Putting ontology alignment in context: usage scenarios, deployment and evaluation in a library case

Antoine Isaac[1,2], Henk Matthezing[2], Lourens van der Meij[1,2], Stefan Schlobach[1], Shenghui Wang[1,2], and Claus Zinn[3]

[1] Vrije Universiteit Amsterdam
[2] Koninklijke Bibliotheek, Den Haag
[3] Max Planck Institute for Psycholinguistics, Nijmegen

Abstract. Thesaurus alignment plays an important role in realising efficient access to heterogeneous Cultural Heritage data. Current ontology alignment techniques, however, provide only limited value for such access as they consider little if any requirements from realistic use cases or application scenarios. In this paper, we focus on two real-world scenarios in a library context: thesaurus merging and book re-indexing. We identify their particular requirements and describe our approach of deploying and evaluating thesaurus alignment techniques in this context. We have applied our approach for the Ontology Alignment Evaluation Initiative, and report on the performance evaluation of participants' tools wrt. the application scenario at hand. It shows that evaluations of tools requires significant effort, but when done carefully, brings many benefits.

1 Introduction

Museums, libraries, and other cultural heritage institutions preserve, categorise, and make available a tremendous amount of human cultural heritage (CH). Many indexing schemes have been devised to describe and manage the heritage data. There are thesauri[4] specific to fields, disciplines, institutions, and even collections. With the advent of information technology and the desire to make available CH resources to the general public, there is an increasing need to facilitate interoperability across collections, institutions, and even disciplines.

By providing representational standards (such as SKOS [1]) as well as generic tool support [2], Semantic Web technology has recently taken a more prominent role in this facilitation. A technology that can help with some of the CH interoperability problems is ontology alignment [3]. Ontology alignment aims at aligning classes (and properties) from different ontologies, by creating sets of correspondences between these entities. Applied to CH vocabulary cases, this could help, for instance, to access a collection via thesauri it is not originally

[4] Here we use the word *thesaurus* to refer to all controlled vocabularies that are used in the Cultural Heritage field: classification schemes, subject heading lists *etc.* To denote the elements contained in these vocabularies, we will use the word *concept*.

indexed with, to interconnect distributed, differently annotated collections on the object level, or to merge two thesauri to rationalise thesaurus maintenance.

Unfortunately, experience shows that existing ontology alignment tools often under-perform in applications in the CH domain [4]. We believe that striving for generality of alignment technology is part of the problem. To this end, we argue that the *generation* and the *evaluation* of thesaurus alignments must take into account the application context. Current alignment research within the Semantic Web community, unfortunately, underestimates the importance of requirements from real applications. Evaluation efforts, such as those of the Ontology Alignment Evaluation Initiative[5] mostly favour "application-independent" settings, where, typically, manually-built gold standards are created and used. Such gold standards are actually biased towards – at best – one single usage scenario (*e.g.*, vocabulary merging), and can be of little use for other scenarios (*e.g.*, query reformulation). Efforts leading to an application-specific assessment are under way [5], but further work is required.

The following questions need to be answered to successfully deploy and evaluate alignment techniques:

- What kind of usage scenarios require alignment technology?
- For a given scenario,
  1. What is the meaning of an alignment, and how to exploit it?
  2. How to use current tools to produce the required type of alignment?
  3. How to evaluate an alignment appropriately?

Our aim is to illustrate how to answer these questions from a realistic application perspective. We focus on analysing application requirements and user needs as well as determining practical processes.

By answering these questions (both methodologically and empirically) we will validate two general hypotheses regarding the evaluation of ontology alignment:

- Evaluation results can depend on the evaluation strategy, even when applied to a same scenario.
- Evaluation results can depend on the scenario, even when the most appropriate evaluation strategy is applied.

These hypotheses, although quite obvious, have nevertheless been rather neglected. The ontology alignment community needs to better take on board requirements that come from a wide variety of real application contexts, and also evaluate the performance of their tools in the light of such requirements. Our aim is also to gain more insight with regard to the comparison between performances of different alignment techniques. The hypothesis, already formulated in alignment research, is that some techniques will be more or less interesting to pursue, depending on the application scenario at hand.

The next section introduces our application context, situated at the National Library of the Netherlands, where two thesauri need to be aligned for various scenarios.

[5] http://oaei.ontologymatching.org

2 The need for thesaurus alignment at KB

The National Library of the Netherlands (KB) maintains two large collections of books. The *Deposit Collection* comprises all Dutch printed publications (one million items), and the *Scientific Collection* has about 1.4 million books on the history, language and culture of the Netherlands. Each collection is annotated – *indexed* – using its own controlled vocabulary. The Scientific Collection is described using the GTT thesaurus, a huge vocabulary containing 35,194 general concepts. The books in the Deposit Collection are mainly described against the Brinkman thesaurus, which contains a large set of headings (5,221) for describing the overall subjects of books. Currently, around 250,000 books are shared by both collections and therefore indexed with both GTT and Brinkman concepts.

The two thesauri have similar coverage (2,895 concepts actually have exactly the same label) but differ in granularity. Represented in SKOS [1] format, each concept has one preferred label, synonyms and other alternative labels, extra hidden labels and scope notes. Also, both thesauri are structured by *broader*, *narrower* and *related* relations between concepts, but this structural information is relatively poor. GTT (resp. Brinkman) contains only 15,746 (resp 4,572) hierarchical *broader* links and 6,980 (resp. 1,855) associative *related* links. On average, one can expect at most one parent per concept, for an average depth of 1 and 2, respectively. GTT has 19,752 root concepts.

The co-existence of these different systems, even if historically and practically justified, is not satisfactory. First, both thesauri are actively maintained but independently from each other, which doubles the management cost. Second, disconnected thesauri do not support unified access to both collections. Except the 250,000 dually indexed books, books can only be retrieved by concepts from the particular thesaurus they were originally indexed with.

In order to achieve better interoperability and reduce management cost, thesaurus alignment plays a crucial role, with regard to the following scenarios.

1. **Concept-based search:** support the retrieval of GTT-indexed books using Brinkman concepts, or *vice versa*.[6]
2. **Re-indexing:** support the indexing of GTT-indexed books with Brinkman concepts, or *vice versa*.
3. **Integration of one Thesaurus into the other:** support the integration of GTT concepts into the Brinkman thesaurus, or *vice versa*..
4. **Thesaurus Merging:** support the construction of a new thesaurus that encompasses both Brinkman and GTT.
5. **Free-text search:** support the search for books using free-text queries that match user search terms to GTT or Brinkman concepts.
6. **Navigation:** support users to browse both collections through a merged version of the two thesauri.

Different scenarios have different requirements with regard to the usages of the alignment. In the following sections, we will focus on two scenarios –

[6] This is a simple version of query reformulation using links between thesauri.

thesaurus merging and book re-indexing – and investigate their different impact on the general thesaurus alignment problem.

3 The thesaurus merging scenario

To reduce thesaurus management and indexing costs, KB considers to merge their Brinkman and GTT thesauri into a single unified thesaurus. The question is whether the thesaurus merging task can be supported by ontology alignment tools, and how. Clearly, this application scenario requires ontology matchers to recognise semantic relations (in particular, equivalence) between the concepts of the two thesauri. These alignment links, together with the respective thesaurus-internal semantic relations, will constitute a semantic network that can then be exploited to create the unified thesaurus. This scenario can be likened to *ontology engineering* use cases for alignment, as presented in [3], chapter 1.

Until now, thesaurus merging scenarios have been rather neglected by the research community [6], since this task, like multilingual thesaurus building, can raise a multitude of languages-specific and cultural issues. GTT and Brinkman are both Dutch thesauri, and they describe similar and quite general domains. Therefore, such issues will not arise, and intuitively, alignments shall rather cover large parts on the input thesauri. The largest problem to address here is the different semantic granularity of the thesauri. GTT and Brinkman have different sizes for a similar subject coverage. Many concepts of one thesaurus, thus, will not have equivalent concepts in the other thesaurus. For instance, due to different indexing usages — *cf.* our discussion on post-coordination in section 4 — Brinkman and GTT both have the terms “gases” and “mechanics”, but only Brinkman has the *compound* concept “Gases; mechanics”.

3.1 Formulation of thesaurus merging problem

In the thesaurus merging scenario, we define an alignment as a function that states whether two concepts are linked to each other by a semantic relation:

$$A_{merge} : \mathcal{G} \times \mathcal{B} \times \mathcal{SR} \rightarrow \{true, false\},$$

where $\mathcal{G}$ and $\mathcal{B}$ denote the sets of GTT and Brinkman concepts. $\mathcal{SR}$ denotes the set of semantic relations, containing the *equivalence* link that is used to merge concepts, as well as the *broader*, *narrower*, and *related* semantic links that are proposed in standard thesaurus building guidelines [7] and which are also used at the KB.

Alignments between *combinations of concepts* could help a thesaurus engineer to determine whether a complex subject from one thesaurus is covered by several simpler concepts from the other thesaurus. Note, however, that such alignments are not explicitly relevant for the task at hand. The GTT and Brinkman thesauri do not deal with complex concepts in their respective internal formats, and there is no reason why a unified thesaurus should entertain such structure.

3.2 Proceeding with existing alignments for thesaurus merging

Standard ontology matchers can give results that do not fit the function specified above. For instance, instead of typed relations, one tool can use only one generic symmetric similarity link, eventually coming with a certainty degree – typically in the $[0, 1]$ interval. Such certainty information could be computed, say, by counting the number of books that share Brinkman and GTT indices.

Such results would have to be re-interpreted and *post-processed*, *e.g.* by defining a threshold that filters out the weakest links, or validates some links as cases of *equivalence* and others as cases of mere *relatedness*. In some variants of this scenario, human experts can be involved, using the certainty information to accept or reject candidate links. Consider, for instance, the terms "making career" (denoting a series of actions) and "career development" (the result of these actions). A matcher might relate them via an equivalence link, but with a weak probability. But in some contexts, a thesaurus builder could nevertheless decide to merge them into a single concept, or to make one a specialisation of the other.

3.3 Evaluation Method

Alignments are evaluated in terms of their individual mappings as follows:

- If the mapping link is *related*, *broader* or *narrower*, then assess whether it would hold within one unified thesaurus, given both concepts were to be included in it;
- If the mapping link is *equivalence*, then assess whether the two concepts should be merged in such a thesaurus.

An evaluation must consider two aspects: (i) *correctness*: what is the proportion of correct (or acceptable) links in the results? (ii) *completeness*: what proportion of the links required by the scenario did the results contain? Two standard Information Retrieval measures, precision and recall, are normally used. But there are alternative options, which could fit well specific – supervised – scenario variants. For example, a semantic version of precision and recall, as proposed in [8], would help to discriminate between near misses and complete failures, when a human expert editing the proposed alignment links can transform these near misses into correct matches.

Ideally, the evaluation should be based on a complete reference alignment. If no complete gold standard is available, then absolute recall cannot be computed. In this case, especially valid when the focus is on comparing the relative performances of several alignments produced by different tools, one can measure *coverage*. For each alignment, we define the coverage as the proportion of all good mappings found by this alignment divided by the total number of distinct good mappings produced by all alignments. This coverage is proportional to the real recall, and in any case it provides an upper bound for it — as the correct mappings found by all participants give a lower bound for the total number of correct mappings.

The thesaurus merging evaluation actually resembles classical *alignment evaluation*, as presented, *e.g.*, in [9]. In the next section, we discuss the re-indexing scenario, where deployment and evaluation, formulated in terms of specific *information needs*, is more in line with the “end-to-end” approach described in [9].

4 The book re-indexing scenario

To streamline the indexing of Dutch scientific books, currently described with both Brinkman and GTT, thesaurus alignment can be used as follows:

- Computer-supported book indexing with the following workflow: first, a new book is manually described with GTT by a human expert; subsequently, thesaurus alignment technology is asked to generate a Brinkman index, given its GTT annotation. In a supervised setting, the expert, not necessarily the same person, can then accept or adapt this suggestion.
- KB decides to terminate their use of GTT in favour of the Brinkman thesaurus. All books that have been indexed with GTT concepts shall be re-indexed with Brinkman using thesaurus alignment technology. Again, this re-indexing could be fully automatic or supervised. In the latter, a human expert takes a book’s new Brinkman indexing as suggestion, possibly changing it by removing or adding Brinkman concepts.

This scenario is about *data migration.* Similar to the “catalogue integration” use case in [3], chapter 1, some tool transforms descriptions of objects — in our case book indices — from one vocabulary to the other.

Re-indexing books is fundamentally a non-trivial activity. Consider the following two books and their respective index in the GTT and in the Brinkman thesaurus:

- Book *Allergens from cats and dogs*
  - Brinkman: “allergie,” (*allergy*) “katten,” (*cats*) “honden” (*dogs*)
  - GTT: “allergenen,” (*allergens*) “katten,” “honden,” “immunoglobulinen” (*immunoglobulins*)
- Book *Het verborgen leven van de kat*
  - Brinkman: “katten”
  - GTT: “diergedrag,” (*animal behaviour*) “katten,” “mens-dier-relatie” (*human-animal relation*)

As we can see, the same concept used in different indices should be jointly aligned to different sets of concepts. Some of these required alignments are obvious, while some are more complicated, sometimes even reflecting different analysis levels on a same book.

These phenomena are related to the use of *post-coordinate indexing.* As above examples show, when a book is annotated with several GTT subject concepts, these concepts are considered in combination, each being a factor of the subject of the whole book. The re-indexing function must therefore deal with more than just the (arbitrary) co-occurrence of concepts.

4.1 Formulation of the book re-indexing problem

A book is usually indexed by a set of concepts; an alignment shall specify how to replace the concepts of a GTT book indexing with conceptually similar Brinkman concepts to yield a Brinkman indexing of the book:

$$A_{reindex} : 2^{\mathcal{G}} \to 2^{\mathcal{B}},$$

where $2^{\mathcal{G}}$ and $2^{\mathcal{B}}$ denote the powersets of the GTT and Brinkman concepts. Note that the sets of proposed Brinkman concepts would also be preferably small. Observation of usage reveals that 99.2% of the Deposit books are indexed with no more than 3 Brinkman concepts and that 98.4% of the GTT-indexed books have no more than 5 concepts.

The (informal) semantics of the required alignments can be determined the following way. First, consider the simple case – concerning 18.7% of KB's dually indexed books – where the GTT index of a given book consists of one GTT concept, and its Brinkman index book consists of one Brinkman concept. Here, our function needs to translate a single GTT concept g into a single Brinkman concept b. The re-indexing can be information-preserving if the concepts g and b are judged equivalent; the re-indexing can loose information if concept b is judged broader than concept g; and using a narrower concept or related concept B, additional information may be introduced, which could be wrong but not necessarily so. These cases correspond to well known mapping situations as described at the semantic level by [6] and given representation formats [1].

The simple case of one-to-one mappings can be generalised to many-to-many (set-to-set) mappings. A complex subject built from GTT concepts by means of post-coordination can be replaced by another complex subject built from Brinkman concepts (or a simple one) if these two complex subjects have equivalent meanings, or, to a lesser extent, if the meaning of the first subsumes the meaning of the second or if they have overlapping meaning.

4.2 Proceeding with existing alignments for re-indexing books

As mentioned before, off-the-shelves matchers usually produce only one-to-one mappings, possibly with a weight.[7] To meet the specific requirements of book re-indexing, a post-processing step is required to obtain multi-concept book indices. In our earlier research, we have presented a procedure – and several options – to do this [11]. The first step consists in grouping concepts based on the mappings they are involved in, so as to obtain translations rules of the form $\{g_0, g_1, \ldots, g_m\} \mapsto \{b_1, b_2, \ldots, b_n\}$. The set of GTT concepts attached to each book is then used to decide whether these rules are *fired* for this book. Given

[7] The *block matching* approach [10], which maps together sets of concepts, is an important exception. However, the nature of block matching prevents us from using such tools in our domain as it constructs sets of semantically close concepts rather than sets of semantically distinct concepts that can be used together. Furthermore, its computational complexity makes it difficult to apply to large datasets.

a book with a GTT annotation G_t, there are several conditions which can be tested for firing a given rule $G_r \mapsto B_r$: (1) $G_t = G_r$; (2) $G_t \supseteq G_r$; (3) $G_t \subseteq G_r$; (4) $G_t \cap G_r \neq \emptyset$. If several rules can be fired for a same book, several strategies can also be chosen for creating the final Brinkman annotation. The most simple one is to consider the union of the consequents of all the rules.

Note that the number of options may be further increased by considering scenarios where human experts are involved in the production of indices. Here, for instance, similarity weights, as given by alignment tools, could be used to generate probability of appropriateness for each candidate concept. An expert would validate the proposed indices using this information.

4.3 Evaluation Method

In the re-indexing scenario, evaluating an alignment's quality means assessing, for each book, the quality of its newly assigned Brinkman index. This assessment gives an indication of the quality of the original thesaurus alignment in the context of the $A_{reindex}$ function that was built from it. We can consider the following evaluation variants and refinements.

Evaluation settings

Variant 1: Fully automatic evaluation. Reconsider the corpus of books that belong both to KB Scientific and Deposit collections. The corpus comprises 243,887 books that are already manually indexed against both GTT and Brinkman. In this variant, the existing Brinkman indices are taken as a gold standard that the evaluated re-indexing procedure must aim to match. That is, for each book in the given corpus, we compare its existing Brinkman index with the one that has been computed by applying $A_{reindex}$. The similarity between these two Brinkman concept sets can be computed, yielding a measure that indicates the general quality of $A_{reindex}$.

Variant 2: Manual evaluation In this variant, a human expert is asked to judge the correctness and completeness of candidate Brinkman indices for a sufficiently large set of books, hence producing a reference indexing. This assessment will vary depending on the scenario that defines how alignment technology is being deployed. Notice that in an unsupervised setting, strict notions of completeness and correctness apply. Here, instead of testing *e.g.* strict set equality, a human expert is likely to accept semantically close Brinkman concepts. The notions for correctness and completeness are thus different and possibly less strict. In this variant, experts are further asked to indicate the concepts which they may eventually use for indexing this book. If the proposed concepts are not part of their ideal choice, then experts can add those concepts. Ideally, this list should contain all the concepts that the human expert expects to describe a given book properly.

Having a human expert in the loop further helps dealing with three important evaluation issues:

– *Indexing variability.* Usually, there is more than one correct indexing of a given book, and two experts might index a given book in two different ways. Having an expert to complement a machine-produced Brinkman index with her own, might make explicit this variability. Also, asking a human expert on the *acceptability* – as opposed to strict validity – of a machine-generated index may increase completeness and correctness results, as human judgement is more flexible and open-minded than automatic measures.
– *Evaluation variability.* Along the same line, the assessment of a book index itself may vary among human evaluators. A manual evaluation allows us to compare several judgements on the same alignment results. One can attempt to address the reliability of the chosen evaluation measure, and then devise new approaches to compensate for the weaknesses that were found.
– *Evaluation set bias.* The corpus of dually indexed books that is needed for variant 1 might have some hidden specific features, while manual evaluation with human experts can be performed on any part of the KB collections.

Evaluation measures All evaluation variants depend on a test set of books indexed with GTT terms. Applying the re-indexing procedure for each book will then produce a set of Brinkman terms. These terms can then be compared against the reference set (or gold standard) that either stems from the existing Brinkman annotation or is set by human experts.

First, we measure how well the generated Brinkman book indices match the correct ones. Correctness and completeness of returned indices are assessed by *precision* and *recall* defined at the indexing level as follows:

$$P_a = \frac{\sum \frac{|\{b_1,\ldots,b_n\}\cap A_{reindex}(\{g_1,\ldots,g_m\})|}{|A_{reindex}(\{g_1,\ldots,g_m\})|}}{\#books\_fired}\ , \quad R_a = \frac{\sum \frac{|\{b_1,\ldots,b_n\}\cap A_{reindex}(\{g_1,\ldots,g_m\})|}{|\{b_1,\ldots,b_n\}|}}{\#books\_total},$$

where $\{b_1, \ldots, b_n\}$ (resp., $\{g_1, \ldots, g_m\}$) is the set of correct Brinkman (resp., existing GTT) concepts for the book; *#books_total* is the number of books in the evaluation set; and $\#books\_fired$ is the number of books for which a re-indexing has been provided. We also use, as a combination of the precision and recall defined above, a Jaccard overlap measure between the produced annotation (possibly empty) and the correct one:

$$J_a = \frac{\sum \frac{|\{b_1,\ldots,b_n\}\cap A_{reindex}(\{g_1,\ldots,g_m\})|}{|\{b_1,\ldots,b_n\}\cup A_{reindex}(\{g_1,\ldots,g_m\})|}}{\#books\_total}$$

Second, we measure the performance of the re-indexing at a broader level, in terms of book retrieval. We consider that a book is retrievable when its correct and generated indices overlap, that is, $\{b_1, \ldots, b_n\} \cap A_{reindex}(\{g_1, \ldots, g_m\}) \neq \emptyset$ — we then call it a *matched* book. Here, *precision* is defined as the fraction of books which are considered as matches according to the previous definition over the number of books for which a new index was generated; and *recall* is defined by the fraction of the "matched" books over the total number of books:

$$P_b = \frac{\#books\_matched}{\#books\_fired} \quad , \quad R_b = \frac{\#books\_matched}{\#books\_total}.$$

Note that in all these formulas, results are counted on a book and annotation basis, and not on a rule basis. This reflects the importance of different rules: a rule for a frequently used concept is more important for the application than a rule for a rarely used concept.

5 Implementing application-specific evaluation for the OAEI Library Track

Since 2004, the Ontology Alignment Evaluation Initiative (OAEI) organises campaigns to review the performance of current state-of-the-art ontology alignment technologies in different cases. Among the six data sets of the 2007 campaign, the *Library* track[8] proposed to align the GTT and Brinkman thesauri, made available in the SKOS [1] and OWL [12] formats. Participants, who had no *a priori* knowledge of the evaluation procedures, were required to deliver SKOS mapping relations: `exactMatch`, `broadMatch` and `relatedMatch`. Three OAEI participants sent results for this track: **Falcon** [13] – 3,697 `exactMatch`, **DSSim** [14] – 9,467 `exactMatch` – and **Silas** [15] – 3,476 `exactMatch` and 10,391 `relatedMatch`.

5.1 Thesaurus merging evaluation

As there was no reference alignment available that maps the complete Brinkman thesaurus to the GTT thesaurus, we used coverage instead of recall for comparing the alignments, as presented in section 3.3. Moreover, to minimize the number of alignments that had to be evaluated, we decided to automatically construct a reference alignment based on a lexical procedure. The method compares labels with each other (literal string matching), but also exploits a Dutch morphology database to recognise variants of a word (*e.g.*, singular and plural). As a result, 3,659 correct equivalence links were obtained.

We only evaluated the `exactMatch` mappings, as only one participant provided another link type. For a representative sampling, the three sets of `exactMatch` mappings were partitioned into sections, one for each combination of the four considered sources (participant alignments plus reference set). For each of the resulting sections that were not in the lexical reference alignment, a sample of mappings was selected and evaluated manually. A total of 330 mappings were assessed by two Dutch native speakers.

From these assessments, precision and coverage were calculated with their 95% confidence intervals, taking into account sampling size and evaluator variability. The results are shown in Table 1.

Clearly, Falcon outperforms the other two systems. Falcon's high precision expresses in the following numbers: 3,493 links are common to Falcon's alignment and the reference alignment; Falcon's alignment has 204 mappings not in the reference alignment (of which 100 are judged correct); and the reference alignment has 166 mappings not in Falcon alignment.

[8] `http://www.few.vu.nl/~aisaac/oaei2007`

Participant	Precision	Coverage
Falcon	0.9725 ± 0.0033	0.870 ± 0.065
Silas	0.786 ± 0.044	0.661 ± 0.094
DSSim	0.134 ± 0.019	0.31 ± 0.19

Table 1. Precision and Coverage for the thesaurus merging scenario

Like Falcon, DSSim also uses a lexical approach for ontology alignment. However, its edit-distance-like approach is more prone to error: only between 20 and 400 of its 8,399 mappings not in the reference alignment were judged correct. In fact, given a selection of 86 mappings from the set of 8,363 mappings unique to DSSim, not a single one was evaluated as correct by the human evaluators. The Silas tool succeeds most in adding mappings to the reference alignment: 234 of its 976 "non-lexical" mappings are correct; nevertheless, it fails to reproduce one third of the reference mappings, and therefore, its coverage is relatively low.

5.2 Book re-indexing evaluation in OAEI 2007

Automatic evaluation and results The automatic evaluation relies on comparing, for the dually indexed books, existing Brinkman indices with the ones that were generated using the alignment. Following the procedure of section 4.2, rules were generated to associate one GTT concept with a set of Brinkman concepts, using a simple grouping strategy. When considering exact matches only, this gives 3,618 rules for Falcon, 3,208 rules for Silas and 9,467 rules for DSSim. One rule is then fired on a given book if its GTT concept is contained in the GTT annotation of this book, *i.e.*, using the firing condition (4) introduced in Section 4.2. When several rules can be fired for a book, the union of their consequents forms the Brinkman re-indexing of the book, which can then be compared to the existing annotation.

Participant	P_b	R_b	P_a	R_a	J_a
Falcon	65.32%	49.21%	52.63%	36.69%	30.76%
Silas	66.05%	47.48%	53.00%	35.12%	29.22%
DSSim	18.59%	14.34%	13.41%	9.43%	7.54%
Silas+related	69.23%	59.48%	34.20%	46.11%	24.24%

Table 2. Performance of book re-indexing generated from mappings.

Table 2 shows the evaluation results when only `exactMatch` mappings are exploited. Interestingly, comparing these results with Table 1, Silas performs as well as Falcon does here. The exploitation of the Falcon alignment resulted in at least one correct Brinkman term per book for nearly half of the test set. At the annotation level, half of the generated Brinkman concepts were judged incorrect, and more than 60% of the gold standard was not found. As mappings from Falcon are mostly generated by lexical similarity, these figures clearly indicate that lexical approach is not sufficient for the book re-indexing scenario.

The results also confirms the sensitivity of mapping evaluation methods to certain application scenarios. Among the three participants, only Silas generated relatedMatch mappings. We combined these mappings with the exactMatch ones to generate a new set of 8,410 rules. As shown in the *Silas+related* row of Table 2, the use of relatedMatch mappings increases the chances of a book being given a correct annotation. However, unsurprisingly, the precision of annotations decreases as noisy results were introduced.

Manual evaluation and results

Evaluation process A sample of 96 books was randomly selected from the dually annotated books indexed by KB experts in 2006. For each of these books we applied the annotation translation rules derived from each participants' results, using only the exactMatch links. For each book, the results of these different procedures were merged into lists of candidate concept annotations. We also included the original annotations in the candidate lists. On average this procedure resulted in five candidate concepts per book.

To acquire experts' assessments of the candidate annotations, paper forms were created for each book in the sample. A form presented the book's cataloguing information — author, title, year of publication *etc.* — plus the candidate annotations found for this book.

Given a book's description and annotation, experts were then asked to judge the *acceptability*[9] for each and every annotation concept. The experts were also asked to select from the candidates the ones they would have *chosen as indices*. For this, experts had the opportunity to add terms to the candidate list they found most appropriate to describe the book.

A preliminary version of the evaluation form was tested with two professional indexers. The experts agreed with our notion of "acceptability" and also found the average number of candidate concepts adequate. Four professional book indexers from the Depot department at the KB, all native Dutch speakers, took part in the final evaluation. Each expert assessed the candidate annotation for every element of the sample set.

Results Table 3 presents the results averaged over the four experts. Interestingly, these human assessments are significantly higher than the figures obtained from our automatic evaluation. It suggests that the chosen application context requires an evaluation that takes into account the indexing variability of human experts.

To assess *evaluation variability*, we computed the *Jaccard overlap* between evaluators' assessments. On average, two evaluators agreed on 60% of their assessments. Using Krippendorff's α coefficient, a common measure for computational linguistics tasks [16], the overall agreement between two evaluators is $\alpha = 0.62$. According to standard interpretation, this indicates large variability.[10]

[9] As it was hinted in Section 4.3, this formulation aims to avoid too narrow judgements. The evaluator can here anticipate situations where other indexers might have selected indices different from hers, *e.g.* when the subject of the book is unclear, the thesaurus contains several concepts equally valid for the book.

[10] Although, the tasks usually analysed with this coefficient (part-of-speech tagging, for instance) are less variable than subject indexing.

Participant	P_a	R_a	J_a	P_a	R_a	J_a
Falcon	74.95%	46.40%	42.16%	52.63%	36.69%	30.76%
Silas	70.35%	39.85%	35.46%	53.00%	35.12%	29.22%
DSSim	21.04%	12.31%	10.10%	13.41%	9.43%	7.54%

Table 3. Performance of mappings as assessed by manual evaluation (left), compared to automatic evaluation results (right, from Table 2).

For measuring *indexing variability* between evaluators, we computed the average Jaccard overlap between their chosen indices as well as the α. Again, we have quite a low overall agreement value – 57% for the Jaccard, 0.59 for the α – which confirms the high intrinsic variability of the indexing task.

Additionally, evaluators assessed the original Brinkman indices for the books of the sample, which we had added in the candidate concepts to be evaluated. These concepts are the results of a careful selection of a human expert. Therefore, they cannot capture all the acceptable concepts for a book and recall is, unsurprisingly, relatively low (R_a=66.69%). More interestingly, almost one in five original index concept were judged not acceptable (P_a=81.60%), showing indeed that indexing variability matters considerably, even when the annotation selection criteria are made less selective.

6 Discussion and conclusion

We reported on application scenarios that require the exploitation of thesaurus alignment. Existing off-the-shelves ontology alignment technology may be of limited practical use. It needs better characterisation for deployment and evaluation. We have studied these problems for the applications at hand, focusing on thesaurus merging and re-indexing scenarios.

All scenarios have some important common features, such as benefiting from alignments links that have thesaurus-inspired semantics. But there are also important differences, such as the emphasis on certain type of relations, or cardinality aspects. Furthermore, depending on the deployment strategy or the degree of human supervision, different levels of precision or recall can also be expected for alignments. Some cases actually hint at using less standard measures for correctness and completeness, or cautiously interpreting the evaluation results in the light of the specific characteristics of the application – *e.g.*, indexing variability.

The results we obtained for the OAEI Library track confirm the importance of considering applications when deploying and evaluating alignments. The practical usefulness of a certain alignment can vary from one scenario to the other, and from one setting to another – even within one scenario. Evaluation needs to be done carefully.

Our approach is related to existing work on solving heterogeneity problems in thesaurus applications [6] or in wider controlled vocabulary contexts, including index translation and query reformulation, either from a general expert perspective [17] or with a strong emphasis on formalisation [18]. Yet, none of these efforts really study the gap between application requirements and alignments such as produced by state-of-the-art techniques. Our work started to investigate this problem, aiming at the alignment research community where application requirements have only recently come under consideration [19, 5].

Our experiments show that the abovementioned gap is manifold. For instance, it is important to obtain asymmetric hierarchical alignment links – *e.g.* broader – instead of a plain similarity measure to address thesauri's different semantic granularity. Aligning sets of combined concepts instead of the standard one-to-one mappings will also be crucial for some data conversion scenarios. The lack of such capacities raises the need for a post-processing step. In such a step, decisions could be made that do not fit the assumptions guiding the computation of the alignment.

Current state-of-the-art alignment tools come with limited options with regard to the specific type of mapping that they can generate (limited mapping relations; 1-1 mappings). That makes it hard, if not impossible to use, evaluate, and deploy them in real-world contexts. We hope indeed that this paper will guide researchers from the Semantic Web domain to continue enhancing their existing tools, possibly by taking into account the diversity and richness of applications contexts and their requirements, some of which we reported here.

Our evaluations have also demonstrated that the compared usefulness of specific alignment strategies is dependent on the application scenario, confirming our last hypothesis. For the merging scenario, Falcon, which relies more on lexical matching when the structure of vocabularies is poor, outperforms the other two participants. While in the translation scenario, Silas, which detects links based on extensional information of concepts, performs as well as Falcon does. This is in line with the current trend in alignment research that investigates ways to perform case-specific selection of alignment strategies [19, 20]. This also gives further reasons to keep up application-specific evaluation efforts in OAEI-like campaigns.

Evaluation when done carefully brings many benefits. We will therefore continue our own effort on determining deployment and evaluation contexts, including the cases we have only briefly mentioned here, as well as other cases outside the KB context. We also plan to investigate alignment methods that better match application requirements, extending for example our previous work on producing multi-concept alignment using instance-based similarity measures [11].

Acknowledgements

This work is funded by the CATCH programme of NWO, the Dutch Organisation for Scientific Research (STITCH project) and by the European Commission (TELplus project). The evaluation at KB would not have been possible without the contribution of Yvonne van der Steen, Irene Wolters, Maarten van Schie, Erik Oltmans and Johan Stapels. We also thank the referees for their comments.

References

1. Isaac, A., Summers, E.: SKOS Simple Knowledge Organization System Primer. W3C Working Draft (2008)
2. Schreiber, G., et al.: Multimedian e-culture demonstrator. In: International Semantic Web Conference (ISWC2006), Athens, USA (2006)
3. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer-Verlag (2007)
4. van Gendt, M., Isaac, A., van der Meij, L., Schlobach, S.: Semantic web techniques for multiple views on heterogeneous collections: a case study. In: European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), Alicante, Spain (2006)
5. Šváb, O., Svátek, V., Stuckenschmidt, H.: A study in empirical and casuistic analysis of ontology mapping results. In: European Semantic Web Conference (ESWC 2006), Innsbruck, Austria (2007)
6. Doerr, M.: Semantic problems of thesaurus mapping. Journal of Digital Information **1**(8) (2001)
7. International Standards Organisation: ISO 2788-1986 Documentation - Guidelines for the establishment and development of monolingual thesauri (1986)
8. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: International Joint Conference on Artificial Intelligence (IJCAI 2007), Hyderabad, India (2007)
9. van Hage, W.R., Isaac, A., Aleksovski, Z.: Sample evaluation of ontology-matching systems. In: Fifth International Workshop on Evaluation of Ontologies and Ontology-based Tools, ISWC 2007, Busan, Korea (2007)
10. Hu, W., Qu, Y.: Block matching for ontologies. In: International Semantic Web Conference (ISWC2006), Athens, USA (2006)
11. Wang, S., Isaac, A., van der Meij, L., Schlobach, S.: Multi-concept alignment and evaluation. In: Second International Workshop on Ontology Matching, ISWC 2007, Busan, Korea (2007)
12. McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language Overview. W3C Recommendation (2004)
13. Hu, W., Zhao, Y., Li, D., Cheng, G., Wu, H., Qu, Y.: Falcon-AO: results for oaei 2007. In: Second International Workshop on Ontology Matching, ISWC 2007, Busan, Korea (2007)
14. Nagy, M., Vargas-Vera, M., Motta, E.: DSSim – managing uncertainty on the semantic web. In: Second International Workshop on Ontology Matching, ISWC 2007, Busan, Korea (2007)
15. Ossewaarde, R.: Simple library thesaurus alignment with SILAS. In: Second International Workshop on Ontology Matching, ISWC 2007, Busan, Korea (2007)
16. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology, 2nd edition. Sage, Thousand Oaks,CA (2004)
17. British Standards Institution: Structured Vocabularies for Information Retrieval – Guide. Part 4: Interoperability between vocabularies. Working Draft (2006)
18. Miles, A.: Retrieval and the semantic web. Master's thesis, Oxford Brookes university (2006)
19. Euzenat, J., Ehrig, M., Jentzsch, A., Mochol, M., Shvaiko, P.: Case-based recommendation of matching tools and techniques. KnowledgeWeb Project deliverable D1.2.6 (2007)
20. Tan, H., Lambrix, P.: A method for recommending ontology alignment strategies. In: International Semantic Web Conference (ISWC 2007), Busan, Korea (2007)