The value of usage scenarios for thesaurus alignment in Cultural Heritage context

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 is important for efficient access to heterogeneous Cultural Heritage data. Current ontology alignment techniques provide solutions, but with limited value in practice, because the requirements from usage scenarios are rarely taken in account. In this paper, we start from particular requirements for book re-indexing and investigate possible ways of developing, deploying and evaluating thesaurus alignment techniques in this context. We then compare different aspects of this scenario with others from a more general perspective.

1 Introduction

Museums, Libraries, and other cultural heritage institutions (CHI) preserve, categorise, and make available a tremendous amount of human cultural heritage. For this, curators, librarians and others have been devising many indexing schemes to describe and manage their assets. There are thesauri[4] specific to fields, disciplines, institutions, and even collections. While specific thesauri are very well manageable at the micro level, providing a good mirror of an institution's purpose and assets or a collection's scope, they hamper access for those not familiar with their structure and content. With the advent of information technology and the desire to make available CH resources to the general public, there is also an increasing need to facilitate access across collections, institutions, and even disciplines and fields.

Accepting the wealth, diversity and value of Cultural Heritage Institutions' assets, technology is required that facilitates its access and exploitation by curators, librarians, archivists (the CH keepers) as well as researchers and the general public (the CH consumers). Such technology has to be able to process thesauri of various domains of knowledge, size, structure, quality, and granularity of modelling. Common to CH thesauri is their practical and large-scale use in managing CH collections. In fact, the corpus of collection items that is described with a

[4] Here we use the word *thesaurus* to refer to all controlled vocabularies that can be used in the Cultural Heritage field: classification schemes, subject heading lists, etc. Later on, to denote the elements contained in these vocabularies, we will however use the word *concept*, and not *term*, as often found in thesaurus-related literature.

thesaurus adds to the description of the thesaurus itself; it assigns meaning to each of the thesaurus' concepts.

One technology that can help solving some of the CHI access problems is ontology alignment [1]. Ontology alignment aims at aligning classes (and properties) from different ontologies, by creating sets of correspondences between these entities. Applied to the thesaurus case, this could help for instance to exploit one thesaurus *via* another one, or to merge two thesauri together. However, the description of ontology alignment is rather vague from a practical perspective, and this is often mirrored in existing research publications in this area. In this paper, we argue that *thesaurus alignment* is an interesting research problem where ontology alignment has to be adapted to concrete usage scenarios. While there is value in describing the problem of alignment in purely abstract terms, our research results show that it must also be approached and complemented by generating and *exploiting* alignments for well-defined problems at hand.

Here, we argue that existing alignment methods often fall short on three points. First, the generation of thesaurus alignments must take into account the application context. The exploitation of thesaurus structure, approaches that use the lexical characteristics of a thesaurus' concepts, or corpus-based methods that analyse instances being annotated with thesaurus concepts — without being informed by the usage scenario at hand — may only return results of limited value.

Second, the evaluation of alignment techniques (and the quality of the alignment being generated with them) must also take into account the application context. Existing research in this area has underestimated the dependence of alignment requirements on the applications that use them. In a number of alignment tool evaluations like the Ontology Alignment Evaluation Initiative (*http://-oaei.ontologymatching.org*) the focus has mostly been on vocabularies and some "application-independent" meaning. This typically results in using manually-built gold standards that are supposed to be neutral. However, such a gold standard is necessarily biased by considering a scenario (*e.g.*, vocabulary merging), and could thus be of limited use to assess the relevance of an alignment for another scenario (*e.g.*, query reformulation). Efforts leading to an application-specific assessment of alignment results are under way, see [2], but further work is required.

Third, the deployment of technology based on ontology alignment is uncharted territory – most research efforts have been devoted to finding alignment techniques and tools – and should be neglected.

We believe that the following questions need to be answered to successfully develop, deploy, and evaluate thesaurus alignment techniques in the CH domain:

- What kind of usage scenarios require thesaurus alignments?
- For a given scenario, how will the alignment be used?
- For a given scenario, can we elicit requirements for alignments (for instance, with regard to the semantics of alignment links?)
- For a given scenario, what kind of problems occur when using the results of current alignment tools?

– For a given scenario, what are the options available for evaluating alignments?
– Across several scenarios, can we identify common points in terms of alignment semantics and evaluation criteria?

Our aim is to illustrate how to answer these questions from a given application perspective. We focus on analysing application requirements and user needs as well as determining realistic processes and tools. Our application context is situated at the National Library of the Netherlands, where two thesauri need to be aligned to enhance their interoperability and management (Section 2). To clarify the different uses that can be made of such an alignment, we are gathering scenarios, which we also describe. A more detailed account of one of these scenarios – book re-indexing – and its impact on alignment development, deployment and evaluation will be given in Section 3. In Section 4 we discuss the differences that can be observed from one scenario to another, trying to determine whether commonalities can be found, and in Section 5 we conclude.

2 The need for thesaurus alignment at KB

The National Library of the Netherlands (KB) maintains a large number of collections. Its *Deposit Collection* comprises all Dutch printed publications (one million items), and its *Scientific Collection* has about 1.4 million books on the history, language and culture of the Netherlands. Each of these two collections has its own indexing system. The Deposit Collection is indexed in terms of the *Brinkman thesaurus*, a set of approximately 5000 concepts; and the Scientific Collection is described with the *GTT*, a huge vocabulary of about 35,000 general concepts ranging from *Wolkenkrabbers* (skyscrapers) to *Verzorging* (care).

The two thesauri have similar coverage but differ in granularity. Also, both thesauri are structured by *broader than*, *narrower than*, and *related to* relations between concepts, but they differ in their structural complexity.

2.1 Thesaurus maintenance and interoperability

The co-existence of these different systems raises issues with regard to maintenance costs and interoperability issues. First, the cost issue becomes obvious when one considers that KB has approximately 250,000 books which have been indexed with both GTT and Brinkman. Both vocabularies are also actively maintained to ensure that new topics (say, Semantic Web) are described with appropriate Brinkman and GTT concepts. Since the thesauri are managed independently from each other, this adds to the duplication of work in terms of thesaurus engineering (addition of new concepts and their proper integration in thesaurus with the definition of relations to existing concepts). Second, as the thesauri are disconnected from each other, there is also no unified access, in particular for thesaurus-based book retrieval. That is, except for the 250,000 dually indexed books, Brinkman concepts need to be used to retrieve books from the Deposit Collection, and GTT concepts need to be used to retrieve books from the Scientific Collection.

2.2 Streamlining thesaurus management at KB

In the long term, KB aims at developing and deploying methods that help streamlining collection and thesaurus management. One option proposes to develop a new thesaurus for the Humanities, the core of KB's interests, by integrating and restructuring the relevant parts of Brinkman and GTT. This new thesaurus will then replace Brinkman and GTT, assuming that legacy data (the description of the collection using Brinkman and GTT) is properly dealt with. There are other options, which we will explain next.

2.3 Use cases for streamlining thesaurus management at KB

Within the KB, there are several application scenarios that require thesaurus alignment:

1. **Concept-based search:** support the retrieval of GTT-indexed books using Brinkman concepts, or *vice versa*. This scenario is aimed at librarians with an intricate expertise of Brinkman or GTT.
2. **Re-indexing:** support the indexing of GTT-indexed books with Brinkman concepts, or *vice versa*. This scenario is aimed at annotators with an intricate expertise of Brinkman or GTT.
3. **Integration of one Thesaurus into the other:** support the integration of GTT elements into the Brinkman thesaurus, or *vice versa*, yielding a "Brinkman-ized" version of the GTT or a "GTT-ized" version of Brinkman. This scenario is aimed at thesaurus experts.
4. **Thesaurus Merging:** support the construction of a new, better-quality thesaurus that encompasses both Brinkman and GTT, and privileges none of the input thesauri. This scenario is also aimed at thesaurus experts.
5. **Free-text search:** support the search for books using free-text queries that would aim at matching user search terms to both GTT or Brinkman concepts. This scenario is aimed at the layman.
6. **Navigation:** support users to browse the Deposit and Scientific collections through a merged version of the two thesauri. This scenario is aimed at the layman.

3 The book re-indexing scenario

3.1 Application scenario

We will now discuss the second use case in more detail, as it exemplifies well the complexity of the problems encountered and has benefited from the most important part of our implementation effort until now. To streamline the cataloguing of Dutch scientific books, currently described with both the Brinkman thesaurus and GTT, KB management may consider the following two options:[5]

[5] In the following, the roles of the Brinkman thesaurus and GTT are interchangeable.

– 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.

Example. 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 alignments are obvious, while some are more complicated, sometimes even reflecting different analysis levels on a same book. The sets of 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.

In both cases, having a human expert in the loop makes a difference. When recommending a Brinkman indexing, thesaurus alignment technology is only required to generate a list of concepts, potentially complemented by a probability that indicates the appropriateness of each candidate. This list may contain concepts that the human will not use in finalising a Brinkman indexing, but it should contain all the concepts that the human expert expects to properly describe a given book. In automatic mode, the re-indexing of books from GTT to the Brinkman thesaurus should be correct and complete, and the margin of error should be negligible.

This scenario is about *data migration*. Similar to the "catalogue integration" use case in [1], chapter 1, some tool transforms description of objects — in our case book indices — from one vocabulary to the other. Obviously, this tool must have access to some alignment of concepts from the GTT thesaurus to the Brinkman one; this primary and enabling resource, however, is complemented by tool elements that decide *how* to read and exploit alignment information. In the interactive scenario with a human expert, the tool will also need to be

complemented with user interface components that facilitate the selection of concepts.

We now further refine the problem at hand, and the requirements it imposes on alignments.

3.2 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:

$$align_{reindex} : 2^{GTT} \rightarrow 2^{Brinkman}$$

where GTT and $Brinkman$ denote the sets of GTT and Brinkman concepts, and 2^{GTT} and $2^{Brinkman}$ denote the powersets of these.

The function $align_{reindex}$ would satisfy the automatic re-indexing option.[6] In the supervised scenario, we need to attach to the resulting concepts a probability that marks their appropriateness, generalising the function as follows:

$$align'_{reindex} : 2^{GTT} \rightarrow 2^{Brinkman \times [0,1]}$$

Interpretation of the required alignment links Let us first consider the simple case where the GTT index of a given book consists of one GTT concept, and the Brinkman index of the same book consists of one Brinkman concept. In this case, our function needs to translate a single GTT concept g_1 into a single Brinkman concept b_1.[7]

A human expert may interpret this re-indexing in one of the following ways:

- g_1 and b_1 are equivalent or nearly equivalent concepts; so that there is no loss of information.
- The concept b_1 is more general than the the concept g_1. If the book is about a given subject, then one can consider that it is also about a subject which is more general. This solution ensures proper recall and precision if b_1 is the "most specific subsumer" of g_1 that can be found in Brinkman, following the *indexing specificity rule* which says a book shall be annotated by the most precise subject that can be applied, and only this one. However, the initial information is only partly kept.
- The concept b_1 is more specific than g_1. In this case $align_{reindex}$ adds information to the new index that was not present in the original one. The newly added information can be false, but not necessarily so. This situation is acceptable if Brinkman contains neither equivalent nor more generic concept, or if the "semantic distance" between g_1 and the more specific b_1 is smaller than the one between g_1 and any more generic concept.

[6] There is also the following option that we have not considered so far: an expert selects between several possible indices. In this case, $align_{reindex}$ becomes a relation.

[7] This situation would actually fit 18.7% of the dually indexed books.

– The concept b_1 and the concept g_1 have overlapping meanings, but one cannot be said to be a specialization of the other. In this case, $align_{reindex}$ may introduce information to the resulting index that could be false. This alignment could be used, however, in the absence of any satisfactory equivalent or broader concept in Brinkman for the concept g_1.

These cases correspond to well known mapping situations as described at the semantic level by [3] and given draft representation formats [4].

The simple case of one-to-one mappings can be generalised to many-to-many (set-to-set) mappings if we take into account *post-coordinate indexing*: when a book is annotated with several subjects, it is about these subjects considered in combination.[8] When two or more GTT concepts are used together, the re-indexing function must be able to deal with more than just the (arbitrary) co-occurence of concepts. That is, re-indexing a GTT book with several GTT concepts is different from re-indexing GTT books indexed with these individual GTT concepts.

Apart from this, however, the following is still valid. A complex subject built from GTT concepts by means of post-coordination can be replaced by another complex subject built from Brinkman concepts 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.

3.3 Exploiting the results of a thesaurus alignment

There are a number of off-the-shelf tools and techniques for ontology alignment. However, their generic nature makes it hard to use them in practise for real-world problems. A first point is the difficulty to interpret the links these tools draw to connect items from one thesaurus to items of a second thesaurus. For instance, they may use mapping constructs with no clear semantics ("=") or constructs with semantics that go beyond what can be used in a specific scenario (*e.g.*, boolean disjunctions of concepts).

A second point regards the use of post-coordination in some situations. If books which are annotated with more than one thesaurus concept are considered as annotated with a more complex, virtual subject, the concepts shall be translated as groups and not as single entities. However, alignment tools generally tend to focus on one-to-one correspondences and not on the ones that involve concept combinations.

Alignments may as a consequence fail to meet the specific requirements that stem from the application scenario.

As an example, in [5] we followed an instance-based approach for alignment construction, measuring the similarity between any two concepts of the two given thesauri. Based on co-occurrence of concepts in the same annotations, we

[8] With GTT, if a book is indexed with the concepts "historische geografie" and "Nederland", then one should expect the book to about a more complex "historical geography of the Netherlands" subject, of which historical geography and Netherlands are facets.

have generated one-to-one GTT-Brinkman correspondences with some similarity measure attached: (g_i, b_j, m). These, as simple one-to-one correspondences, are not sufficient to properly address the re-indexing scenario. Nevertheless, one could exploit alignments that contain one-to-one mappings by post-processing these to yield *aggregate* concepts, and consequently, multi-concept alignments.

Grouping concepts based on one-to-one mappings. For a concept C_0 (from either thesauri), we use the aforementioned weighted correspondences to generate the list of the top k most similar concepts, for a chosen threshold k:

$$C_0 \rightarrow (C_1, C_2, \ldots, C_k).$$

Note that this list may contain concepts from both thesauri.

These top k concepts, together with the concept C_0, are expected to form a group of concepts closely related from an extensional point of view. We then take this set of concepts and split it into two separate thesaurus-specific sets, which we use to define an m to n mapping between both thesauri (with respect to the $k+1$ concepts). That is, if

$$g_0 \rightarrow (b_1, g_1, \ldots, b_n, g_m),$$

where $m + n = k$, then we generate a translation rule

$$\{g_0, g_1, \ldots, g_m\} \mapsto \{b_1, b_2, \ldots, b_n\}.$$

Partitioning concepts based on clustering. Another alternative to produce the translation rules requested by the scenario is to apply a similarity-based clustering technique [6], using the information found in the dually indexed books in order to partition the concepts into clusters. If one cluster contains k concepts

$$(b_1, g_1, \ldots, b_n, g_m),$$

where $m + n = k$, then we generate a translation rule

$$\{g_1, \ldots, g_m\} \mapsto \{b_1, b_2, \ldots, b_n\}.$$

We now have rules that allow to translate groups of concepts. When the GTT annotation of a book (G_t) matches the left-hand-side of a rule $G_r \mapsto B_r$, one annotates this book with the set of Brinkman concepts found in the right-hand-side of the rule – the rule is "fired". However, such a strategy would lead to a low coverage of the book collection: depending on the techniques used, our experiments resulted in a number of rules ranging from 717 to 8334.

We therefore decided to test different rule firing strategies, which amounts to defining several translation functions. Given a book with a GTT annotation G_t, the following conditions are tested for firing a given rule $G_r \mapsto B_r$: (1) $G_t = G_r$; (2) $G_t \supseteq G_r$; (3) $G_t \subseteq G_r$; (ALL) $G_t \cap G_r \neq \emptyset$.

3.4 Evaluation Design

In the chosen scenario, evaluating the quality of an alignment means assessing, for each book, the quality of its newly assigned Brinkman index, independently from the GTT index that was used to produce it. It is important to note that this assessment, thus, also judges the quality of the re-indexing function $align_{reindex}$ and how it exploits the alignment under scrutiny. We argue that this evaluation method is more informative than assessments that are detached from any practical use case.

An evaluation must consider two complementary aspects: (i)*completeness*: does the $align_{reindex}$ function return a Brinkman index for every book's GTT index of the Scientific Collection? — this corresponds to the notion of *recall* in Information Retrieval, in terms of the books contained in the collection; (ii)*correctness*: for each book, is the Brinkman index that has been produced for it correct (or acceptable)? — this corresponds to the notion of *precision* in Information Retrieval, in terms of the concepts that are contained in the generated indices. For the re-indexing scenario, we can consider the following evaluation variants and refinements.

Variant 1: Fully automatic evaluation. Reconsider the corpus of books that belong both to KB Scientific and Deposit collections. The corpus comprises approximately 250,000 books that are already indexed against the GTT and the Brinkman thesauri. The existing Brinkman indices are taken as a gold standard that any automatic procedure must aim to match. That is, for each book in the given corpus, we compare its existing (and manually constructed) Brinkman index with the one that has been computed by applying $align_{reindex}$ to the book's existing GTT index. The similarity between these two Brinkman concept sets can be computed. Averaging this similarity over the set of all books of this corpus will yield a measure that indicates the general quality of the original thesaurus alignment in the context of the $align_{reindex}$ function that was built from it.

Variant 2: Manual evaluation I. In this variant, a human expert is asked to assess the correctness and completeness of Brinkman indices for a sufficiently large set of books. This assessment will differ depending on the scenario that defines how alignment technology is being deployed. In an unsupervised setting, the margin of error should be negligible, and therefore, strict notions of completeness and correctness apply, although they will be less strict than in Variant 1. Instead of testing strict set equality, a human expert is likely to accept semantically close Brinkman concepts. Moreover, the human expert might also take the original GTT index for a book into account, especially when she is asked to do so.[9]

[9] The task of the expert could then be formulated as follows: "Given your knowledge of both GTT and Brinkman thesauri, would you judge the following Brinkman index is an appropriate index for this given book, considering its GTT index?"

In a supervised mode, where alignment technology "only" needs to be deployed to support the human annotator, the notions for correctness and completeness are different, and possibly, less strict. Here, a human expert is asked whether the set of Brinkman concepts that is being suggested to index a book includes all the concepts that she may eventually use for this task, and whether the suggestion set contains concepts that are clearly incorrect. In addition, a human expert may assess the size of the suggestion set as too many suggestions might have a negative contribution to the book indexing effort. As in the automatic mode of Variant 2, we may want to give human experts the GTT index of a book to consider in their judgement.

Variant 3: Manual evaluation II. In this variant, the human expert plays a more active role than in Variant I. She is first asked to produce a Brinkman index for this book by herself (given its GTT index — and not the book and its contents). Only then will the expert be given the automatically generated Brinkman index, which she then compares with hers, along similar criteria as in Variant 2. The completeness and correctness can then be measured by the number of additions and corrections the expert performs on both Brinkman sets. Alternatively, in a "Turing Test"-like variation of this option, we could ask a second person to identify which Brinkman book index comes from a human expert, and which from a machine; or more precisely, mark both Brinkman book indices for correctness and completeness.

The advantage of having a human expert in the loop is three-fold:

- Indexing variability. Usually, there is no 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 this variability explicit (within the given problem setting).
- Evaluation variability. Along the same line, the assessment of a book index itself may vary among human evaluators. A manual evaluation thus allows us to compare several judgements on the same alignment results. Asking a human expert on the acceptability of a machine-generated index may increase completeness and correctness results as human judgement is more flexible and open-minded than automatic measures. One can also 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 complete Deposit and Scientific collections.

3.5 Evaluation Results

In the experimental context presented in Sect. 3.3, we have performed an evaluation according to Variant 1. For each book which has a dual indexing in terms

of both GTT and Brinkman thesauri, we have automatically compared its existing Brinkman index with the one that has been produced by our alignment technology.

First, we measure how well the generated Brinkman book indices match the existing ones, in order to obtain *precision* and *recall* at the indexing level:

$$P_a = \frac{\sum \frac{\#good\_found}{|align_{reindex}(\{g_1,\ldots,g_m\}|}}{\#books\_fired} \quad , \quad R_b = \frac{\sum \frac{\#good\_found}{|\{b_1,\ldots,b_n\}|}}{\#books\_total},$$

where $\#good\_found$ is the number of existing Brinkman concepts that were found in the generated set of Brinkman concepts; $\#books\_total$ is the number of books in the evaluation set; and $\#books\_fired$ is the number of books for which a translation, or re-indexing, has been provided.

Second, we measure the performance of the re-indexing at the book level, which is more appropriate for the supervised setting. The set of books for which a re-indexing was successful defines the set of books found, and we consider that a book is a match when its original and generated indices overlap, *i.e.* $\{b_1,\ldots,b_n\} \cap align_{reindex}(\{g_1,\ldots,g_m\}) \neq \emptyset$. 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:

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

We have computed these measures for alignment production and deployment in different settings – alignment techniques, grouping methods, firing strategies.Depending on the chosen setting, the value of P_a ranges from 4.16% to 25.17%. This makes the candidate Brinkman indices unusable for the automatic scenario.

Nevertheless, such low figures would be less of a problem in the context of a supervised scenario, where human experts can select correct Brinkman concepts from a larger set of less suited other concepts. Yet, using our alignment technology in this scenario is made almost impossible by a low recall value for R_a, which is at best 16.41%. The book-level recall R_b shows that at most 19.54% of the books in the evaluation set are given one good Brinkman concept. This means that a human expert will often need to add Brinkman concepts to the ones suggested for indexing a given book. Consequently, the techniques that we have explored so far are not sufficiently mature to cope with the re-indexing of books over an entire collection.

Our evaluation has however not taken into account the variability of indexing. To compensate for it, we could investigate semantic measures that take into account the knowledge contained in the Brinkman thesaurus. For instance, following some of the proposals in [7], we could increase the score for completeness and correctness for a given produced index, if the concepts it contains are related to the ones in the reference index via the links found the Brinkman thesaurus.

We will also need to compute precision and recall for Variant 2 and 3, where we expect our alignment technology to score higher, given that human expert take the semantic context of a concept (that is, its relations to other concepts) into account.

4 Comparing alignment requirements and evaluation criteria between scenarios

In the previous section, we have investigated thesaurus alignment requirements for a book re-indexing scenario, and found that requirements varied depending on the level of human involvement. The use of post-coordination in book indexing imposes a difficult constraint on the nature of any produced alignment, or its exploitation. It suggests that alignment technology must support the generation of many-to-many (concept set) correspondences between thesauri rather than (individual) one-to-one mappings. We also saw that the nature of these correspondences should convey specific semantic information such as equivalence, subsumption, and overlapping of concepts or concept sets.

Existing off-the-shelf tools do not satisfy all requirements at once, and hence need to be adapted; and in Sect. 3.3 we sketched two possible post-processing methods to transform 1-1 alignments into many-many ones.

Finally, we have shown how to design evaluations that can clearly indicate whether a given alignment technology "works", given its intended (and specific) application context. We sketched various evaluation variants to match their specific problem settings and then found that their notions (and the respective importance) of correctness and completeness differ.

Now, we compare the use case *re-indexing* with the other five use cases mentioned in Sect. 2.3.

4.1 Semantics of required alignments

Thesauri are usually structured, relating their concepts along three basic types of relations: *broader than*, *narrower than*, and *related to*. Clearly, technology that aligns two (or more) thesauri to each other should exploit such thesaurus-internal structural elements and transfer them to describe relations between concepts *across* thesauri. In addition to the three relation types, however, an alignment should offer an equivalence relation between a concept of one thesaurus and some concept in another thesaurus. Possibly, one would also need a relation type expressing that a concept of one thesaurus semantically overlaps with a concept from another thesaurus (although *related to* could be used to express this).

It seems that alignment-based solutions for each of the other five scenarios would profit from the availability of these relation types. In scenario 1 (concept-based search), for instance, one would first attempt to translate a book search query consisting of GTT concepts with semantically equivalent Brinkman concepts (as indicated by the alignment function). If the alignment does not contain

such equivalences, one then would search the correspondences along an alignment's *broader than* relations to construct a more general Brinkman query; if this fails, one could then exploit an alignment's *related to* links.

Once a book query in terms of Brinkman concepts has been constructed, its result set could still be unsatisfactory. In this case, one could do one of the following to improve the search quality: for instance, check existing relations *in the alignment*, say, by replacing an *equivalence* link by a *broader than* link to yield a larger result set.

Alignment-based methods, however, need to take scenario-specific requirements into account that refine or change the semantics of relation types. Concept equivalence in the book re-indexing scenario must not only consider subject similarity but also indexing policies (as different collections can be indexed at different levels). In a search context, the content of the collections being searched also plays a role. In our KB collections, a statistical analysis of dually indexed books reveals that the Brinkman concept closest to the GTT concept "excavations" is "Archeology; the Netherlands". Concept-based search would profit from exploiting this equivalence while a thesaurus engineer would rather search for better correspondences for scenario 3 or 4.

Additionally, different kinds of links such as *equivalence* and *related-to* will need to be applied differently in specific scenarios. In the navigation and search scenarios mixing both alignment types may prove useful. One can consider two concepts equivalent if they have highly overlapping meanings, like "making career" denoting a series of actions and "career development" rather denoting the result of these actions. This allows for serendipity when searching, and also compensates for the indexing variation phenomenon a user (especially a layman) cannot easily deal with.

At a first glance, such variations in usage and interpretation hamper the value of a common semantic characterisation of alignment links. Yet, this characterisation still lies within what is usually found in CH experts' practice. Links established in a thesaurus will indeed often be influenced by the use of the vocabulary in a given reference collection, while at other times thesaurus engineers will try to stick to a more "neutral" approach, privileging their domain expertise.

4.2 Multi-concept alignments

In the re-indexing scenario, post-coordination suggests that alignments between concept sets are the most appropriate. The same is true for the book search scenarios 1 and 5 in Sect. 2.3. Users often use two or more concepts in a query to find material that is best described by their combination. The situation is different for thesaurus engineering (scenarios 3 and 4). Concept combinations can help a thesaurus engineer determine whether a complex subject from one thesaurus is covered by several concepts from the another thesaurus. However, combinations of concepts are not formally required for thesaurus integration and merging tasks: In fact, the GTT and Brinkman thesauri do not deal with them. A different feature, correspondences between individual subjects applying to a same element, can nevertheless be useful. In an integrated thesaurus, a semantic

equivalence link between "Dutch geography" on the one side, and "geography" and "the Netherlands", on the other side, should indeed be dealt with by introducing "Dutch geography" as a specialisation of both "geography" and "the Netherlands".

4.3 Coverage needs

The need for the deployment of multi-concept alignment technology stems from the coverage requirement specific to re-indexing. Ideally, all possible indices should be translated, which concerns all thesaurus concepts, and every complex subject built from them. Similar coverage is needed for search scenarios, since every possible query should be given an appropriate reformulation.

For the scenarios which integrate individual concepts from one thesaurus into another (3,4 and in some measure 6), it would suffice to give every concept one (or several) corresponding concept(s) from the target thesaurus.

4.4 Precision and recall requirements for evaluation

Alignment technology should optimise precision and recall for the task at hand. As seen for the re-indexing scenario, the optimisation depends on whether alignment is expected to provide directly the results for the scenario (*i.e.*, new indices for books) or whether the alignment is an intermediate step during a general process where humans are involved.

Sub-scenarios that rely on humans to create or validate the output of alignment (*e.g.*, choosing among several candidate indices or query elements) can afford lower precision, but need high recall, as human experts would prefer not to search for information elsewhere. Additionally, a layman will often accept weaker precision than an expert. For instance, a layman may be less demanding regarding the quality of a hierarchy when browsing a collection (scenario 6), but an expert uses this hierarchy as a important guiding resource when indexing books. Precision could also be less important for search scenarios that produce large result sets: wrong results are statistically less significant. In such cases, recall is likely to be more important because missing valid results is considered worse than a few false hits.

5 Conclusion

In this paper, we reported on application scenarios that require thesaurus alignment. For these, ontology alignment technology can be of help, but needs better characterisation for deployment and evaluation. We have studied these problems for the scenarios at hand, focusing on a re-indexing use case. We also compared requirements across scenarios. As our section 4 has shown, some of our findings – *e.g.*, about expected levels of precision and recall – acknowledge the importance of specific application settings. Nevertheless, there are some important

commonalities, such as all scenarios potentially benefiting from alignments links that have thesaurus-inspired semantics.

Our observations are in line with existing work on solving heterogeneity problems in thesaurus engineering scenarios [3, 8] or in wider contexts, including index translation and query reformulation, either from a general expert perspective [9] or with a strong emphasis on formalization [10]. 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 [11, 2]. We hope this paper will encourage researchers and practitioners from the Cultural Heritage domain to report about more case studies, so as to help the alignment research community to enhance existing solutions. We will continue our effort regarding this aspect, 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 work on producing multi-concept alignment using instance-based similarity measures.

References

1. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer-Verlag (2007)
2. Šváb, O., Svátek, V., Stuckenschmidt, H.: A study in empirical and casuistic analysis of ontology mapping results. In: Proc. of the European Semantic Web Conference (ESWC), Innsbruck, Austria (2007)
3. Doerr, M.: Semantic problems of thesaurus mapping. Journal of Digital Information **1**(8) (2001)
4. Miles, A., Dan Brickley, E.: Skos mapping vocabulary specification. W3C Working Draft (work in progress) (2004)
5. 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. (2007) To appear.
6. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science (February 16 2007)
7. Euzenat, J.: Semantic precision and recall for ontology alignment evaluation. In: Proceedings of IJCAI 2007. (2007) 348–353
8. International Standards Organisation: ISO 5964-1985 Guidelines for the establishment and development of multilingual thesauri (1985)
9. British Standards Institution: Structured Vocabularies for Information Retrieval – Guide. Part 4: Interoperability between vocabularies. Working Draft (2006)
10. Miles, A.: Retrieval and the semantic web. Master's thesis, Oxford Brookes university (2006)
11. 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)