the common extension of concepts requires the existence of a sufficient number of shared instances. In our case, we thus need books that can be found in two libraries and are described using the two vocabularies to match. This is not a very common situation. However, as instances (in our cases, books) have their own information, such as authors, titles, etc.; it is possible to calculate the similarity between them. Our assumption is that similar books are likely to be annotated with similar subject headings, no matter they are from different collections or described in different languages. When we judge two books from different collections to be similar, we enrich each of them with the concept annotation of the other one, leading to a “virtually doubly annotated” book, which can be exploited by the simple technique of 2.3.1.

The instance matching based method first compares books from both collections, using their metadata records as proxies for them. For each book from Collection A, i_a , there is a most similar book from Collection B, i_b . We then consider that i_a shares the same subject headings as i_b does. In other words, i_a is now an instance of all subject headings which i_b is annotated with. This matching procedure is carried out on both directions. This way, we can again use measures of the subject headings’ shared extensions, such as the one defined in the previous section, even if these extensions are the result of artificial enrichment.

From a bird’s eye view, the algorithm consists thus of two steps:

- 1) Enrich the instances (documents) of one collection with the annotations of the most similar instance(s) of the other collection
- 2) Match the two vocabularies by using a co-occurrence based similarity measure, in our case we re-use the corrected Jaccard overlap coefficient presented in the previous section.

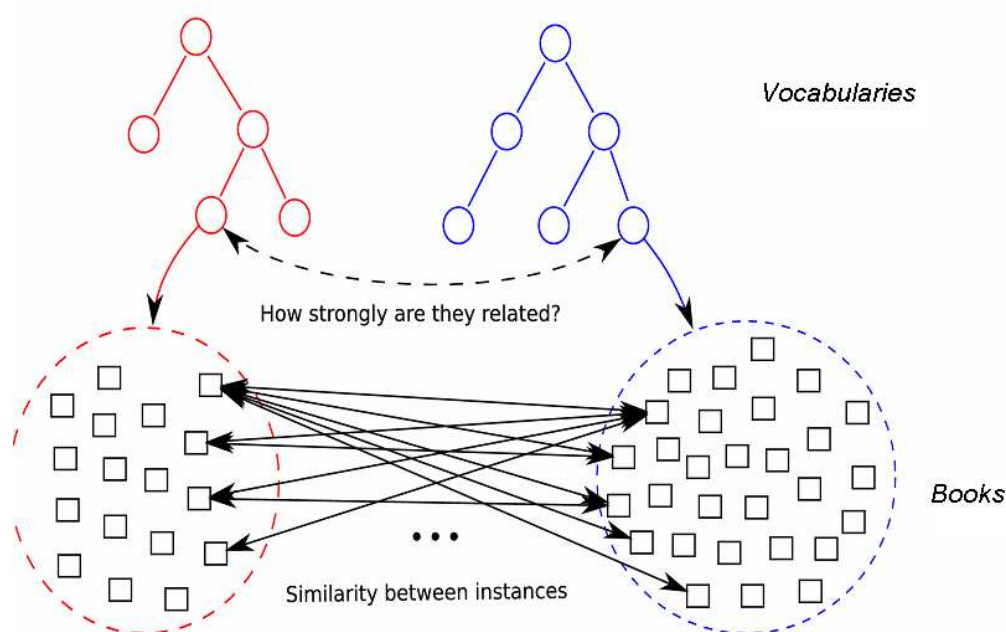


Fig. 2. Instance-based method using instance similarity

There are various solutions to match instances. One is to consider instances as documents (with all their metadata as their feature) and use information retrieval techniques to retrieve similar instances (documents). We base our approach on the *TF-IDF* (Term Frequency – Inverse Document Frequency) weighting scheme which is often exploited in the vector space model for information retrieval and text mining [SM83]. Our implementation is similar to what an information retrieval tool such as Lucene¹ performs to measure similarity between queries and documents—an important difference being that the “query” is the document to be matched.

To apply such a method in a multilingual context, translation is necessary. We follow a naive approach, using the Google Translate service to translate book metadata, including subject labels. The translation was done offline on a word-to-word level. We created a list of all unique words in each book collection. Batches of words were sent via an API to the Google Translate service, which proved to be relatively scalable (still, the translation of all words present in a corpus takes several hours). Every word for which we obtained a translation was then stored in a translation table. During the instance matching process, we translate every word of that instance by looking it up in the translation table and replacing it with the translation if available. We then calculate book similarity scores within a same language, stem the words in metadata records. We consider a local approach to the computation of IDF, reflecting the different information loads of terms in each metadata corpus.

Once we obtain book similarity scores, we can compute for each book the most similar one in the other collection. We then add to this book the concepts found in the indexing of that most similar book. Thus, we obtain a collection of virtually doubly indexed books, on which we can use the simple extensional concept similarity measure described in 3.1.

¹ <http://lucene.apache.org/>

Note that there were a number of choices we had to make when implementing the instance matching process:

- normalizing or not the words of the corpus before starting the similarity computation, e.g., by stemming them (e.g., ‘make’, ‘makes’, ‘making’ all become: ‘mak’)
- filtering some metadata fields which have little value for recognizing book with potentially similar subject, such as the ones that indicate the topmost library collection the books belong to, or numerical values like page numbers;
- comparing records grouping all metadata field content into one single “bag of words”, vs. performing field-by-field comparison;
- filtering most commonly used words, or occurrence of words which are deemed to be less informative, to make the matching process faster or reduce the memory space it uses;
- considering one global set of metadata for computing the inverse document frequency (IDF) of terms, or computing IDFs based on each collections’ metadata corpus;
- matching a given book to its most similar counterpart in the other collection, taking the N most similar books, or taking all books for which the similarity is above a given threshold;
- etc.

For more details, the reader is referred to [S09].

2.3.3 Data pre-processing and experiment settings

Records for books annotated using LCSH, RAMEAU and SWD are gathered respectively from the British Library (BL), the French National Library (BnF) and the German National Library (DNB).

In order to apply instance-based matching methods, the link between each book record and its subjects, must be properly identified. However, except for the German collection, the metadata records do not refer to the identifiers of the concepts from the controlled vocabularies. Instead, librarians often use simple string annotations. This introduces many issues, for example, using the lower case version of a concept’s label that is capitalized. Also, the preferred label is in principle used to refer unambiguously to those subjects. But sometimes librarians have been using alternative labels when indexing. Or as a result of vocabulary maintenance, a preferred label will become an alternative one, and corresponding book descriptions will not be updated... Those issues have to be addressed by using a string look-up and matching procedure to identify the book-concept links, using the information found in the SKOS representations of the vocabularies.

Another problem is that a single field can be used for controlled indexing and uncontrolled one, or for controlled indexing with different vocabularies—in the British collection records, subjects from the MESH vocabulary appear in the same field as LCSH ones, only field indicators distinguish the two.

Furthermore, it is necessary to tackle the *pre-coordination* issue. Librarians often combine concepts into complex subjects when annotating books, e.g., France--History--13th century. Some of these combinations are so often used that they are included into the

subject vocabulary later, while some are generated only at annotation time. In our data pre-processing step, we adopted the following strategy: if the subject combination cannot be recognised as an existing concept, it is then separated into single (existing) concepts, and the book is considered to be annotated by each of these concepts. We are well aware that this choice is not neutral. Hereby, a concept's extension, beyond the instances simply annotated by it, also contains the instances indexed with a compound subject that includes it, if this combination is not an existing concept in the vocabulary. However, it also brings more instances for concepts, which is very important for the statistical validity of the instance-based methods we employ here. Indeed this is made even more important by the low number of annotations we identified from the collections. Not every concept is used in the collections we have, cf. Tab. 4 and Tab. 5. This issue, which is mostly caused by the SHLs being designed and used for collections beyond the ones we got, will cause a mapping coverage problem for the instance-based methods, which we see in the next section.

Vocabulary	Number of concepts	Concepts used in collection
LCSH	339,612	138,785
RAMEAU	154,974	87,722
SWD	805,017	209,666

Tab. 4. Size of SHLs and number of concepts used to annotate books in collections

Collections	Total records	Rec. with valid subject annotations	Individual book-concept links
English	8,430,994	2,448,050	6,250,859
French	8,333,000	1,457,143	4,073,040
German	2,385,912	1,364,287	4,258,106

Tab. 5. Three collections and identified records with valid subject annotations (i.e., there is at least one link between these books and one SHL concept)

Another related, important decision we made is to restrict ourselves to match only individual concepts, or combinations that are reified in the vocabulary files. This drastically reduces the problem space, while keeping it focused on the arguably most important part of the potential book subjects. In fact this is rather in line with what is done in MACS, where very few mappings (up to 3.8% for SWD-concepts mappings) involve coordinations of concepts that are not already in the vocabularies.

A last pre-processing step is identifying the common books in two collections. The ISBN number is a unique identifier of one book. By comparing the ISBNs in both collections, we found three dually annotated datasets between the three pairs of SHLs, as shown in Tab. 6. The number of common books is extremely small compared to the size of collections. This is not unexpected, but causes a serious problem of concept coverage for the simple instance-based method that completely relies on these common instances.

Collection pair	Common books
French-English	182,460
German-English	83,786
German-French	63,340

Tab. 6. Common books between different collections

2.3.4 Results

When applied to the three pairs of vocabularies and their associated collections, the two instance-based methods give raw results summed up in Tab. 7. Note that the list of original similarity assessments is longer: any concept which has a book in common with another one will cause a proposed mapping to appear in the list, even if the similarity computed is extremely low. We kept mappings with a concept similarity measure above a 0.001 threshold. As we see in Section 4, the aim is then to find a correct filter, selecting the best mappings either based on their ranks or their raw confidence measure.

Instance-based technique	LCSH-RAMEAU	RAMEAU-SWD	LCSH-SWD
Simple instance-based	1,794,722	1,745,474	2,353,079
Using instance similarity	1,817,450	1,758,939	2,378,344

Tab. 7. Common books between different collections

3 Evaluation of automatic alignments

In the previous section, we have presented the four matching techniques that we have used to align RAMEAU, LCSH and SWD:

- lexical without translation,
- lexical with translation,
- extensional using real dually annotated books (hereafter called “real dual”),
- extensional based on instance matching and enrichment

In this section, we detail how we evaluated the results we obtained using these different techniques. First, we did a raw comparison with the existing manual mappings created with MACS, to have a first idea of the precision of the automatic alignment, and assess how well those automatic alignments can reproduce what is obtained manually. Second, to assess the quality of those automatic mappings that are not judgeable using the MACS material, we perform a manual evaluation on samples of the automatic alignments.

3.1 Comparing with MACS manual mappings

3.1.1 Method

Our first evaluation approach uses the MACS manual mappings as reference mappings.¹ Tab. 8 gives the concept coverage of mappings between each pair of vocabularies:

Vocabulary pair	Number of mappings	Concepts involved
LCSH-RAMEAU	57,663	16.4% of LCSH and 36.1% of RAMEAU
RAMEAU-SWD	13,420	7.8% of RAMEAU and 1.6% of SWD
LCSH-SWD	12,031	3.2% of LCSH and 1.4% of SWD

Tab. 8. Simple statistics of MACS manual mappings and concepts involved

Obviously, there is a serious lack in terms of concept coverage if using MACS as a gold standard, as MACS is still work in progress. For example, only 12.7% of LCSH concepts and 27.0% of RAMEAU concepts are both involved in MACS mappings and used to annotate books in the collections we gathered. The situation is worse for the other two pairs of thesauri, where only 1 to 3% concepts are both considered by MACS and used to annotate books.

That can raise issue if one counts as false all automatic mappings that are not confirmed by MACS. To perform a relatively fair evaluation on our matchers' accuracy, we separated the generated mappings as *judgeable* and *non-judgeable*. A mapping is judgeable if at least one concept of the pair is involved in a MACS manual mapping, otherwise, it is not judgeable – that is, no data in MACS allows us to say whether the mapping is correct or not. We measure precision as the proportion of the correct mappings over all generated and judgeable mappings.

To measure the completeness of the found alignments, we would need to compute recall, that is, the proportion of the correct mappings over all possible correct mappings. Unfortunately it is very difficult to get all correct mappings in practice. Manual alignment efforts are time consuming and result in a limited number of mappings if the two vocabularies to align are big. Despite the lack in concept coverage for MACS, we decided that these manual mappings were still useful to exploit. Indeed, measuring how well we can reproduce manual mappings with the help of automatic tools is valuable per se. As a proxy for completeness, we thus measure the coverage of MACS, that is, the proportion of MACS mappings we find in the automatically produced alignments.

As mentioned in Section 2, our matchers return candidate mappings with a confidence value based on lexical considerations or the extensional overlap of two concepts. This allows us to rank the mappings, and, moving from the top of the ranked list, to measure precision and coverage up to certain ranks.

¹ Those mappings were communicated by the MACS team at several dates. The latest shipment, with which the results presented here were obtained, was sent in April 2009.

3.1.2 Results

Fig. 3 (a and b) gives the performance of the four different alignment methods for matching LCSH and RAMEAU. Here the x-axis is the global rank of those mappings—by “global ranking,” we take the non-judgeable mappings into account; however, they are not considered when computing precision. Note that our lexical mapper provides three confidence levels. Mappings with the same value are given the same rank; they are therefore measured together.

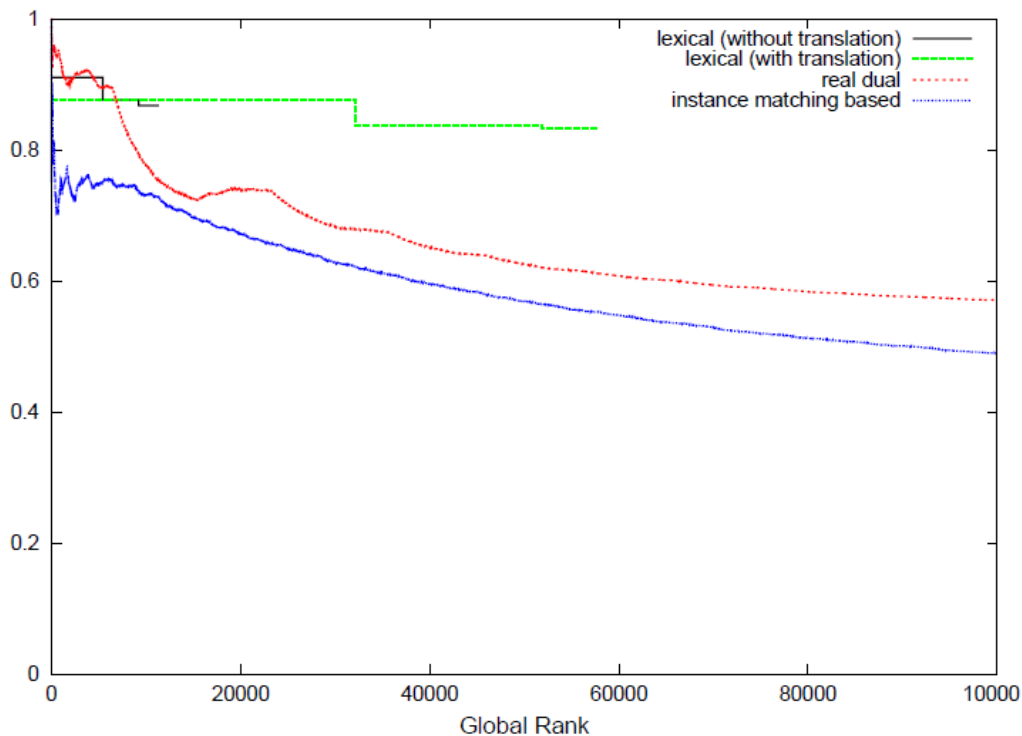


Fig. 3(a). Performance for matching LCSH and RAMEAU: Precision

The lexical method applied on non-translated LCSH and RAMEAU gives a very limited number of mappings: in total, 86% of these mappings are in MACS, but they only represent 13% of the MACS mappings. By naively using Google Translate, the automated lexical method already recovers 56% of MACS mappings, while the precision decreases by 3%. The main reason for this precision decrease is that the translation is not perfect nor stable. For example, in SWD the German name *Aachen* occurs in several subject headings. However, it was sometimes (rightly) translated to the French name *Aix-la-Chapelle* and in other cases it was not translated at all.

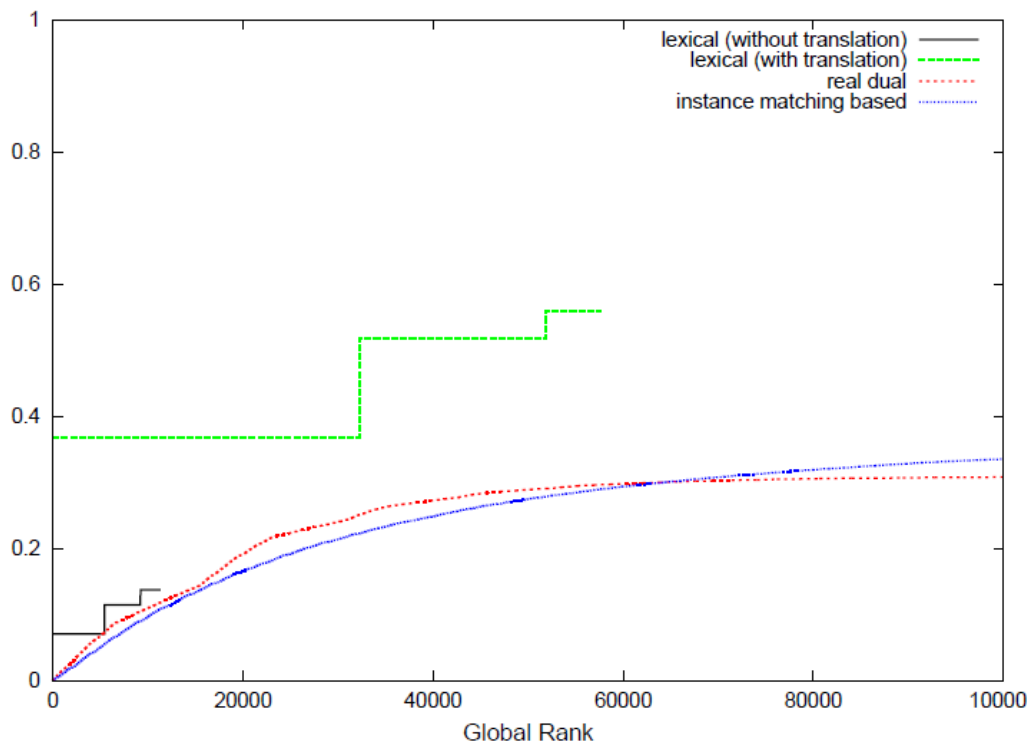


Fig. 3(b). Performance for matching LCSH and RAMEAU: Coverage

From Fig. 3, the precision and coverage of the first 7K instance-based mappings generated from the real dually annotated dataset (1% of total book records in two collections) are similar to the lexical method on non-translated thesauri. Gradually, the precision decreases and the coverage increases; and both level after approximately 60K mappings.

The sheer number of instances (cf. Sec. 2.3.3) inevitably influences the performance of this method. Another possible reason is that instance-based methods focus on the extensional semantics of those concepts, i.e., how they are used in reality. Some mapped concepts are not really equivalent from the point of view of their intension, but they are used to annotate the same books in two collections. For example, according to MACS, the RAMEAU concept *Cavitation* is mapped to the LCSH concept *Cavitation*; however, our instance-based method maps it to another LCSH concept, *Hydraulic machinery*, because they both annotate the same books. Such mappings can therefore be very useful in the query reformulation or search applications, and of course would require further evaluation. This also confirms that the intensional and the extensional semantics (i.e., the meaning of a concept attached by people and its actual use in the libraries) may differ significantly.

The method based on instance matching performed worse. The loss of nearly 10% in precision can have two reasons: 1) the translation of book information is not good enough; 2) the similarity between books is calculated purely based on weighted common words, where we ignore the semantic distinction between different metadata fields, which could potentially help to identify similar books. Meanwhile, by matching similar instances on top of using real dually annotated books, we gradually include new concepts, which increases coverage.

3.2 Sample manual evaluation

3.2.1 Motivation and method

As introduced earlier, not every mapping can be evaluated using the MACS comparison method presented above, as neither of their concepts is considered in MACS before. For example, up to rank 50K, 29% of MACS mappings (16,644) between LCSH and RAMEAU are found, and the precision is 63%. However, only less than 26K mappings were actually judgeable.

Fig. 4 compares the distribution of different kinds of mappings, where the shaded area shows the number of non-judgeable mappings. The coverage issue for the SWD-related cases is more serious, even for lexical mappings, as shown in Fig. 5. Among those non-judgeable mappings, we expect to find valid ones, given the precision of the judgeable mappings that are found at neighbouring ranks.

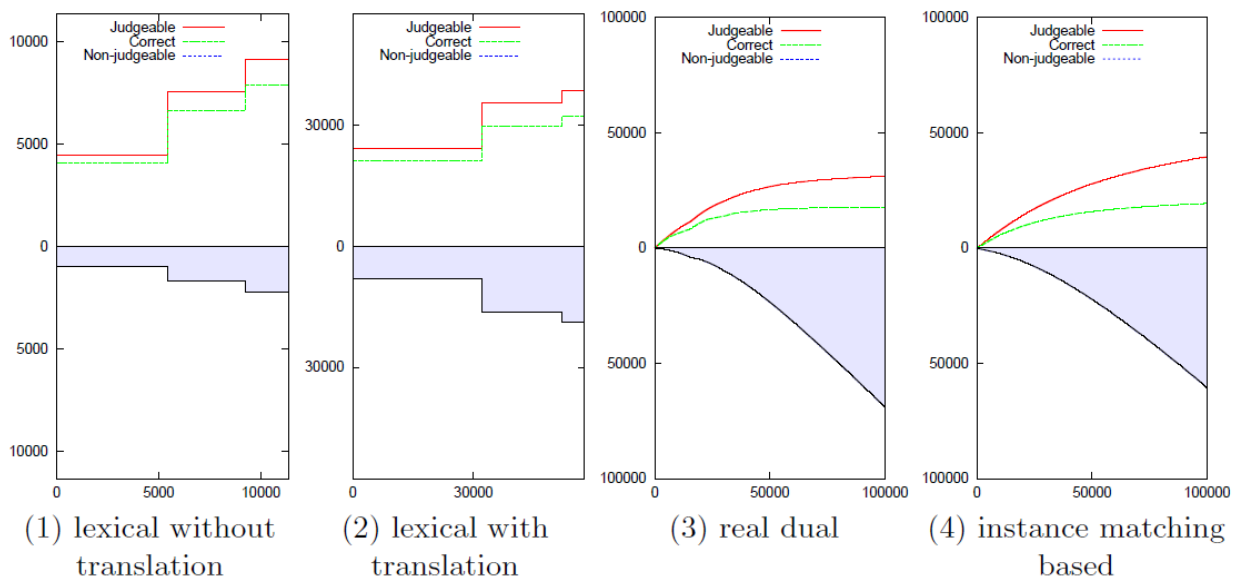


Fig. 4. The distribution of mappings generated by different methods: LCSH-RAMEAU

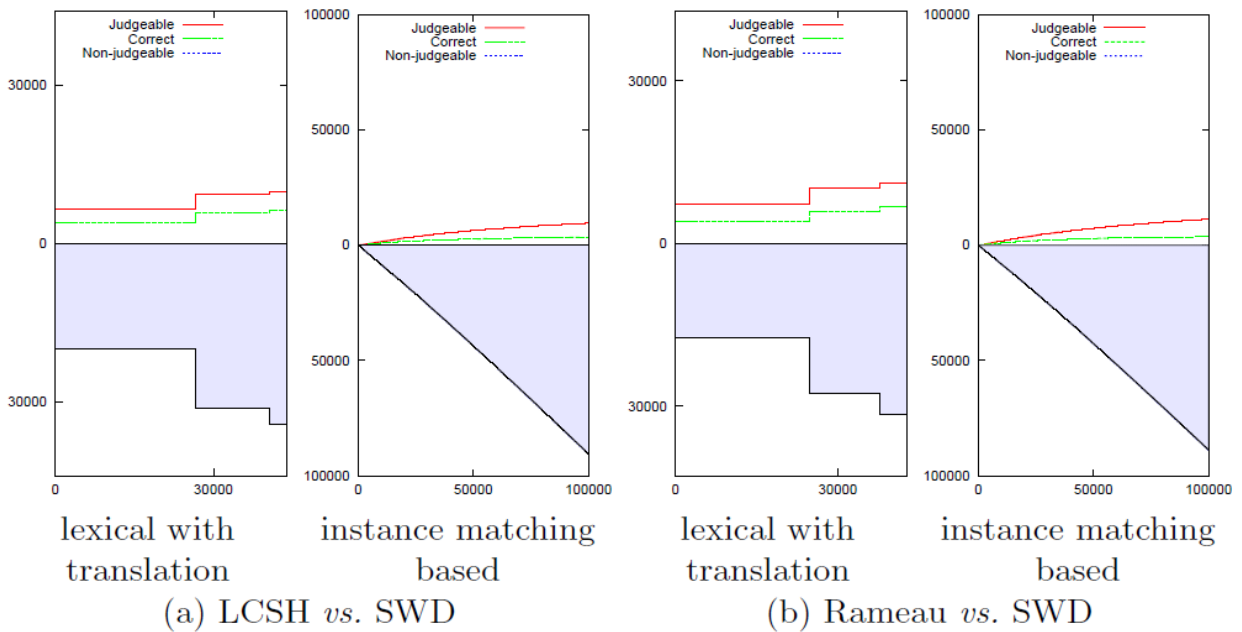


Fig. 5. Coverage issue for the LCSH-SWD and RAMEAU-SWD cases

We thus carried out manual evaluation of samples of non-judgeable mapping. For lexical mappings, we took 50 random mappings within each of the three confidence levels. For instance-based mappings, we first ranked them based on their confidence values, and then chose every 10th mapping among the first 1000 mappings, every 100th mapping from 1000 to 10,000 mappings, and every 1000th mappings from 10,000 to 100,000 mappings. In all cases, we kept for manual evaluation only the mappings that are not judgeable according to MACS. Depending on the actual sample size, the corresponding error bar was also calculated.

3.2.2 Results

Fig. 6 (a, b, and c) shows the precision of the manual evaluation proportionally combined with that from comparing with MACS reference mappings. For the LCSH-RAMEAU case, the precision, which is consistent with Fig 3(a), indicates that our methods indeed provide a large number of valid mappings. More importantly, they provide mappings that complement the MACS manual mappings. For the LCSH-SWD and REAMU-SWD cases, the global precision is also comparable with the precision measured using MACS alone. It also confirms that all methods perform worse in these two cases, which we can relate to the fact that LCSH and RAMEAU headings are quite similar in the way they are designed and used, and less similar to SWD. Finally, the performances of the two instance-based methods cannot be distinguished anymore. Yet, due to the small sample size, it is impossible to say whether this is caused by statistical uncertainty, or by that fact that the method using instance matching may suffer less from a very small overlap of collections. This aspect should be investigated in the future.

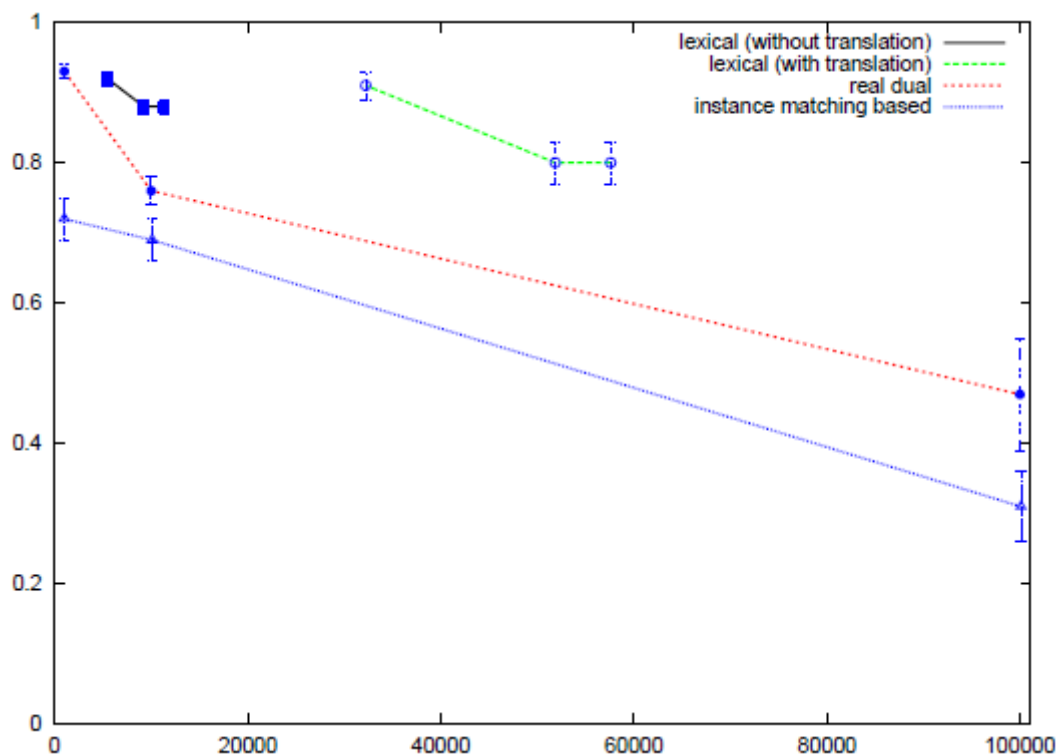


Fig. 6(a). Overall precision combining MACS and manual evaluation: LCSH-RAMEAU

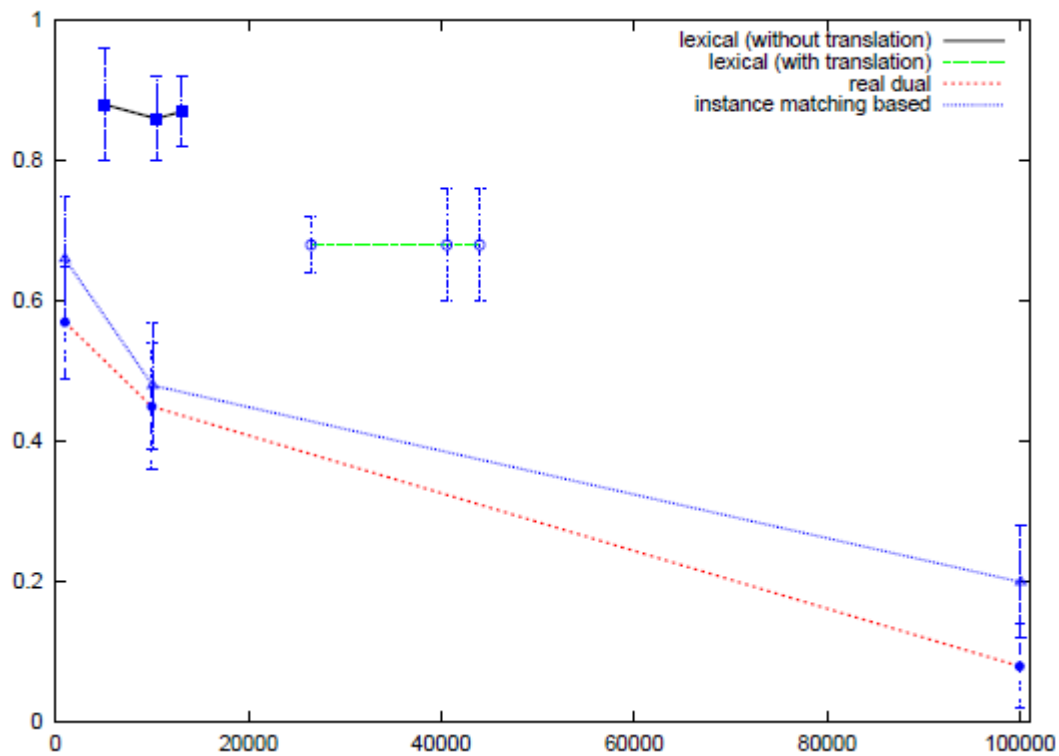


Fig. 6(b). Overall precision combining MACS and manual evaluation: LCSH-SWD

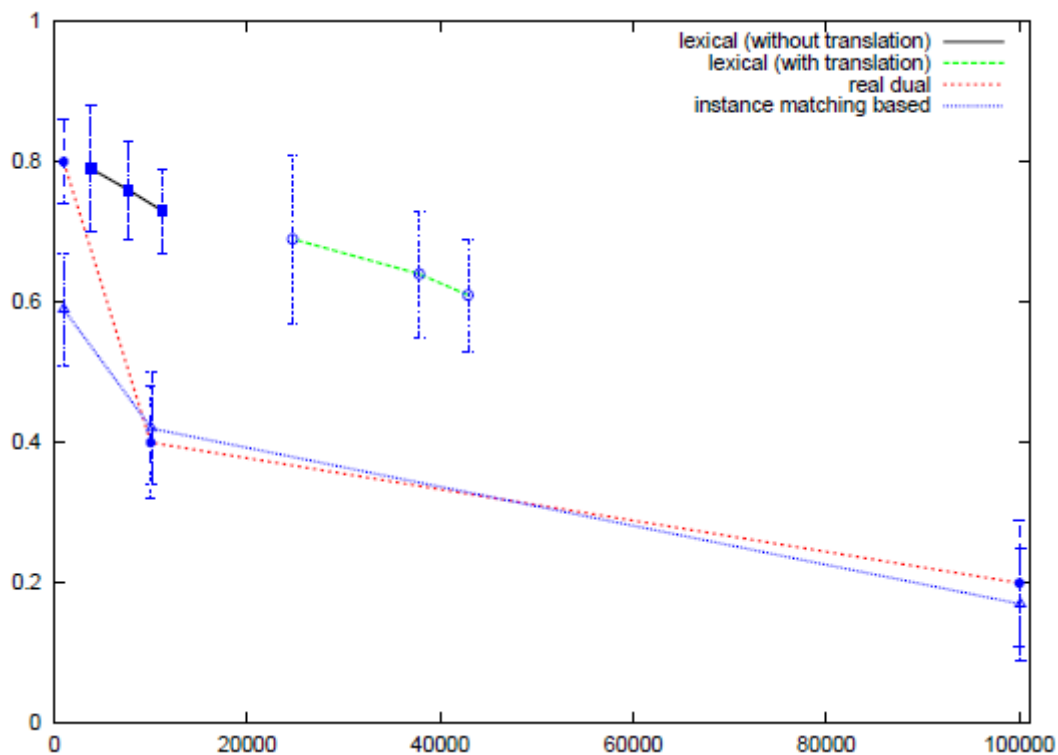


Fig. 6(c). Overall precision combining MACS and manual evaluation: RAMEAU-SWD

4 Feasibility assessment of automatic alignments

As said in the introduction to this report, it is not expected that automation of alignment processes is entirely reachable. In fact, considering the intrinsic difficulty of the alignment case at hand for existing matching techniques, we already had to focus on simple, pragmatic alignment methods. In the previous sections, we have respectively presented the alignment strategies that we have implemented, and an evaluation of their results.

The aim of the present section is to assess the feasibility and costs of the chosen methodological and technical approaches, especially focusing their practical ease of deployment as well as on their extendibility. We discuss each of the techniques we followed:

- lexical matching without translation,
- lexical matching with translation,
- extensional matching using real dually annotated books,
- extensional matching based on instance matching and enrichment

Readers must be aware that this report is not intended at a definitive guide on vocabulary matching. It is a rather collection of informal statements and lessons learned, from the perspective of researchers who had to find and implement practical solutions to a complex problem with limited resources.

4.1 Lexical matching without translation

4.1.1 Cost of implementation and running

Basic lexical matching was fairly easy to specify and implement. Having vocabularies available using a common format and model (in our case, RDF and the SKOS model) was especially useful. This makes accessing and exploiting different kinds of labels easy. In particular, articulating the comparison of different features (preferred labels, alternative labels) with different comparison options (direct comparison, using lemmatized or split word forms) was straightforward. Furthermore, there are very few parameters to tune in the script that we created, making it quite easy to run. On our server, matching each pair of vocabularies took at most an hour, depending on the size of the vocabulary, and including the time to pre-load the external lexical resources that we used.

It is important to mention using different types of lexical knowledge is by far the most costly component to implement in our process. Lexical databases can be fairly complex. This is especially true for the CELEX database, which requires to understanding an important part of the whole database's structure to access links between word forms and lemmas. Such effort must be reproduced for every language the matcher has to be applied with. Even though CELEX provides with lexical information for English and German, the structure of the database was slightly different for the two languages. And for French, we had to go for another dictionary, DELA,¹ which has a different structure.

4.1.2 Applicability to the specific TELplus case

The three vocabularies we have to match in our case are provided with relatively extensive lexical information, as shown in Tab. 9:

Vocabulary	Number of concepts	Number of alternative labels	Number of labels per concept
LCSH	339,612	309,653	1.91
RAMEAU	154,974	200,811	2.30
SWD	805,017	1,078,299	2.34

Tab. 9. Lexical information in the three SHLs²

Without translation, the simple lexical matcher can only perform well on concepts which happen to have the same lexicalization in different languages. We found however that more such concepts than expected were present in our three vocabularies. Especially, for places, persons, or very specific topics that are provided with language-independent names (such as Latin names for biological species), interesting mappings can be found. Still, the coverage of

¹ <http://infoling.univ-mlv.fr/DonneesLinguistiques/Dictionnaires/telechargement.html>

² Note that each concept is provided with exactly one *preferred* label in the language of the vocabulary, in addition to the *alternative* ones. The number of preferred labels is thus equal to the number of concepts.

those cases is not wide: slightly more than 10K mappings can be found for each pair of vocabularies (cf. Sec. 2.2.1).

4.1.3 Genericity & extendability – potential applications to other cases

The simple lexical approach can be applied to all cases where vocabularies are in the same language, or a large set of concepts with common lexicalizations in different languages can be expected (see above). Additionally, the naming strategies must be consistent. If for instance complex labels are permuted, more complex techniques are required. External lexical knowledge base can also be sought if the set of synonyms provided in the vocabularies to palliate for such lexical variations is not rich enough. In general, using simple lexical methods is expected to lead to better results in more constrained domains, like science, as opposed to more common-use topics.

Note that from the perspective of lexical processing, the use of lemmatization, for which we had convenient solutions at hand, can be replaced by turning to more common stemming tools. Stemming can be less precise, but robust implementations are available for more languages, and many come free of charge, as in the Lucene platform.¹

4.2 Lexical matching with translation

4.2.1 Cost of implementation and running

Our translation-enabled matching process is based on the same matching tool as the basic lexical matching. The main difference is that a vocabulary is translated into the language of another vocabulary prior to being matched to it.

As explained in Section 2.3, we used the Google Translate service. This translation step is straightforward to implement, as Google's API is fairly simple and remains the same from one couple of languages to the other. It requires however a large amount of time—approximately one day—to run, given the high number of concept labels to translate (cf. Tab. 3). Also, calling Google's web service turned out to be not fully reliable—some concept labels not being assigned a translation as a result of time outs. We thus had to do several runs of service calls in order to ensure that every translation query had been given an answer.

4.2.2 Applicability to the specific TELplus case

As shown in Tab. 3, our vocabularies are provided with relatively extensive lexical information (preferred labels and alternative labels). Translating allows finding many more mappings than with the basic lexical approach. As mentioned in Sec. 2.3, matching with translation produces between 40K and 60K mappings, depending on the vocabulary pair. The results significantly overlap with the MACS ones, hinting that this approach may be used for helping manual matching. However, it still falls short with respect to coverage of the vocabularies.

¹ <http://lucene.apache.org/>

Interestingly, automatic translation gives results of good precision for the English-French case, but precision drops when the German language is involved. We can hypothesize that this is related to the design of the German headings, which differ from English and French ones, while French and English headings follow relatively close design rules [L99]. In addition, Google Translate may have failed to appropriately translate some compound German words.

In fact, one must keep in mind that such online translation service based on text corpus statistics is of course not a perfect solution, and that the quality varies from one language pair to the other.¹ Also, a same word in different labels (and thus different lexical contexts) can lead to different results. For example, in SWD the German city name Aachen occurs in several subject headings. We found out that it had been sometimes (rightly) translated to the French Aix-la-Chapelle, but that in other cases it had not been translated at all.

4.2.3 Genericity & extendability – potential applications to other cases

Our naïve translation approach is of course easily deployable for all languages managed by the translation service we used. Plugging in other translations is very simple from a technical perspective. It may however provide results of lower quality, depending on the language, for the reasons mentioned in the above section.

We are well aware that other, more sophisticated approaches exist to tackle the translation issue. But deploying them requires thorough expertise in the use of linguistic resources, which makes it more difficult to assess their practical feasibility with limited means. Also, a number of industry translation options exist. But they come at a price, which may hamper their use in a project such as TEL. Adopting them would also require to take into account business-level criteria, which falls beyond the (technical) scope of our present experiment.

4.3 Extensional matching using real dually annotated books

4.3.1 Cost of implementation and running

The basic extensional matching method is in principle extremely simple to implement and deploy. It requires to store, for each concept used as a subject of book collection, a link to its extension, namely the books that have that concept as a subject. Similarity assessment can then be done by taking two concepts and using the formula shown in Sec. 2.3, with the extensions of these two concepts as input.

The most demanding task was in fact to obtain the metadata sets for books, and pre-processing it to feed the (ISBN-based) extension comparison process. For deploying the instance-based matching method, two elements are indeed crucial:

- finding identifiers of books that are present in different collections—this is done by retrieving common ISBNs or ISSNs;

¹ In fact Google Translate's historical focus on UN official languages (http://en.wikipedia.org/wiki/Google_Translate, accessed November 2009) may also explain the discrepancy with the English-French case.

- finding occurrences of concepts from the SHL SKOS conversions inside subject annotations.

Regarding both aspects, consistency issues occurred. First, ISBNs can be wrongly documented. Sometimes they are incompletely typed, or typed differently (e.g., with or without spaces in-between groups of digits). They can also be presented in different metadata fields, even when the metadata of two collections are structured according to a same schema. The British and French records, as we got them, mostly use the Dublin Core metadata fields, while DnB metadata comes in the MABxml format. But in the French records the ISBN is encoded in the `dc:description` field, while `dc:identifier` is used in the English ones.

Second, retrieving an exploitable link between books and vocabulary concepts is not trivial, as seen in Sec. 2.3.3. All the problems mentioned there forced us to implement rather ad-hoc solutions. That could have been avoided if the link between books and concept had been clearly represented in the data exports we got.

4.3.2 Applicability to the specific TELplus case

Our evaluation (see Sec. 3) shows that considering the usage strategy for book concepts can bring interesting results, complementing the manually obtained mapping. However, as just mentioned, applying that method to the TELplus case is not easy.

As regards the identification of common books across collections based on ISBNs, the problem is dual: as seen, the description of ISBNs in the metadata records may be incomplete, or impossible to get. ISBNs may also not exist, especially for older books.

Further, there is an important indexing coverage issue. The book collections we have to deal with are huge; and subject indexing has not been done for all books. It usually covers a specific part of the collection, e.g., starting from 1986 for the German collection. The number of books available for instance-based matching is thus lower than the number of books in the whole collections, cf. Tab. 5. Also, the SHLs at hand are designed for use in several collections: some concepts may thus be used outside the collections we have. As a result, not every concept is used in the metadata we got, cf. Tab. 4. And a lot of concepts are very rarely used. This causes an important data sparseness problem.

Due to the aforementioned subject indexing and ISBN coverage issues, the corpus of common books obtained using ISBNs matching is thus quite small compared to the original collections' sizes. And that corpus covers only a small part of the vocabularies to align, cf. Tab. 6. For example, in the French and English collections, there are 128,423 pairs of book record with the same ISBN, which involve 33,391 RAMEAU concepts and 43,863 LCSH concepts. This low numbers are not unexpected, but cause a serious problem of concept coverage for the simple instance-based method that completely relies on common instances.

To palliate this issue, we tried to exploit not only identical ISBNs, but ISBNs that correspond to books that are *related*, resulting for example from translation or re-publishing. ThingISBN¹

¹ http://www.librarything.com/thingology/2006/06/introducing-thingisbn_14.php

from the LibraryThing community site provides with such links. Yet, this community effort is still much work in progress. In our case, the additional number of ISBNs in one collection having related ISBNs in another collection is small: we gathered less than 8,000 additional links for the three collections. We thus did not include them in our process. In the near future, this data may however be more interesting to exploit. Depending on business opportunities, the xISBN service from OCLC¹ may also be exploited, as a more complete but paying alternative.

4.3.3 Genericity & extendability – potential applications to other cases

We can only apply basic instance-based matching approaches when the concepts to be matched are used in collections that share a substantial number of elements—enough to give sufficient statistical ground to overlap measures. This is often not the case in Cultural heritage collections. Libraries may constitute an exception, as a same book can almost always be found in several libraries. Yet, in the case of libraries from different countries, the overlap between collections will be smaller, and one should seek books that are related to each other, instead of books that are exactly the same.

Additionally, our experiment has shown that some tedious data ingestion and pre-processing has to be carried out so as to recognize those identical or related books. This is likely to be also true for collections coming from libraries other than the ones under scrutiny in our experiment. There is no hint that the English, French and German national libraries are less careful than other libraries when creating and exporting their metadata.

A final remark, our approach considers only *one-to-one* mappings between subjects that already exist as such in the vocabularies. To cope with that initial choice, we had to ignore some pre-coordination cases at the book level. In a more general effort, it may be more fruitful to adapt the process so as to detect mappings between combinations of concepts (*many-to-many* mappings). This requires changing the formula presented in the Sec. 2.3, as we have done in [IKMW09].

4.4 Extensional matching based on instance matching and enrichment

4.4.1 Cost of implementation and running

In the state-of-the art, very few efforts exist that employ document similarity assessments (inspired from Information Retrieval practice) to derive similarity between the subjects those documents are about (a matter rather focused on in the Semantic Web and Knowledge Organization System research communities). Obtaining a satisfactory procedure for instance matching-based extensional matching required thus a lot of efforts, both from the research and implementation perspectives (see Sec. 2.3.2 and [S09]). Altogether, maybe one person-year has been devoted to the whole effort of implementing this approach in the TELplus case.

¹ <http://www.worldcat.org/affiliate/webservices/xisbn>

Additionally to running the tool, appropriate data must be obtained and pre-processed, even though, in the specific case at hand, this work was common with the work required for the basic extensional technique. Also, the tuning of the method is far from trivial. As hinted in Sec. 2.3.2, there were many parameter setting choices possible, which required extensive testing.

Finally, it is important to mention that instance matching is a complex task, which requires a lot of computing power. This is especially true when the collections of documents to match are large, which was of course the case here. Tab. 7 shows the performance of the tool when applied to collection samples of different sizes, for the indexing step,¹ which is the most time-consuming one. Together with the translation of book metadata, the computation of each book's most similar book in the other collection, the enrichment of books based on that matching, and the computation of the overlap measures based on the enrichment, the whole process can take more than a dozen hours.

Number of indexed books	time to enrich 100K instance (hh:mm)	memory usage
524K	0:17	1,294 MB
2,506K	0:32	7,279 MB

Tab. 7. Run-time performance of the indexing step

4.4.2 Applicability to the specific TELplus case

Our evaluation (see Sec. 3) shows that the basic extensional method performs slightly better than the instance matching-based method for the mappings in the first ranks. However, the precision quickly becomes similar as we go further down the ranked mappings. The more complex but more widely usable instance matching-based approach can thus be equally applied here to provide with mappings complementary to the manual ones.

One may hypothesize that in the TELplus context, it would be very interesting to further test this specific method using the full-text data available from book OCRing, as performed in WP1. Using metadata, as we did, enables to leverage the higher level of control of the descriptions exploited. But full text may be more amenable to detecting similarity between books from different collections using information retrieval approaches. The basic translation we have used may also perform better when larger chunks of text are being processed. When translating a given (group of) word(s), complete sentences give more appropriate data for automatic translation methods than metadata records in which values cannot be considered to be “real” text. However, the timing of the project did not allow carrying out such experiments. Further work on that issue is certainly required.

4.4.3 Genericity & extendability – potential applications to other cases

The main benefit of instance matching-based extensional techniques over basic extensional techniques is that the former can be used even when the book collections available are

¹ In information retrieval approaches, indexing is the step that allows creating representations of documents (in our case, as vectors in a vector space) that are amenable to similarity computation.

disjoint, or when the overlap is very small. They are in fact applicable whenever there is data for the objects connected to the concepts to be matched.

Of course there must be enough such objects, as well as enough data to provide sufficient material for the information-retrieval techniques we use. Also, the process we implemented may have to be adapted to the case of other collections. That can be time-consuming, given the number of parameters available. However, we observed in our various experiments that the various settings show minor differences wrt. the quality of their results. And, importantly, the best results are almost always obtained using a simple baseline configuration. The need for tedious parameter setting may thus be minimal. In fact, we have already re-applied our tools to another communication science case [WSTA09].

4.5 Concluding remarks

Implementing automatic matching techniques that fit a case such as the TELplus one is a costly exercise. Automatic vocabulary matching is still an active research issue, and is likely to remain so for the coming years, given the number and hardness of the challenges at hand. Specific efforts are thus required to adapt existing techniques, or to implement them from scratch.

We naturally hope that our experiments have helped to make these issues clearer, especially regarding instance matching-based extensional techniques. This specific approach had not been explored so far in cases such as ours. We hope that the results we obtained and the lessons we learnt will help future practitioners' efforts.

It is also worth noting that all automatic matching efforts greatly benefit from vocabulary and metadata sources' being available in convenient standard exchange formats. Regarding the vocabularies to match, creating RDF/SKOS datasets as part of WP3.2 was helpful to devise matching processes that are as generic as possible. These could indeed be seamlessly applied to the three matching cases we have. The situation was a bit different for the book metadata. All metadata came as XML, but the German set was using a different schema from the French and English ones. And while the French and English sets used the same Dublin Core schema, some subtle variations in the way fields are used required further efforts before exploiting them.

Finally, even though the basic lexical method was much less demanding, all automatic methods require computational resources, both in terms of computation time, memory and disk space usage. This definitively prohibits the use of such techniques at final application run-time. When the vocabularies to match and the associated book collections are so large, alignments have to be computed beforehand.

5 Usefulness assessment of automatic alignments

This section investigates what are the conditions and the cost of integrating automatically produced alignments into production processes, as well as the benefits users could expect from this integration. Two aspects are focused on: how the mappings can complement and

support a manual mapping process, and how they can be employed in the multilingual search tool reported upon in D3.4.

5.1 Usefulness of automatic alignments for manual matching

As said, manual alignments are costly, and benefiting from automatic input could help a lot. Especially, while the precision of manual mappings is good, their coverage is more worrying. It took years to build the MACS alignments, which are at the time of writing this report still not covering all concepts of LCSH, RAMEAU and SWD. The question we seek to answer in that sub-section is whether our automatic alignments can be of any help to a manual matching project like MACS.

Answering this question requires to consider the strategy used to produce automatic mappings. First, our evaluation (see Sec. 3.1.2) shows that automatic lexical mappings, with or without translation, tend to have high precision and to mirror quite well what has been produced manually. Such mappings may therefore be successfully used as direct input for the manual matching process. The effort required to validate them is expected to be small, especially considering that they are based on evidence (label comparison) which is easy to understand for a human validator.

The issue is that those alignments have a rather weak coverage. In fact, as MACS manual alignments already contain thousands of mappings, they can add only little on top of them. Applying those techniques seem therefore more valuable at the start of a matching project, when the low-hanging fruits have not yet been picked.

The two automatic alignment techniques that exploit the usage of concepts in collections arguably perform better with respect to coverage. They virtually allow finding a mapping for every concept that co-occurs (directly or by means of instance enrichment) with a concept from another vocabulary. Additionally, we have seen that to a great extent those mappings complement the MACS manual ones. They can thus prove very useful to augment the coverage of partial manual alignments.

However, this potentially higher coverage comes at the cost of precision. When the overlap between the extensions of concepts goes thinner, the precision level drops quite highly. This would hamper the work of a human operator in charge of validating such automatic mappings. An appropriate trade-off should therefore be determined between the extra coverage potentially provided by extensional alignments, and the extra effort required to correct them.

Luckily, our extensional approaches rank the mappings according to their confidence measure. This enables to fine-tune such a balance depending on the priorities of the mapping process at hand: it is possible to select mappings corresponding to a certain precision or coverage level. However, this requires some prior sample evaluation, so as to determine the expected quality level of the mappings belonging to a given rank interval.

Finally, a crucial assumption for extensional matching is that similar books are likely to be annotated with similar subject headings, no matter they are from different collections, or are

described in different languages. As seen in Sec. 3, this assumption is partly validated in our case: considering usage strategies for concepts brings interesting results that complement manual alignments. However, we can observe discrepancies between the indexing policies of the three libraries involved. This results in mappings that do not correspond to strict equivalence between the intuitive meanings of concepts, as shown in the *Cavitation* example on p. 15.

This raises issues when comparing with a manual approach such as MACS. MACS in fact also aims at taking into account the usage of concepts. However, that usage is quite difficult to grasp for a human operator without the use of statistic tools that are able to cope with the vastness of the collections involved. MACS alignments thus mostly contain mappings that mirror equivalence of the “intensional” meaning of the concepts involved, as represented by their labels.

Including extensional mappings in the loop may thus question the methodology of the whole alignment process, including the fundamental question of what is a *relevant* mapping. In fact our extensional mappings can be very useful in query reformation or search applications (“did you mean” suggestions, suggestions of related concepts) that intend to bridge across various indexing policies, as we will see below.

5.2 Usefulness of automatic alignments for multilingual search

Our D3.4 report presents how we have selected mappings from our four automatic alignments to include them in a multilingual search prototype, alongside the manually produced MACS mappings.

The main barrier to using automatically produced alignments in the context of an end-user oriented prototype is their quality. As we measured it (cf. Sec. 3), this quality can be low, especially for the alignments obtained using extensional techniques, which produce a huge number of correspondences with greatly varying confidence levels.

5.2.1 Selection of automatic alignments

As already mentioned, our automatic mappings are ranked according to their confidence levels, putting in first positions those who are expected to have better quality. This enables to easily proceed with a meaningful selection.

We opted for selecting mappings that correspond to an expected level of precision comprised between 50% and 60%, based on the results of our manual evaluation (cf. Sec. 3.2). This threshold, which may appear quite low, has two main motivations. First, as automatic mappings are sorted according to their expected quality, a search tool can still present results that correspond to good mappings ahead of the results obtained using less good mappings. This mirrors a strategy that is very common in search engines, such as Google, where users are first provided with results expected to be the best fitting ones, but have also access to a “long tail” of worse results that they can explore, should they not be satisfied by the first ones.

The second reason stems from the nature of the extensional mappings. Those mappings are based on concrete statistical evidence drawn from the actual usage of concepts in book description. Even if they do not always correspond to a strict semantic equivalence, most of them—at least in the first dozen thousands—denote a semantic relation which can be judged relevant by a user and thus bring an interesting form of serendipity. As a complementary example to the *Cavitation* case presented in the previous section, we can mention that the *Théâtre sanscrit* concept in RAMEAU is mapped to *Sanskrit drama--History and criticism* in LCSH. User interested in books indexed using the first subject may often be interested in books indexed using the second one. Extensional mappings can thus be very useful in the query reformation or search applications. The resulting serendipity effect may in fact be precious for compensating the effect of controlled indexing policies. Sometime those policies can result in discrepancies between what trained indexers consider to be the subject of a book and the looser connections end users accustomed to modern search engines will judge useful for their information need. This claim, even if reasonable at first sight, remains of course to be confirmed by further evaluation on the search prototype built: we could not carry out such evaluation due to lack of time.

The selection process resulted in the inclusion of following numbers of mappings in the tool. For the lexical alignment without translation, no selection has been made. The alignment contains:

- 11,345 LCSH-RAMEAU mappings;
- 13,119 LCSH-SWD mappings;
- 11,215 RAMEAU-SWD mappings.

No selection was made also for the lexical mappings with translation, as all mapping configurations are expected to be above our 60% threshold for precision. The alignment thus contains:

- 57,595 LCSH-RAMEAU mappings;
- 44,017 LCSH-SWD mappings;
- 42,869 RAMEAU-SWD mappings.

For this basic extensional alignment based on real dually indexed books, a selection has been made, resulting in:

- 60,000 LCSH-RAMEAU mappings (corresponding approximately to an expected 60% precision level)
- 5,500 LCSH-SWD mappings (corresponding approximately to an expected 50% precision level);
- 7,250 RAMEAU-SWD mappings (corresponding approximately to an expected 50% precision level).

Finally, for the extensional alignment using instance matching, we selected:

- 31,800 LCSH-RAMEAU mappings (corresponding approximately to an expected 60% precision level)
- 9,100 LCSH-SWD mappings (corresponding approximately to an expected 50% precision level);
- 5,500 RAMEAU-SWD mappings (corresponding approximately to an expected 50% precision level).

MACS manually created alignments contain approximately 130,000 trilingual mappings, which involve approximately 100,000 RAMEAU subjects, 90,000 LCSH subjects and 15,000

SWD subjects. Some of our automatic mappings are common with the MACS ones. However, based on the figures showing this overlap (Fig. 4 and Fig. 5), we can argue that the automatic alignments contribute a number of complementary mappings of the same order of magnitude, which is a significant achievement. The multilingual search tool is thus very likely to benefit from those alignments, as they will increase its coverage—i.e., the number of books it can return as a result of a given language-specific query.

5.2.2 Ingestion of the mappings in the prototype

From a technical perspective, the selection and ingestion of selected automatic mappings in the search prototype described in D3.4 was straightforward. All our matching tools indeed use a standard output format, namely the one used in the context of the Ontology Alignment Evaluation Initiative.¹ This greatly enhances interoperability, notably when it comes to ingest mappings in another tool, like our search prototype.

In fact, to comply with choices made in the early development of the prototype, we had to convert to the XML format used in the MACS approach. But this was easy, as the MACS format is geared towards the representation of similar mappings, and our SKOS representations for the vocabularies contain all the necessary information to generate MACS-compatible mappings—which in particular requires the preferred labels of mapped concepts to be included along with their identifiers.

The only adaptation in the search tool ingestion procedure was to make it compatible with a small extension of the MACS format, representing the confidence measure granted to mappings. This enables the tool to be aware of the ranking, and thus present to the user mappings in a decreasing quality order. An example of ingested mapping can be found below:

```
<doc>
  <field name="id">RealDual_LCSH-RAMEAU.3</field>
  <field name="RAMEAU">Analyse mathématique non standard</field>
  <field name="RAMEAU_number">FRBNF121361738</field>
  <field name="LCSH">Nonstandard mathematical analysis</field>
  <field name="LCSH_number">sh85082118</field>
  <field name="measure">0.966091783079296</field>
</doc>
```

5.3 Concluding remarks

Our matching experiments on RAMEAU, LCSH and SWD show that we can automatically produce mappings of surprisingly good quality, even when using quite naive translation and matching methods. The lexical methods produce a relatively high coverage of the MACS manual mappings, which indicates that the use of such very simple algorithms can already support the creation of manual mappings. The instance-based mapping methods provide mappings that are significantly complementary to manual ones. This is interesting, as it indicates that, while each approach brings useful results, the intensional and extensional semantic links are largely different. More efforts would however now be required to turn these findings into an appropriate methodology to assist manual alignment, or to evaluate to which

¹ <http://oaei.ontologymatching.org/2009/align.html>

extent “approximate” mappings can still benefit to multilingual collection access for end users in the TELplus case.

From a more practical perspective, automatically produced alignments can seamlessly be included in a multilingual search prototype. In particular, it is easy to perform a meaningful selection of the mappings, based on their expected quality, both regarding precision of the mappings and their coverage. The main drawback is that this requires a prior evaluation of the mappings. Even if it is possible to do it on a selected sample, this can be regarded as a significant effort, depending on the constraints that apply to the context where alignments have to be employed.

References

[E07] Euzenat, J., Shvaiko, P.: *Ontology Matching*. Springer Verlag (2007)

[IKMW09] Antoine Isaac, Dirk Kramer, Lourens van der Meij, Shenghui Wang, Stefan Schlobach, Johan Stapel. *Vocabulary Matching for Book Indexing Suggestion in Linked Libraries – A Prototype Implementation & Evaluation*. Proceedings of the 8th International Semantic Web Conference (ISWC 2009).

[IMSW07] Isaac, A., van der Meij, L., Schlobach, S., Wang, S.: *An empirical study of instance-based ontology matching*. In: Proceedings of the 6th International Semantic Web Conference (ISWC 2007).

[L99] Landry, P.: *Multilingual Subject Access Project – Comparative analysis of titles indexed using LCSH, RAMEAU and SWD / RSWK*. Technical Report (1999)

[MIGB07] Malaisé, V., Isaac, A., Gazendam, L., Brugman, H.: *Anchoring Dutch Cultural Heritage Thesauri to WordNet: two case studies*. In: *ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*

[SM83] Salton, G., McGill, M.J.: *Introduction to Modern Information Retrieval*. McGraw-Hill (1983)

[S09] Schopman, B.: *Instance-Based Ontology Matching by Instance Enrichment*. Master’s Thesis, Vrije Universiteit Amsterdam, 2009. Available at <http://sites.google.com/site/bschopman/master-thesis>

[S00] Svenonius, E.: *The Intellectual Foundation of Information Organization*. MIT Press (2000)

[V97] Vizine-Goetz, D.: *Popular LCSH with Dewey Numbers: Subject headings for everyone*. *Annual Review of OCLC Research* (1997)

[WISS09] Shenghui Wang, Antoine Isaac, Balthasar Schopman, Stefan Schlobach, Lourens van der Meij. Matching multi-lingual subject vocabularies. Proceedings of the 13th European Conference on Digital Libraries (ECDL2009)

[WSTA09] Shenghui Wang, Stefan Schlobach, Janet Takens, and Wouter van Atteveldt. Mapping-Chains for studying Concept Shift in Political Ontologies. Proceedings of the Third International Workshop on Ontology Matching (2009).

Annex – system specifications

All experiments were carried out using a server with the following characteristics:

Processors	AMD Opteron™ 8220
Number of processor cores	8
Processor clock frequency	2800 MHz
Internal memory	32GB
Operating system	Linux version 2.6.18-6-amd64, Debian 4.1.1-21
Java™ VM	Java HotSpot™ 64-Bit Server VM, build 1.5.0 14-b03, mixed mode

Note that we have not applied multi-threading in our experiment implementations, so the main program always uses a single thread and thus a single processor core.