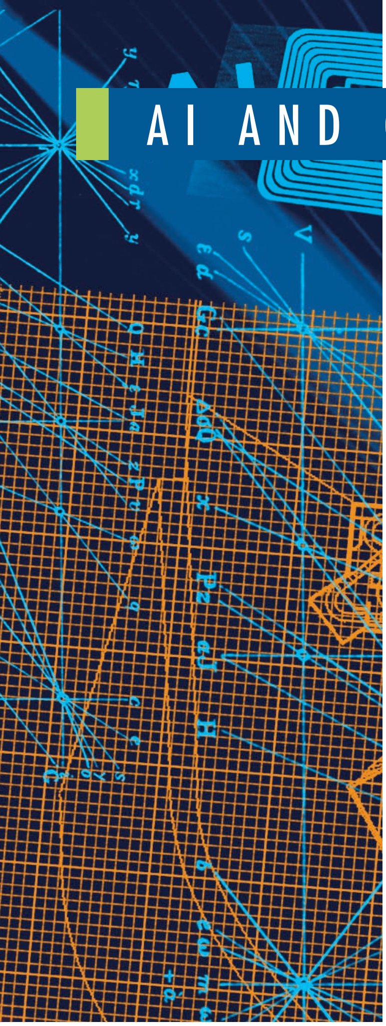

Evaluating Thesaurus Alignments for Semantic Interoperability in the Library Domain

Antoine Isaac, Shenghui Wang, Lourens van der Meij, and Stefan Schlobach, *Vrije Universiteit Amsterdam*

Claus Zinn, *Max Planck Institute for Psycholinguistics*

Henk Matthezing, *Koninklijke Bibliotheek*

Bridging the gap between theoretical study and practical applications would allow more efficient access to heterogeneous cultural heritage data using thesaurus alignments.

Museums, libraries, and other institutions preserve, categorize, and make available a tremendous amount of human cultural heritage (CH). Many indexing schemes have been devised to describe and manage heritage data, including thesauri (classification schemes, subject heading lists, and other controlled vocabularies) specific to fields, institutions, and even collections. The desire to make CH resources available to the general public (for example, see www.europeana.eu) increases the need to facilitate interoperability across different contexts.

By providing representational standards such as the Simple Knowledge Organization System (SKOS; www.w3.org/2004/02/skos) and generic tool support, the Semantic Web community has taken a prominent role in this facilitation. Its ontology-matching branch aims at developing technology to produce alignments—that is, sets of semantic mappings between elements from different vocabularies.[1] One can exploit alignments, for instance, to access a collection via thesauri it is not originally indexed with, to interconnect distributed, differently annotated collections on the object level, or to merge two thesauri to rationalize thesaurus maintenance.

Unfortunately, our experience shows that existing matching tools often do not perform well in CH applications.[2] We believe part of the problem is that they strive for generality. To this end, we argue that the generation and evaluation of thesaurus alignments must take into account well-understood real-world application contexts and their specific requirements.

Ontology Matching

Ontology matching aims at determining the semantic relations between elements of two given knowledge organization systems (for example, ontologies and thesauri).[1] The set of semantic relations usually comprises concept equivalence, hierarchical concept links

1541-1672/09/$25.00 © 2009 IEEE
Published by the IEEE Computer Society

(broader or narrower than), and mere relatedness links. Various research projects in the CH context have tackled the matching task manually—notably, Multilingual Access to Subjects (MACS; http://macs.cenl.org). These projects demonstrated the complexity and cost of manually aligning large vocabularies in realistic collections, and thus the need for computer assistance. Tools under development in the Semantic Web community address these ontology-matching issues (for an overview, see chapter 6 of *Ontology Matching*[1]). We can decompose their complex machinery into a mix of basic techniques:

- *lexical*, detecting similarities between labels and other lexical information of concepts;
- *structural,* using the structure of the ontologies;
- *extensional,* using classified instance data; and
- *background knowledge,* using external knowledge sources such as WordNet (http://wordnet.princeton.edu).

Ontology-matching tools normally take an application-neutral perspective on the matching problem. They typically apply one or several of the techniques we've listed to compute similarity between concepts; the resulting alignments consist of mappings that relate entities via semantic relations, sometimes with a measure attached. However, there are various degrees of freedom:

- Entities can consist of simple concepts or complex constructions of several concepts.
- One entity can appear in only one or any number of mappings (*m*:*n* links).
- The type of mapping relations can range from vaguely to formally specified.
- The measure may denote a probability or an objectively measurable similarity degree.

Decisions on these options influence the quality and usability of the alignments in given application contexts.

Evaluation

Since 2004, the Ontology Alignment Evaluation Initiative (OAEI) has organized campaigns to review the performance of matching technologies in various domains. As most tools are still highly experimental and not used in practical applications, the first evaluation efforts favored mostly "application-independent" methods.[3] Researchers typically created and used manually built reference alignments (or gold standards) that were often biased toward—at best—one single usage scenario (for example, vocabulary merging), with little use for other scenarios. More recent evaluation approaches have adopted more realistic assessments by using, for instance, application-specific sampling methods and measures.[4,5] Further work, however, is needed to better understand real-world use cases, their requirements, and the proper use and evaluation of matching technology.

> We must better understand real-world use cases and their requirements when using and evaluating matching technology.

For this article, we took the OAEI 2007 Library Track[6] as our evaluation context. We gave three participating systems—Falcon, DSSim, and Silas—real-world data from the National Library of the Netherlands (KB) and required them to focus on SKOS mapping relations (in particular, `exactMatch` for equivalence mappings, `broadMatch` for generic-specific relationships, and `relatedMatch` for simple associations). We then evaluated the resulting alignments in the context of realistic scenarios involving the KB.

The Need for Thesaurus Alignment at the KB

The 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 defined by the GTT (Joint Subject Headings) thesaurus, a large vocabulary containing 35,194 concepts standing for general topics, and the Deposit Collection is mainly described using the Brinkman thesaurus, which contains 5,221 headings referencing general subjects. Currently, around 250,000 books are shared by both collections and indexed with both GTT and Brinkman concepts. Both thesauri are in Dutch and have similar coverage, but differ in granularity. Represented in SKOS, each concept has one preferred label, one or more synonyms and alternative labels, extra hidden labels, and scope notes. Both thesauri are structured by `broader`, `narrower`, and `related` relations among concepts, but this structural information is relatively poor. The average depths of the GTT and Brinkman hierarchies are 0.69 and 1.03 respectively, and nearly 20,000 of the GTT concepts have no parents.

The coexistence of these different thesauri, even if historically and practically justified, is not satisfactory. First, both thesauri are actively but independently maintained, which doubles

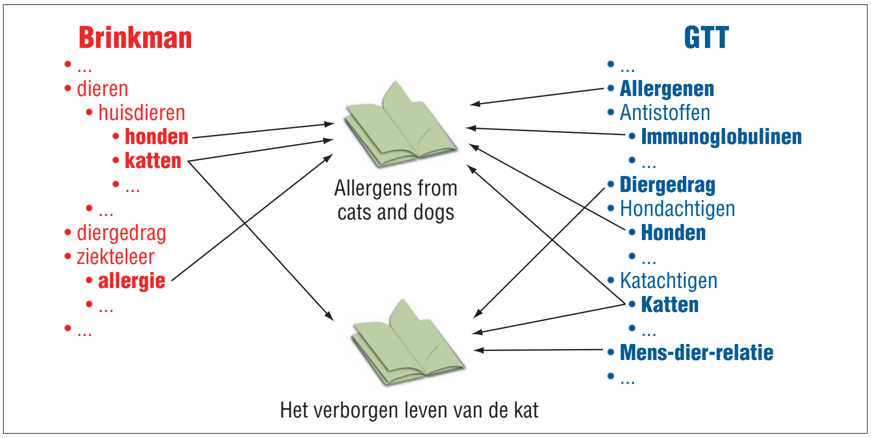

Figure 1. Indexing two books using two thesauri: Brinkman and GTT (Joint Subject Headings). Two books that are indexed by both systems can have different indexings in one system or the other, even when the thesauri feature equivalent concepts.

management costs. Second, disconnected thesauri do not support unified access to both collections. Books can only be retrieved by concepts from the particular thesaurus they are indexed with, except the 250,000 dually indexed books.

To achieve better interoperability and reduce costs, matching technology plays a crucial role, with regard to the following scenarios:

1. *Reindexing*—supporting the indexing of GTT indexed books with Brinkman concepts, or vice versa.
2. *Concept-based search across vocabularies*—supporting the retrieval of GTT indexed books using Brinkman concepts, or vice versa.
3. *Navigation across thesauri*—supporting the exploration of concept spaces across thesauri, and giving (exploratory) access to collection items indexed with selected concepts.
4. *Thesaurus merging*—supporting the construction of a new thesaurus that encompasses both Brinkman and GTT, or the integration of one thesaurus into the other.

These scenarios enable public access to library resources by exploiting information created by librarians (front-office use), and help librarians create new information to enhance such access or improve the library processes themselves (back-office).

We believe these scenarios well represent the potential use of matching technology in the library context. Scenario 4, for instance, considers thesauri and their constituents (concepts and how they are interrelated) as its objects of study and investigates how they can be manipulated to yield other structures. Scenarios 1 and 2 take a more instance-based view where the investigation centers around the use of thesauri for the description, retrieval, and exploration of book collections. Scenario 3 actively interoperates among different concept spaces and collections.

Scenarios

This section takes the viewpoint of the library community, and details use cases that provide representative examples for the exploitation, deployment, and evaluation of alignments. As such, they bring to life and make more concrete problem statements that the ontology-matching community specifies only in abstract terms.[1]

Book Reindexing

To streamline the indexing of books currently described with both Brinkman and GTT thesauri, the KB is considering computer-supported reindexing methods. The conservative approach consists of maintaining both thesauri: a (newly acquired) book is first indexed with GTT by a human expert; given its GTT annotation, a tool automatically generates a corresponding Brinkman annotation, which—in a supervised setting—a human expert can then accept or correct. A more revolutionary approach consists of terminating the use of one thesaurus, say GTT, altogether. Here, all legacy books indexed with GTT will be reindexed with adequate Brinkman concepts. This *data migration* for information integration[1] could be fully automated or supervised.

Reindexing books is not trivial. Figure 1 shows the annotation of two books (in bold) from both thesauri. At first glance, some reindexing seems straightforward as lexically identical or similar concepts occur in both GTT and Brinkman annotations, for example, "katten"—"katten" (cats), "honden"—"honden" (dogs), and "allergie"—"allergenen" (allergy). However, the lexically identical correspondences are not necessarily always correct. For example, "diergedrag" (animal behavior) occurs in both thesauri, but the one in Brinkman does not index the same book.

Moreover, correct reindexing requires more than the identification of one-to-one mappings, as shown in the second book example with three GTT concepts versus one Brinkman concept. Clearly, the librarians' annotation reflects diverging analysis levels or even thesaurus-inherent indexing policies. These examples also highlight the issue of *postcoordinate indexing*: when a book is annotated with several concepts, these concepts can be considered in combination, each being a factor of the subject of the whole book. Reindexing must therefore deal with more than just the arbitrary co-occurrence of concepts.

Problem statement. We need to specify a function that translates the concepts of a GTT-indexed book into Brinkman concepts to yield a Brinkman indexing of the book:

$$f_r : 2^G \rightarrow 2^B,$$

where 2^G and 2^B denote the powersets of GTT and Brinkman concepts. Note that to do this, we must take an idealized and pragmatic stance: the KB's corpus of dually annotated books contains cases where two books with identical GTT annotation have different Brinkman annotations.

Ideally, the GTT index of any given book should be translated into a semantically equivalent or similar Brinkman index. If the latter has broader (more general) semantics, then the translation is not information preserving, but instead loses information. If the latter has narrower (more specific) semantics, then the translation adds information that may be wrong.

Evaluation method. The quality of an alignment is assessed in terms of, for each book, the quality of its newly assigned Brinkman index. We thus evaluate only those parts of the alignment relevant to the task at hand. We measure the correctness and completeness of the reindexing as follows: we define *precision* as the average proportion, for the books provided with a Brinkman reindexing, of the new indices that also belong to a reference (gold standard) set of Brinkman indices. *Recall* is the average proportion, for all books, of the reference indices that were also found using the alignment. The Jaccard similarity—the overlap measure of candidate indices and reference ones—provides a combination of precision and recall.

Automatic evaluation is possible on the 250,000 books manually indexed against both GTT and Brinkman. This gold standard allows us to evaluate any reindexing procedure at the book level: for each book, we compare its existing Brinkman index with the one computed by applying f_r.

For the OAEI Library Track evaluation, the alignments of Falcon, Silas, and DSSim, all consisting of one-to-one mappings, were exploited straightforwardly: for each concept used in a considered GTT annotation, the best mapped concept available (as determined by the strength of the mapping) was added to the Brinkman reindexing. While Falcon and DSSim only produced `exactMatch` mappings, Silas also produced `relatedMatch` ones. Silas was thus evaluated twice, first considering only `exactMatch` mappings, and then adding `relatedMatch` ones to those. At first sight, the results were disappointing.[6] For the best systems, nearly half of the generated Brinkman concepts were incorrect (precision = 54 percent for Silas without `relatedMatch`). Recall is weak as well, as more than 60 percent of the gold standard was not found (recall = 39 percent for Silas with `relatedMatch`).

We need to provide a single access point to multiple collections, each of which is described by a different thesaurus.

Manual evaluation requires human experts to judge the correctness and completeness of candidate Brinkman indices for each book of a sufficiently large sample. We randomly selected a sample of 96 books from the dually annotated set. Each book was reindexed using only `exactMatch` links. We then formed a *candidate index* for each book, combining the book's original annotation with those resulting from the reindexing. Given a book's general description (author, title and so on), experts were asked to judge the *acceptability* of each proposed concept. They were also asked to select—or add—the ones they would have chosen as indices. Four professional book indexers from the KB Depot Department independently assessed each of the 502 new conceptual annotations proposed.

Their assessment yielded significantly better values for both precision and recall. For instance, Falcon's precision and recall respectively improved from 53 to 75 percent and from 36 to 46 percent. This indicates that human experts are more likely to accept semantically close Brinkman concepts. Instead of testing for strict set equality, they applied less strict notions for correctness and completeness. This confirms the subjective nature of the indexing task. Using Krippendorff's α coefficient, the overall agreement between two evaluators on acceptable indices is 0.62, indicating a rather large evaluation variability. We also obtained quite a low overall agreement value on the *chosen* indices (0.59), showing a high level of intrinsic indexing variability. One should consider this aspect carefully when exploiting (and tuning) alignments for this scenario.

Book Search across Collections

In the search scenario, we need to provide a single access point to multiple collections, each of which is described by a different thesaurus:

- to search for a particular book, a librarian formulates a query that consists of a set of, say, GTT concepts;
- the GTT concepts in the query are translated into Brinkman concepts; and
- the search engine executes both queries and gives the results to the librarian.

Conversely, a librarian might want to formulate a query in terms of Brinkman to get access to books that are only annotated with GTT concepts. This scenario is a typical case of *mediation* for information integration.[1]

Problem statement. As the librarian may specify an arbitrary set of GTT concepts to search for a given book, we need to specify a function that translates any member of the GTT concepts powerset to some set of Brinkman concepts, which is then passed to a search routine to retrieve the book(s) in question:

f_s: $2^G \rightarrow 2^{SC} \cup 2^{DC}$.

Here, *DC* denotes the set of all books from the Deposit Collection (indexed with Brinkman concepts), and *SC* denotes the set of all books from the Scientific Collection (indexed with GTT concepts).

The simplest query reformulation takes each concept g_i in the search query $g_1 \ldots g_j$ and searches the alignment for a single semantically equal Brinkman concept b_i, yielding a reformulated query $b_1 \ldots b_j$ (when for each concept g_i such a concept b_i is found). If not every corresponding Brinkman concept is found, we obtain a reformulated query $b_1 \ldots b_m$ with $m < j$.

Evaluation method. An evaluation of this scenario would profit from a representative set of search queries, including whether the results obtained were used (say, by clicking on them). Unfortunately, we have neither the KB's log of concept-based searches nor any information on results used. Alternatively, a realistic evaluation could ask librarians to compare the search result stemming from the given GTT-based query with the one returned by executing an automatically constructed Brinkman-based query in terms of quality or relevance. As this is too labor-intensive, it is not pursued here.

Our automatic evaluation for search builds upon the one for reindexing. For each book *i* of the dual corpus, let G_i be the set of existing GTT concepts of the book, B_i^* the existing Brinkman concepts of the book, and B_i the predicted Brinkman concepts. We then execute the search queries for Q_{G_i}, $Q_{B_i^*}$, and Q_{B_i} and compare their answer sets, where Q_{G_i} denotes the set of books annotated with all concepts in G_i, and similarly for $Q_{B_i^*}$ and Q_{B_i}.

Note that this setup makes three assumptions: that the difference between the Brinkman and GTT indexing policies is negligible; that indexing policies are consistently applied; and that library experts have the talent to specify, for any given book, its correct and complete GTT annotation. Clearly, all three assumptions give an idealized view that is rarely found in library practice.

It's preferable that reformulation of a query fails to establish a correct mapping, rather than giving a wrong mapping.

In this setup, we have different definitions for precision and recall. Instead of computing the intersections between annotations, we compute the average overlap of the answer sets $Q_{B_i^*}$ and Q_{B_i} that result from instructing the search routine with the respective annotations. Similarly, we can adapt the use of the Jaccard measure to compute the similarity between answer sets.

In the dual collection, by definition, it holds that both Q_{G_i} and $Q_{B_i^*}$ return book *i*, and potentially other books that share the same Brinkman and GTT annotations. Clearly, $i \in Q_{B_i}$ if query formulation succeeds in translating at least one GTT concept of g_i into a Brinkman concept $b_j \in B_i^*$, and there is no single incorrect mapping from one GTT index to a Brinkman one. It is therefore preferable that reformulation fails to establish a mapping between one GTT index and a Brinkman index, rather than giving a wrong mapping. In the first case, the Brinkman-based search is less constrained and should thus return an answer set with equal or higher cardinality. Search failure could also result from indexing policies that vary across thesauri. Consider the case where a book is originally annotated with *n* GTT concepts and $m < n$ Brinkman concepts. If query reformulation maps each GTT concept correctly to one Brinkman concept, then the search will fail because it overspecifies the book in the Brinkman context.

We performed automatic evaluation on all books of the dually indexed corpus using the results of the three OAEI participants. Falcon and Silas perform comparably with a precision of around 36 percent, recall of 33 percent, and an overlap of retrieved books of 19 percent using Jaccard. The third participant, DSSim, performed significantly worse (P: 9 percent, R: 7 percent). On average, a book's given Brinkman annotation consists of 1.65 concepts whereas a GTT annotation has 2.3 concepts. Also, on average, a given book's GTT annotation is by no means unique but shared by 76 other books; a book's given Brinkman annotation is shared by 157 books. Reindexing GTT concepts using Falcon's alignment, we computed on average 1.14 Brinkman concepts per book, and those concepts on average identified 124 books. The average intersection size between a book's original Brinkman annotation and the one computed is 0.56. The intersection between the book sets returned by original Brinkman concepts with those obtained by computed Brinkman concepts contains 38.34 books.

Given the size of the answer sets, we may expect f_s and the search engine to optimize the number of search hits by strengthening or relaxing search queries. The engine could first attempt to exploit equivalences between GTT and Brinkman concepts. If this yields no satisfactory results, then it could try broader mappings; if this also fails, it could consider mappings with any relation holding between concepts. Moreover, it is also possible to generalize a given GTT concept by following the structural links within the GTT thesaurus, and then subject the concept's generalization to query reformulation. These strategies are discussed in the following navigation scenario.

Navigating with Multiple Vocabularies

KB might consider *faceted browsing* functionalities that let users easily access multiple collections classified along various vocabularies.[7] With a *single view*, users can access multiple collections via a single thesaurus. With a *combined view*, users can access multiple collections through their respective vocabularies, allowing users to browse the integrated collections as if they were a single collection indexed against two complementary points of view. And with a *merged view*, users can access multiple collections via a merged vocabulary that combines the respective vocabularies of the single collections.[2]

Faceted browsing is a *navigation* scenario, where users are presented with vocabulary terms (that is, concepts) that guide browsing activity through collections. Users can refine, extend, or change the items in focus by selecting more general or more specific concepts or by changing from one concept to another related one. This attracts users with limited expertise in formulating search queries (see "Book Search across Collections" subsection) as well as users who want to explore collections along several dimensions (facets) rather than quickly finding a specific item of interest. To support the exploratory character, faceted browsers often expand user requests; that is, when a concept is selected, all items that are indexed by the concept's specialization or generalization are also retrieved.

The overall nature of exploratory navigation across collections lowers the barrier for an alignment's coverage and precision.

Matching technology can help to support such navigation, which mixes aspects of *ontology merging* and *data mediation*.[1] In the single view, for instance, when users select a concept, an alignment can be exploited to return the corresponding concepts of the other thesaurus (or those that are in a broader or narrower relation to it). A subsequent search can then retrieve all items that are indexed with concepts from either vocabulary.

The overall nature of exploratory navigation across collections lowers the barrier for an alignment's coverage and precision. The failure of matching one concept to its equivalent in the other vocabulary can be covered by considering more loosely related concepts in the other thesaurus. In fact, even in the presence of exact mappings, query reformulation may take into account less precise associations between concepts as it clearly adds serendipity to find items of interest without explicitly searching for them. To address coordination issues, and because faceted browsing environments often use several hierarchies in parallel, it will also be necessary to find mappings between a single concept of one thesaurus and a combination of several concepts of another thesaurus.

Problem statement. Navigation using multiple vocabularies is the dual task of fetching collection items that are indexed against selected concepts from multiple vocabularies as well as proposing new concepts to add to this selection. Formally, navigation in the KB case thus is a function:

$$f_n : 2^{B \cup G} \to 2^{SC \cup DC} \times 2^{B \cup G}.$$

Technically, the extraction of collection items in thesaurus-based navigation is reduced to thesaurus-based search. Each concept selection that results from a user's browsing action triggers a concept search for documents described with this concept. The alignment is thus used for the purpose of reformulating a search query from one vocabulary to another. In practice, this search must be complemented with an adequate GUI that gives access to a click-based selection of search queries, term generalization, and refinement. For the merged view, processes similar to the ones discussed in the thesaurus merge scenario will be needed.

Evaluation method. Whereas search and reindexing can use automatic evaluation, evaluation in the navigation context requires human users, and thus realistic end-to-end evaluation settings.[4] When matching technology, for instance, has been used to produce a merged view, the evaluation will need to emphasize specific interface requirements such as the need for generating balanced hierarchies. The heavy impact of GUI-related design issues (concept selection, GUI-based tree navigation, and so on) on the overall navigation experience makes any evaluation hard; further discussion falls outside the scope of this paper.

Thesaurus Merging/Integration

To reduce thesaurus management and indexing costs, KB considers merging Brinkman and GTT into a single unified thesaurus. When two thesauri cover a similar conceptual space but differ significantly in granularity, the merging task is all but trivial, as many concepts of one thesaurus do not have exactly equivalent ones in the second, and vice versa.

Thesaurus merging becomes more complex when taking into account very practical ontology engineering—in particular, thesaurus design issues. For instance,

- concepts should be kept or removed depending on usage frequencies;
- local or global structures may need to be preserved or reorganized for the sake of hierarchical balance;
- potential merging conflicts may arise and need to be resolved; and
- last but not least, legacy data should still be easily accessible via the resulting thesaurus.

Thesaurus merging is thus a complex cognitive endeavor where matching technology can play only a supporting role—namely, at the local level—by suggesting interthesaurus concept correspondences, as in *ontology merging*.[1] Clearly, given significant differences in the granularity of input thesauri, matching tools must complement concept equivalences with relations that specify broader, narrower, and related links. Information that stems from the alignment can be regarded as merging suggestions that a thesaurus engineer can then accept or reject to form a coherent unified thesaurus. In practice, there are two merging approaches.

Thesaurus integration aims at keeping the structure of one thesaurus (target) and weaving into it the concepts of the other thesaurus (source). The target will preserve (most of) its structure, but its content will be enriched. Enrichment will take the form of concept specializations, and thus broader/narrower mappings are of particular interest. Associative related mappings may also be considered useful, though less crucial.

When two thesauri cover a similar conceptual space but differ significantly in granularity, the merging task is all but trivial.

Thesaurus merging of two input thesauri into a third thesaurus is different, as the output thesaurus can differ dramatically from the originals. Here, the alignment, in combination with intrastructural thesaurus information, can be used to group related concepts, interlinking them via equivalent, broader, narrower, or related relations into small clusters. A thesaurus manager can then reorganize the concepts and relationships within clusters into coherent hierarchies. Gradually, smaller hierarchies can be composed into larger ones, and finally yield a unified thesaurus.

Problem statement. Specify a function that merges the two input thesauri G and B, each with their concepts and interrelations, into a unified thesaurus GB:

$$f_m: T \times T \rightarrow T,$$

where T is the realm of thesauri. A merged thesaurus contains concepts from two input thesauri, and the relations between them are carefully added, taking into account the original thesaurus information and alignment.

We can support the thesaurus-merging process as follows. First, a filter may eliminate those mappings of an alignment with certainty values below a defined threshold. Second, in the absence of hierarchical mappings (as is the case for the three tools under study), mappings could be combined with thesaurus-internal structural information to yield broader and narrower links between concepts across thesauri. Concept relations directly read from the alignment, together with the derived broader and narrower links, can then be suggested to the thesaurus engineer for further consideration.

Evaluation method. We should evaluate two aspects:

- correctness/precision, the proportion of followed suggestions over all suggestions, and
- completeness/recall, computing the proportion of followed suggestions over all merging operations the thesaurus manager actually performed.

In addition, one might consider semantic versions of precision and recall that aim at discriminating complete failures from near misses.[8] Furthermore, redundancy and inconsistency aspects of merging suggestions should also be measured, partially supported by automated reasoning or performed by human experts.

Given the overall complexity of the thesaurus-merging task, the above aspects are often not measurable. The computation of completeness/recall, for

Table 1. Overview of requirements and evaluation aspects. For each scenario, we list (a) its requirements, broken down in different aspects that are relevant to specifying matching processes; (b) the different evaluation options that are available

		Scenarios			
		Reindexing	**Search**	**Navigation**	**Merging**
(a) Requirements	Hierarchical links	Soft *To recover from lack of equality links*	Soft *To broaden or restrict search*	Hard *To guide navigation across thesauri*	Hard *To (add) structure (to) new thesaurus*
	m:*n* links	Hard *For postcoordination*	Hard *For postcoordination*	Soft	Soft
	Coverage	Active concepts in indexing	Active concepts in search and indexing	Active concepts in search and indexing	All concepts of input thesauri
	Best method	Extensional	Extensional	Combining intensional and extensional	Combining intensional and extensional
(b) Evaluation	Object	Book annotations *Original versus new index*	Query results *Original versus new query*	Alignment *GUI*	Alignment *Links used*
	Form	Automatic *Given dual corpus* Manual *Focus on human factor*	Automatic *Given dual corpus* Manual *Focus on human factor*	Manual *User experience efficacy*	Manual

instance, requires a completed merging process. Consequently, the evaluation in this context directly focuses on the alignments produced, where human experts are left to judge concept relations, while keeping the merging task in mind. In automatic settings, alignments could be compared to a reference alignment, if available, or with each other to determine their agreement. Clearly, this only covers a very small part of the whole merging process, but it directly relates to the main role of an alignment for thesaurus merging.

Evaluation. During OAEI 2007, we evaluated the three participants in the setting of the merging scenario.[6] By comparing concept labels (literal string matching) and by exploiting a Dutch morphology database to recognize word variants (that is, singular and plural forms), we constructed a reference alignment of 3,659 equivalence mappings between the GTT and Brinkman thesauri. We also did a representative sampling from the alignments produced by the three participants that were not in the original reference alignment and manually evaluated an additional 330 mappings.

Clearly, our reference alignment has a strong bias toward the lexical equivalence mappings. This explains the high precision of Falcon (97.3 percent), as almost all its mappings are lexically equivalent pairs. Silas, which performed similarly well to Falcon in the reindexing scenario, has a lower coverage of the reference alignment (66.1 percent compared to 87.0 percent for Falcon). Both Silas and DSSim provided mappings that were not in the initial reference alignment. However, the quality of Silas' findings is much higher, as its general precision reaches 78.6 percent (compared to 13.4 percent for DSSim).

Comparison and Discussion

Table 1 provides an overview of the various scenarios in terms of the requirements they impose and the evaluation forms they suggest.

Adequacy of Matching Techniques

The evaluation of the OAEI participant tools in our different scenarios gives some indication of the appropriateness of the matching techniques they employ. While Falcon and DSSim use a combination of lexical and structure-based approaches, Silas computes ontology alignments with a combination of a lexical approach and an instance-based approach. Although Falcon is the best system overall, the relative performance of the other two systems is telling. In the rather *intensional* exercise of thesaurus merging, Falcon's lead is based on the strength of its lexical component, which produced a large number of correct correspondences between lexically equivalent (or similar) concepts. Its structural component could contribute little, but this is due to the low structural similarity between the GTT and Brinkman thesauri. In comparison, DSSim came in a distant third; the edit-distance algorithm in its lexical component was too prone to error, handing over the second place to Silas. Silas' instance-based module, which takes into account a third-party book collection to identify concept co-occurrences, was likely misled by this data, but its lexical module could partly recover from this.

The situation is different in the more *extensional* scenarios of book reindexing and search. While Falcon's alignment is still the best, its lead over Silas is much less significant, indicating that instance-based methods have some benefits. Here, concept equivalence has other than lexical or structural roots. In the book-reindexing scenario, concept equivalence is measured in terms of extensional overlap—for instance, in the intersection of the books that the concepts index. In our search scenario, concepts are considered equivalent if their use in the search query returns significantly overlapping answer sets. Take, for instance, the Brinkman concept "Archeology; the Netherlands" and the GTT concept "excavations." Though lexically different, they can be considered similar, given that they index almost the same set of books in our dually annotated corpus. The ability of extensional techniques to take into account variations of indexing policies across collections (and thus the use of different concept labels) is key in these scenarios.

Semantics of Required Alignments

Thesauri are structured along three basic types of relations: broader, narrower, and related. An alignment normally offers equivalence relations among concepts across thesauri, but other relations are also required in practice, such as those expressing that a concept of one thesaurus is semantically broader than a concept from another, mirroring internal thesaurus relations.

Alignment-based solutions for the KB problems at hand would profit from the availability of these relation types. In the search and navigation scenarios, query reformulation can strengthen or relax search queries by also harvesting narrower and broader mappings. In the thesaurus-merging scenario, thesaurus engineers need to take into account all sorts of relations—equivalent, broader, narrower, and related—to reorganize a network of concepts.

However, the interpretation and exploitation of such common relation types partly depends on the scenario at hand. As hinted previously, reindexing and concept-based search (as well as navigation relying on search) would profit from exploiting mappings that are based on extensional similarity. A thesaurus engineer, on the other hand, would rather search for more intensional equivalence mappings for his task—which actually questions the relevance of using a single equivalence relation in both situations.

In our search scenario, concepts are considered equivalent if their use in the search query returns significantly overlapping answer sets.

Also, in the search and navigation scenarios, one could exploit related relations. Consider, for instance, the concept "making career," denoting a series of actions, being related to "career development," denoting the result of these actions. Following such related links in search query reformulation adds serendipity to searching. To some extent, this also compensates for indexing variation across collections, which (lay) users cannot easily deal with.

Such variations in usage and interpretation of alignment links better reflect CH experts' practice. These should be taken carefully into account when deploying alignments that stem from general-purpose tools.

Mapping Cardinality

The requirement for many-to-many mappings is scenario dependent. For thesaurus merging, concept combinations can help a thesaurus engineer determine whether a complex subject from one thesaurus is covered by several concepts from the other thesaurus. For instance, an equivalent link, "Dutch geography" = "geography" + "the Netherlands," should result in introducing "Dutch geography" as a specialization of both "geography" and "the Netherlands" in the integrated or merged thesaurus. In the book search scenario, postcoordination suggests that mappings between concept sets are probably more appropriate. Users often use two or more concepts in a query to find material that is best described by their combination. The same is true for the reindexing scenario, considering that, on average, a Brinkman annotation consists of 1.65 indices while a GTT annotation has 2.3 concepts.

Coverage Needs

It is usually taken for granted that an alignment between thesauri should cover most of the concepts in both thesauri. That is, concepts from one thesaurus should have at least one correspondence in the other thesaurus, and vice versa. This may hold for the thesaurus-merging scenario, but it is not required for book search and reindexing. For instance, if a GTT concept is not used to index books, it is rather pointless to require mappings involving it for reindexing legacy data using Brinkman. In fact, our dually indexed corpus shows that 15,495 GTT indices (from a total of 35,194 GTT concepts) are used fewer than 10 times and that as many as 11,134 GTT terms are not used at all.

An alignment, however, should have mappings between frequently used concepts. Alignments provided by the OAEI participants cover 51 to 85 percent of Brinkman concepts, but only 10 to 26 percent of GTT concepts. Clearly, such alignments will provide little exploitable material for thesaurus merging. Still, they can make a significant contribution to support book reindexing or search, as long as the most frequently used concepts are covered. As we measure the performance book by book, our evaluation takes into account the concept usage frequency; when the frequently used concepts have a mapping, then a greater number of correct GTT annotations are identifiable. Our results show 46 percent of recall in the manual evaluation, which exceeds expectations, given the low overall coverage.

Precision and Recall Needs

Ideally, matching technology should optimize precision and recall for the task at hand. However, the requirements of precision and recall across scenarios are significantly different, depending on whether tasks are automated (such as the search scenario) or only computer supported (reindexing, thesaurus merging). Scenarios that rely on human involvement (such as choosing among several candidate indices or query elements) can afford lower precision but need higher recall as human experts cannot afford to search for information elsewhere. Additionally, novice users may often accept weaker precision than experts. A layman, for instance, may be less demanding regarding the global quality of the set of results for a query, while an expert may use this result set as an important resource to assess the content of a collection.

Ontology-matching technology can play a major role in granting uniform access to heterogeneously described CH collections. The tools used in our study, however, did not solve any of the problems to our full satisfaction. A major limitation was the tools' lack of support in providing mappings with thesaurus-inspired semantics and mappings between concept sets rather than individuals. This forced us to carefully interpret and postprocess alignments produced by the participant tools, given the individual problem contexts—a task that not all users will be willing or able to perform. We therefore encourage tool developers to implement such functionality or to fine-tune such functionality where it is already present[9,10] and also to participate in future OAEI tracks to evaluate their systems' performance.

Our study also indicated application contexts favoring particular matching methods; lexical approaches work best in intensional scenarios such as thesaurus merging, while instance-based methods have their strengths in extensional contexts such as reindexing and search. Tool selection, however, also needs to take into account notions such as required coverage, precision, and recall, which in turn depend on other factors such as the level of mechanization sought.

We hope that this article guides researchers from the Semantic Web community to better consider real-world thesauri and their use in realistic application contexts, thus better

THE AUTHORS

Antoine Isaac is a researcher in the Computer Science Department at Vrije Universiteit Amsterdam. His research interests include the representation and interoperability of cultural heritage collections and their vocabularies, particularly the use of Semantic Web techniques for representing and aligning knowledge organization systems. Isaac has a PhD in computer science from the University of Paris-Sorbonne. He participates in the W3C Semantic Web Deployment working group and is involved in the design of the Simple Knowledge Organization System (SKOS). Contact him at aisaac@few.vu.nl.

Shenghui Wang is a researcher in the Department of Computer Science at Vrije Universiteit Amsterdam. Her research interests include the problem of semantic interoperability in the cultural heritage domain, including matching different thesauri and deploying the mappings in various interoperability applications. Wang has a PhD in computer science at the University of Manchester. Contact her at swang@few.vu.nl.

Claus Zinn is an R&D engineer at the Max Planck Institute of Psycholinguistics, where he researches and develops methods and tools to support ontology management and the analysis of language corpora. His research interests include automated reasoning, dialogue management, and knowledge representation. Zinn has a PhD in computer science from the University of Erlangen-Nuremberg. Contact him at claus.zinn@mpi.nl.

Henk Matthezing is a project leader in the R&D Department at the Koninklijke Bibliotheek, the National Library of the Netherlands. He has been involved in numerous library Internet projects—for example, the introduction of Dublin Core in the Netherlands, the experimental stages of KB's e-Depot, digital preservation, digitization, the conversion of collection metadata sets into an open-access environment, and semantic interoperability to enhance access to digital cultural heritage collections. Matthezing has an MA in economic and social history from the University of Amsterdam. Contact him at henk.matthezing@kb.nl.

Lourens van der Meij is a scientific programmer in the Knowledge Representation and Reasoning group at Vrije Universiteit Amsterdam. His research focus is the application of Semantic Web techniques in the cultural heritage domain. Van der Meij has an MS in computer science from Delft University and an MS in physics from Leiden University. Contact him at lourens@cs.vu.nl.

Stefan Schlobach is an assistant professor for AI at Vrije Universiteit Amsterdam. His main research areas are semantic interoperability in the fields of cultural heritage and health, as well as nonstandard reasoning, in particular the study of new methods for symbolic approaches to uncertainty, approximation, and anytime behavior in large ontologies. Schlobach has a PhD from King's College London for research on combining knowledge representation and learning techniques. Contact him at schlobac@few.vu.nl.

balancing their generality design imperative with what is needed in practice for usable and high-performance tools. In this vein, we hope that future OAEI campaigns other than the Library Track provide similarly concrete scenarios and evaluation contexts, which consequently will lead to technology that better addresses real-world problems. We also hope that this article encourages and guides the library sciences community in adopting matching technology. There is much complexity to deal with, and the human factor of technology use should not be ignored. But the benefits for librarians and end users alike are worth it. ■

Acknowledgments

This work is funded by the Catch program of the Dutch Organization for Scientific Research (Stitch project) and by the European Commission (TELplus project). The evaluation at the National Library of the Netherlands (KB) would not have been possible without the contribution of Yvonne van der Steen, Irene Wolters, Maarten van Schie, Erik Oltmans, and Johan Stapel. We are also grateful to the three participants of the Ontology Alignment Evaluation Initiative 2007 Library Track.

References

1. J. Euzenat and P. Shvaiko, *Ontology Matching*, Springer, 2007.
2. M. van Gendt et al., "Semantic Web Techniques for Multiple Views on Heterogeneous Collections: A Case Study," *Proc. 10th European Conf. Research and Advanced Technology for Digital Libraries* (ECDL 06), LNCS 4172, Springer, 2006, pp. 426–437.
3. S. Zhang and O. Bodenreider, "Experience in Aligning Anatomical Ontologies," *Int'l J. Semantic Web & Information Systems*, vol. 3, no. 2, 2008, pp. 1–26.
4. L. Hollink et al., "Two Variations on Ontology Alignment Evaluation: Methodological Issues," *The Semantic Web: Research and Applications, Proc. 5th European Semantic Web Conf.* (ESWC 08), LNCS 5021, Springer, 2008, pp. 388–401.
5. B. Lauser et al., "Comparing Human and Automatic Thesaurus Mapping Approaches in the Agricultural Domain," *Proc. Int'l Conf. Dublin Core and Metadata Applications* (DC 08), Universitätsverlag Göttingen, 2008, pp. 43–53.
6. A. Isaac et al., "Putting Ontology Alignment in Context: Usage Scenarios, Deployment and Evaluation in a Library Case," *The Semantic Web: Research and Applications, Proc. 5th European Semantic Web Conf.* (ESWC 08), LNCS 5021, Springer, 2008, pp. 402–417.
7. M. Hearst et al., "Finding the Flow in Web Site Search," *Comm. ACM*, vol. 45, no. 9, 2002, pp. 42–49.
8. J. Euzenat, "Semantic Precision and Recall for Ontology Alignment Evaluation," *Proc. 20th Int'l Joint Conf. Artificial Intelligence* (IJCAI 07), Springer, 2007, pp. 348–353.
9. B. He and K.C.C. Chang, "Automatic Complex Schema Matching across Web Query Interfaces: A Correlation Mining Approach," *ACM Trans. Database Systems*, vol. 31, no. 1, 2006, pp. 346–395.
10. F. Giunchiglia, M. Yatskevich, and P. Shvaiko, "Semantic Matching: Algorithms and Implementation," *J. Data Semantics IX*, LNCS 4601, Springer, 2007, pp. 1–38.

For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.

Call for Articles

IEEE Software seeks practical, readable articles that will appeal to experts and nonexperts alike. The magazine aims to deliver reliable information to software developers and managers to help them stay on top of rapid technology change.

IEEE Software

Author guidelines: www.computer.org/software/author.htm
Further details: software@computer.org
www.computer.org/software