# Evaluating Thesaurus Alignments for Semantic Interoperability in the Library Domain

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

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

**Abstract.** Thesaurus alignments play an important role in realising efficient access to heterogeneous Cultural Heritage data. Current technology, however, provides only limited value for such access as it fails to bridge the gap between theoretical study and user needs that stem from practical application requirements. In this paper, we explore common real-world problems of a library, and identify solutions that would greatly benefit from a more application embedded study, development, and evaluation of matching technology.

*keywords* thesaurus alignments, ontology matching, (application-specific) evaluation, libraries

## 1 Introduction

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

By providing representational standards (such as SKOS[5]) and generic tool support, the Semantic Web community has recently 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]. Alignments can be exploited, 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 rationalise thesaurus maintenance.

[4] See for instance the Europeana portal, `http://www.europeana.eu`.
[5] See `http://www.w3.org/2004/02/skos`.

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

*Method.* This paper explores a selection of common real-world problems of libraries to manage and give access to their collections and thesauri. For each problem, we identify the beneficial use of alignments to facilitate the problem's solution. In doing so, we not only identify application specific requirements and users' needs, but also show how resulting solutions can (and should) be evaluated. For a good part of the problems, we give solutions, evaluate them, and thus show the strengths and weaknesses of existing matching technology.

*Investigations.* The use of real data at industrial scale (two thesauri with many thousands concepts each; two $10^6$ book collections indexed with these thesauri) allows us to investigate (and quantify) for each scenario

- the interoperability problem and the potential of matching technology to contribute to its solution; and
- the evaluation of alignments (and thus of the tools producing them) in the proper deployment contexts.

In particular, we aim at determining the variability in requirements across scenarios, for instance, with regard to supported mapping link types (*e.g.*, concept equivalence, hierachical links), aligning single concepts $(1:1)$ *vs.* groups of concepts $(m:n)$, and performance related measures (precision, recall).

*Experiments.* We take the OAEI 2007 Library Track as our evaluation context and evaluate the alignments from all three participant tools. We especially compare their performances against specific requirements from different scenarios.

*Findings.* Our expectations, outlined in the Library Track webpage, were not met. None of the participating systems returned hierarchical links, and the suggestion to deliver $m:n$ mappings was not followed either. The tools' generic design prevented us from using them entirely off-the-shelf; their alignments needed to be carefully interpreted and post-processed to fit the problems at hand. With regard to system performance, precision and recall leave much to be improved. Moreover, our various evaluations also confirmed that the way evaluations are conducted – within or across usage scenarios – clearly affects the results, and thus the quality judgement of existing matching technology. In sum, we believe that tool developers need to have a better knowledge of realistic application domains and their requirements to develop, evaluate and optimize their technology.

*Scientific and Practical Relevance.* This paper makes two contributions. First, we inform the Semantic Web community of realistic scenarios,[6] which, although stemming from a specific library context, formalize and generalize on previous practitioners' efforts. Second, we provide the Library Sciences community with hands-on methodological guidelines for the application and evaluation of ontology matching technology.

We expect our work to be highly relevant for practitioners within the wider Cultural Heritage sector, but also for other communities that manage large sets of artifacts with multiple description systems. In fact our work can have direct impact in other fields such as biology, e-commerce, or e-health in which expert knowledge is modeled using controlled vocabularies, as it opens a problem-driven approach to Semantic Interoperability that takes the application contexts, and thus human factors, into account.

## 2 Background

### 2.1 Ontology Matching

Ontology matching aims at determining the semantic relations that hold between the elements of two given knowledge organization systems (*e.g.*, ontologies, thesauri) [1]. The set of semantic relations usually comprises concept equivalence, hierarchical concepts links (*broader/narrower than*) and mere relatedness links. Various research projects in the CH context have tackled the matching task *manually*, most notably, MACS.[7] These projects demonstrated the complexity and cost of manually aligning large vocabularies in realistic collections, and thus the need for computer assistance.

There are now tools, being developed within the Semantic Web community, that address the problem of ontology matching.[8] Their complex machinery can be decomposed 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.[9]

*Ontology Matching Tools* normally take an application neutral perspective on the matching problem. They typically apply one or several of the above techniques to compute similarity between concepts; the resulting *alignments* consist of *mappings*, which in turn relate *entities* via semantic *relations*, with sometimes a *measure* attached. However, there are various degrees of freedom: entities can be simple concepts or complex constructions of several concepts; one entity can

[6] This paper is a major revision and extension of previous work [3]. It clarifies problem descriptions, adds the search scenario to the list of common library problems, and extends evaluation as well as cross-scenario comparison.

[7] See `http://macs.cenl.org`.

[8] For an overview of individual tools, see Ch. 6 of [1].

[9] See `http://wordnet.princeton.edu`.

be involved in only one or any number of mappings; the type of mapping relations can range from vaguely to formally specified; and the measure may denote a probability for the mapping to hold or an objectively measurable similarity degree. Decisions on these options influence the quality and usability of the alignments in given application contexts.

*Evaluation Approaches.* Since 2004, the Ontology Alignment Evaluation Initiative organises campaigns to review the performance of current state-of-the-art matching technologies in different domains. As most tools are still highly experimental and not used in practical applications, the first evaluation efforts favoured mostly "application-independent" methods, as in [4]. Typically, manually built *reference alignments* (or *gold standards*) were created and used, and consequently, often biased towards – at best – one single usage scenario (*e.g.*, vocabulary merging), with little use for other scenarios. Recently, new evaluation approaches adopt more realistic assessments by using, for instance, application specific sampling methods and measures [5, 6]. Further work, however, is needed to better understand real-world use cases, their requirements, and the proper use and evaluation of matching technology.

In this paper, we take the OAEI 2007 Library Track [3] as our evaluation context. Three participating systems, **Falcon**, **DSSim** and **Silas** were given realistic data and required to focus on SKOS mapping relations.[10] The resulting alignments were then evaluated in the context of KB scenarios.

### 2.2 The need for thesaurus alignment at KB

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

The two thesauri are both in Dutch and have similar coverage but differ in granularity. Represented in SKOS, each concept has one preferred label, synonyms and other alternative labels, extra hidden labels and scope notes. Both thesauri are structured by *broader*, *narrower* and *related* relations between concepts, but this structural information is relatively poor. The average depths of the GTT and Brinkman concepts are 0.69 and 1.03 respectively, and nearly 20K GTT concepts have no parents.

[10] In particular, `exactMatch` for equivalence mappings, `broadMatch` for generic-specific relationships, and `relatedMatch` for simple associations.

The co-existence of these different thesauri, even if historically and practically justified, is not satisfactory. First, both thesauri are actively but independently maintained, which doubles 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 250K dually indexed books. To achieve better interoperability and reduce costs, matching technology plays a crucial role, with regard to the following scenarios:

1. **Reindexing:** support the indexing of GTT indexed books with Brinkman concepts, or *vice versa*.
2. **Concept based search across vocabularies:** support the retrieval of GTT indexed books using Brinkman concepts, or *vice versa*.
3. **Navigation across thesauri:** support the exploration of concept spaces across thesauri, and give (exploratory) access to collection items indexed with selected concepts.
4. **Thesaurus Merging:** support 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 creating new information to enhance such access, or to improve the library processes themselves (back-office).

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

## 3 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 were given only in abstract terms by the ontology matching community [1].

### 3.1 The book reindexing scenario

To streamline the indexing of books currently described with both Brinkman and GTT thesauri, KB is considering computer supported reindexing methods. The conservative approach consists of maintaining both thesauri. A (newly acquired)

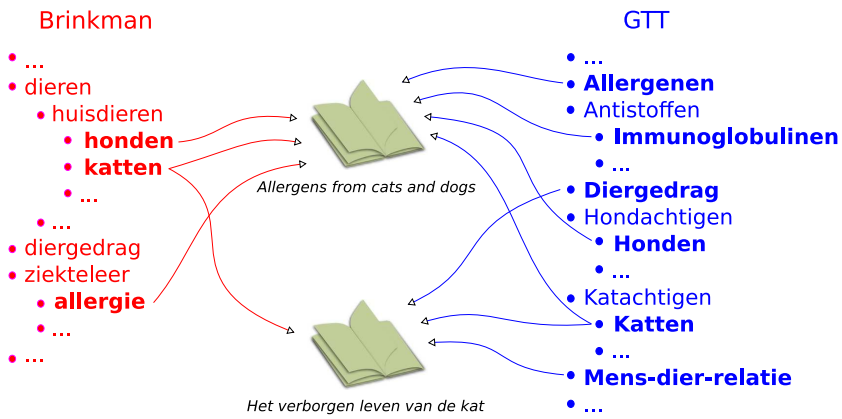


**Fig. 1.** Indexing books using two thesauri: GTT and Brinkman

book is first indexed with GTT by a human expert; given its GTT annotation, a tool shall automatically generate 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 shall be reindexed with adequate Brinkman concepts. This *data migration* for information integration [1] could be fully mechanised or supervised.

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

Moreover, the correct reindexing requires more than the identification of one-to-one mappings, as shown in the second book example, *i.e.*, three GTT concepts *vs.* one Brinkman concept. Clearly, the librarians' annotation reflects diverging analysis levels, or even thesaurus-inherent indexing policies. These examples also highlight the issue of *post-coordinate 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[11] that translates the concepts of a GTT indexed book into Brinkman concepts to yield a Brinkman

[11] Here, we take an idealized and pragmatic stance. KB's corpus of dually annotated books contains cases where two books with identical GTT annotation have different

indexing of the book:

$$f_r : 2^{\mathcal{G}} \rightarrow 2^{\mathcal{B}},$$

where $2^{\mathcal{G}}$ and $2^{\mathcal{B}}$ denote the powersets of GTT and Brinkman concepts. Ideally, the GTT index of any given book should be translated into a semantically equivalent or similar Brinkman index. If the latter has a broader (narrower) semantics, then the translation is not information preserving, but instead loses (gains) information.

**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 for 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 250K 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 $f_r$.

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, results are disappointing [3]. For the best systems, nearly half of the generated Brinkman concepts were incorrect (*precision* = 54% for Silas without `relatedMatch`). Recall is weak as well, as more than 60% of the gold standard was not found (*recall* = 39% for Silas with `relatedMatch`).

*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 book 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 *etc.*), experts were then asked to judge the *acceptability* of each proposed

Brinkman annotations; instance-based methods to learn a function rather than a relation from this data are thus doomed to a certain extent of inaccuracy.

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 assessed independently each of the 502 concepts proposed.

Their assessment yields significant better values for both precision and recall; for instance, Falcon's precision and recall respectively improve from 53% to 75% and from 36% to 46%. This indicates that human experts are more likely to accept semantically close Brinkman concepts. Instead of testing for strict set equality, they applied less stricht notions for correctness and completeness. This confirms the subjective nature of the indexing task. Using Krippendorff's $\alpha$ 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. This aspect shall be considered carefully when exploiting (and tuning) alignments for this scenario.

### 3.2 Book search across collections

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

1. to search for a particular book, a librarian formulates a query that consists of a set of, say, GTT concepts;
2. the query's GTT concepts are translated into Brinkman concepts; and
3. both queries are executed by the search engine, and the results of both queries are given to the librarian.

Conversely, a librarian may want to formulate a query in terms of Brinkman to get access to books that are annotated with GTT concepts only. 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, the task is to specify a function that translates any member of the powerset of GTT concepts to some set of Brinkman concepts which is then passed onto a search routine to retrieve the book(s) in question:

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

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 most simple query reformulation takes each concept $g_i$ in the search query $g_1 \dots g_j$, and searches the alignment for a single semantically equal Brinkman concept $b_i$, yielding a reformulated query $b_1 \dots b_j$ (when for each concept $g_i$ such a concept $b_i$ is found). If no corresponding Brinkman concept is found we obtain a reformulated query $b_1 \dots b_m$ with $m < j$.

**Evaluation Method.** An evaluation of this scenario would profit from the existence of a representative set of search queries, possibly in combination with information whether the search results obtained were further used by users (say, by clicking on them). Unfortunately, we have neither KB's log of concept based searches nor any information on whether their results were 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 labour 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 $\mathcal{G}_i$ be the set of existing GTT concepts of the book, $\mathcal{B}_i^*$ the existing Brinkman concepts of the book, and $\mathcal{B}_i$ the predicted Brinkman concepts.We then evaluate the search queries for $Q_{\mathcal{G}_i}, Q_{\mathcal{B}_i^*}$ and $Q_{\mathcal{B}_i}$, and compare their answer sets, where $Q_{\mathcal{G}_i}$ denotes the set of books annotated with *all* concepts in $\mathcal{G}_i$, and similarly for $Q_{\mathcal{B}_i^*}$ and $Q_{\mathcal{B}_i}$.

Note that this setup makes three assumptions: that the difference between Brinkman and GTT indexing policy 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. It is clear that all three assumptions give an idealised view that is rarely found in library practise.

In this setup, we have different definitions for precision and recall. Instead of computing the intersections between annotations, we compute the normalised overlap of the answer sets $Q_{\mathcal{B}_i^*}$ and $Q_{\mathcal{B}_i}$ 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 $Q_{\mathcal{G}_i}$ and $Q_{\mathcal{B}_i^*}$ return book $i$, and potentially other books that share the same Brinkman and GTT annotations. Clearly, $i \in Q_{\mathcal{B}_i}$ if query formulation succeeds in translating at least one GTT concept of $g_i$ into a Brinkman concept $b_j \in \mathcal{B}_i^*$, *and* there is no single incorrect mapping from one of GTT indices to a Brinkman one. Moreover, it is also 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, 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 might 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 of the GTT concepts correctly to one Brinkman concept, then search will fail as it overspecifies the book in the Brinkman context.

*Automatic evaluation* has been performed on all books of the dually indexed corpus using again the three OAEI participants. **Falcon** and **Silas** perform comparably with a precision of around 36%, recall of 33% and an overlap of retrieved books of 19% using Jaccard. The third participant, DSSim, performed significantly worse (P:9%, R:7%). On average, a book's given Brinkman annotation consists of 1.65 concepts while 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.

*Discussion.* Given the size of the answer sets, we may expect $f_s$ and the search engine to optimise 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 generalise a given GTT concept by following the structural links within the GTT thesaurus, and then subject the concept's generalisation to query reformulation. These strategies are discussed in the navigation scenario below.

### 3.3 Navigating with multiple vocabularies

KB may consider *faceted browsing* functionalities that allow users to easily access multiple collections, which are possibly classified along various vocabularies [7]. With a *single view*, access to multiple collections is given via a single thesaurus; with a *combined view* access to multiple collections is given through their respective vocabularies, allowing users to browse through the integrated collections as if they were a single collection indexed against two complementary points of view; and with a *merged view* access to multiple collections is given 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 a user's browsing activity through collections. Users can refine, extend or change the items in focus by the selection of more general or more specific concepts, or by changing from one concept to another (related) one. It attracts users with limited expertise in formulating search queries (see search scenario), and users with the objective of exploring collections along several dimensions (facets) rather than quickly finding a specific collection item of interest. To support the exploratory character, faceted browsers often expand user requests, *i.e.*, when a concept is selected, all items that are indexed by the concept's specialisation or generalisation are also retrieved.

It is clear that 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/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 of 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 the fact that 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 be added to this selection. Formally, navigation in the KB case thus is a function:

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

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 to another vocabulary. In practise, this search must be complemented with an adequate graphical user interface that gives access to a click based selection of search queries, term generalisation and refinement. For the merged view, processes similar to the ones discussed in the thesaurus merge scenario will be needed (see Sect. 3.4).

**Evaluation method.** While there is an automatic evaluation for search and re-indexing, evaluation in the navigation context requires human users, and thus realistic end-to-end evaluation settings [5]. 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 *etc.*) on the overall navigation experience makes any evaluation hard; a further discussion falls outside the scope of this paper.

### 3.4 The thesaurus merging/integration scenario

To reduce thesaurus management and indexing costs, KB considers to merge 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*.

On a second glance, 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 reorganised for the sake of hierarchical balance; potential merging conflicts may arise and need to be resolved; and last but not least, legacy data shall still be easily accessible via the resulting thesaurus.

Thesaurus merging is thus a complex cognitive endeavour where matching technology can only play a supporting role, namely, at the local level by suggesting inter-thesaurus 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. 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 specialisations, and thus *broader*/*narrower* mappings are of particular interest. Associative *related* mappings may also be considered useful, even though less crucial.

*Thesaurus merging* of two input thesauri into a third thesaurus is different, as the output thesaurus can be dramatically different from the two input. Here, the alignment, in combination with intra-structural thesaurus information, can be used to help grouping together related concepts, interlinked via *equivalent*, *broader*, *narrower* or *related* relations, into small clusters. A thesaurus manager can then reorganise the concepts and relationships within clusters into coherent hierarchies. Gradually, smaller hierarchies can be composed to larger ones, and finally yield a unified thesaurus.

**Problem Statement.** Specify a function that merges the two input thesauri $\mathcal{G}$ and $\mathcal{B}$, each with their concepts and interrelations, into a unified thesaurus $\mathcal{GB}$:

$$f_m : \mathcal{T} \times \mathcal{T} \rightarrow \mathcal{T},$$

where $\mathcal{T}$ 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 the alignment.

The thesaurus merging process can be supported 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.** Two aspects should be evaluated:

- Correctness/precision: the proportion of followed suggestions over all suggestions;
- 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 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 with having 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 is directly related to the main role of an alignment for thesaurus merging.

*Evaluation.* During the OAEI 2007, we evaluated the three participants in the setting of the merging scenario [3]. By comparing concept labels (literal string matching), and by exploiting a Dutch morphology database to recognise word variants (*e.g.*, *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 which were not in the original reference alignment and manually evaluated an additional 330 mappings.

Clearly, our reference alignment has a very strong bias towards the lexical equivalence mappings. This explains the high precision of Falcon (97.3%), as almost all its mappings are lexically equivalent pairs. Silas, which performed similarly well to what Falcon did in the reindexing scenario, has a lower coverage of the reference alignment (66.1%, compared to 87.0% 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% (compared to 13.4% for DSSim).

## 4 Comparison and discussion

This section recapitulates and compares the various scenarios in terms of the requirements they impose and the evaluation forms they suggest. Tab. 1 provides an overview to this discussion.

| | | Scenarios | | | |
|---|---|---|---|---|---|
| | | **Reindexing** | **Search** | **Navigation** | **Merging** |
| Requirements | hierarchical links | soft<br>*to recover from*<br>*lack of equality links* | soft<br>*to broaden or*<br>*restrict search* | hard<br>*to guide navigation*<br>*across thesauri* | hard<br>*to (add) structure*<br>*(to) new thesaurus* |
| | m:n links | hard<br>*for post-coordination* | hard<br>*for post-coordination* | soft | soft |
| | coverage | active concepts<br>in indexing | active concepts<br>in search & indexing | active concepts<br>in search & indexing | all concepts<br>of input thesaurus/i |
| | best method | extensional | extensional | ? | ? |
| Evaluation | object | book annotations<br>*orig. vs. new index* | query results<br>*orig. vs. new query* | alignment<br>GUI | alignment<br>*links used* |
| | form | automatic<br>*given dual corpus*<br>manual<br>*focus on human factor* | automatic<br>*given dual corpus*<br>manual<br>*focus on human factor* | manual<br>*user experience*<br>*efficacy* | manual |

**Table 1.** Overview of Requirements and Evaluation Aspects

### 4.1 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 by a combination of a lexical approach with an instance-based approach. Although Falcon is the best system overall, the relative performance with 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 correspondances between lexically equivalent (or similar) concepts; its structural component could little contribute, but this is due to the *low* structural similarity between the GTT and the Brinkman thesauri. In comparison, however, DSSim came as 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 mislead by this data, but its lexical module could recover partly 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. Here, concept equivalence has other than lexical or structural roots. In the book reindexing scenario, concept equivalence is measured in terms of their extensional overlap, *i.e.*, in the intersection of the books they index. In our search scenario, concepts are considered equivalent if their use in the search query return 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.

### 4.2 Semantics of required alignments

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

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 reorganise 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, while a thesaurus engineer would rather search for more intensional equivalence mappings for his task – which actually questions the relevance of using one single equivalence relation in both situations.

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 (layman) users cannot easily deal with.

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

### 4.3 Mapping Cardinality

The requirement for many-to-many mappings is scenario dependent. For thesaurus merging, concept combinations can help a thesaurus engineer determining 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 specialisation of both "geography" and "the Netherlands" in the integrated or merged thesaurus.

In the book search scenario, post-coordination 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.

### 4.4 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 shall 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 less than ten times; and as much as 11,134 of GTT terms are not used at all.

An alignment, however, shall have mappings between frequently used concepts. Alignments provided by the OAEI participants cover 51–85% of Brinkman concepts but only 10–26% of GTT concepts. Clearly, such alignments will provide little exploitable material for thesaurus merging. However, 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% of recall in the manual evaluation, which exceeds expectations, given the low overall coverage.

### 4.5 Precision and recall needs

Ideally, matching technology should optimise 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 (*e.g.*, the search scenario) or only computer supported (reindexing, thesaurus merging). Scenarios that rely on humans involvement, (*e.g.*, 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.

## 5 Conclusion

In this paper, we described four real-world problems that the KB, but also CH institutions in general, face: how to support the reindexing of books (or other

CH objects) from one vocabulary to another, the search for books or objects across collections, the simultaneous navigation through multiple collections, and the merging of two or more thesauri?

Ontology matching technology can play a major role in giving uniform access to heterogeneously described CH collections. The tools participating 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 post-process 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 hereby encourage tool developers to implement such functionality, or to finetune such functionality where it is already present (see [9, 10]), and to also participate in our future OAEI tracks to evaluate their system's performance.

Our study also indicated application contexts favouring 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 paper guides researchers from the Semantic Web community to better take into account real-world thesauri and their use in realistic application contexts, thus better balancing their generality design imperative with what is needed in practise 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 paper encourages and guides the Library Sciences community in adopting matching technology. There is much complexity to be dealt with, and the human factor of technology use should not be ignored. But the benefits for librarians and end-users alike are worth it.

## Acknowledgements

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

## References

1. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer (2007)

2. van Gendt, M., Isaac, A., van der Meij, L., Schlobach, S.: Semantic web techniques for multiple views on heterogeneous collections: a case study. In: Proc. 10th European Conf. Research and Advanced Technology for Digital Libraries (ECDL). Volume 4172 of LNCS., Springer (2006)
3. Isaac, A., Matthezing, H., van der Meij, L., Schlobach, S., Wang, S., Zinn, C.: Putting ontology alignment in context: usage scenarios, deployment and evaluation in a library case. In: The Semantic Web: Research and Applications, Proc. 5th European Semantic Web Conf. (ESWC). Volume 5021 of LNCS., Springer (2008)
4. Zhang, S., Bodenreider, O.: Experience in Aligning Anatomical Ontologies. Int'l J. Semantic Web & Information Systems **3**(2) (2008)
5. Hollink, L., van Assem, M., Isaac, A., Wang, S., Schreiber, G.: Two variations on ontology alignment evaluation: Methodological issues. In: The Semantic Web: Research and Applications, Proc. 5th European Semantic Web Conf. (ESWC). Volume 5021 of LNCS., Springer (2008)
6. Lauser, B., Johannsen, G., Caracciolo, C., Keizer, J., van Hage, W., Mayr, P.: Comparing human and automatic thesaurus mapping approaches in the agricultural domain. In: Proc. Int'l Conf. on Dublin Core and Metadata Applications, Universitätsverlag Göttingen (2008)
7. Hearst, M., et al.: Finding the Flow in Web Site Search. Communications of the ACM **45**(9) (2002)
8. Euzenat, J.: Semantic Precision and Recall for Ontology Alignment Evaluation. In: Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI). (2007)
9. He, B., Chang, K.C.C.: Automatic Complex Schema Matching across Web Query Interfaces: A Correlation Mining Approach. ACM Trans. Database Systems **31**(1) (2006)
10. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic Matching: Algorithms and Implementation. Journal on Data Semantics IX (2007)

## Authors' information

*Antoine Isaac* obtained in 2005 a computer science PhD. from the University of Paris-Sorbonne for research on the design and use of ontologies in INA, the French National Institute for Audiovisual archives. He is now a researcher at the Vrije Universiteit Amsterdam, focusing on the representation and the interoperability of Cultural Heritage collections and their vocabularies. In particular, he works the use of Semantic Web techniques for representing and aligning Knowledge Organization Systems. He participates in the W3C Semantic Web Deployment working group and is involved in the design of SKOS.

mail address: Vrije Universiteit Amsterdam, Department of Computer Science, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands
phone: +31 (0)20 598 7449
fax: +31 (0)20 598 7653
email: aisaac@few.vu.nl
web: http://www.few.vu.nl/∼aisaac

*Shenghui Wang* is currently a researcher in the Department of Computer Science, Vrije Universiteit Amsterdam. She received her PhD degree in 2007 from

the School of Computer Science of the University of Manchester. Her PhD study focused on the semantics of natural language descriptions of continuous quantities. She is now working on the problem of semantic interoperability in the Cultural Heritage domain, including matching different thesauri, deploying the mappings in various interoperability applications.

mail address: Vrije Universiteit Amsterdam, Department of Computer Science, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands
phone: +31 (0)20 598 7452
fax: +31 (0)20 598 7653
email: swang@few.vu.nl
web: http://www.few.vu.nl/∼swang

*Claus Zinn* is an R&D engineer at the Max Planck Institute of Psycholinguistics (MPI), Nijmegen, The Netherlands. He holds an MSc. and PhD. degree in Computer Science (University of Erlangen-Nuremberg, Germany). He has held positions as Research Associate (later Research Fellow) at the School of Informatics, University of Edinburgh, and as Senior Researcher at the German Research Centre for Artificial Intelligence (DFKI). His interests cover many areas of Artificial Intelligence, in particular, Automated Reasoning, Dialogue Management, and Knowledge Representation. At the MPI, he researches and develops methods and tools to support ontology management and the analysis of language corpora.

mail address: Max-Planck Institute for Psycholinguistics, Wundtlaan 1, PB 310, 6500 AH Nijmegen, The Netherlands
phone: +31 (0)24 352 1473
email: Claus.Zinn@mpi.nl
web: http://www.mpi.nl/Members/ClausZinn

*Henk Matthezing* is project leader at the Research and Development Department in the Koninklijke Bibliotheek, The National Library of the Netherlands. He has been involved in numerous library internet projects, a.o. introduction of Dublin Core in the Netherlands, the experimental stages of KB's e-Depot, Digital Preservation, various digitisation projects and lately in converting collection metadata sets into an open access environment and semantic interoperability to enhance the access to Digital Cultural Heritage Collections.

mail address: Koninklijke Bibliotheek, Prins Willem-Alexanderhof 5, 2509LK Den Haag, The Netherlands
phone: +31 (0)70 314 0687
fax: +31 (0)70 314 0427
email: Henk.Matthezing@kb.nl

*Lourens van der Meij* is a scientific programmer at the Knowledge Representation and Reasoning Group at the Vrije Universiteit Amsterdam. He holds MSc. degrees in Computer Science (Delft University) and Physics (Leiden University). His current research focus is application of semantic web techniques in the Cultural Heritage domain. He has been involved in various projects in the AI

domain and has developed software environments for multi agent systems and simulation.

mail address: Vrije Universiteit Amsterdam, Department of Computer Science, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands
phone: +31 (0)20 598 7449
fax: +31 (0)20 598 763
email: lourens@cs.vu.nl
web: http://www.few.vu.nl/∼lourens

*Stefan Schlobach* is Assistant Professor for Artificial Intelligence at the Vrije Universiteit Amsterdam. He studied Computer Science and Philosophy in Paris and Saarbruecken, and received a PhD from King's College London for research on combining knowledge representation and learning techniques. His main research areas are Semantic Interoperability in the fields of Cultural Heritage and Health, as well as non-standard reasoning, in particular the study of new methods for symbolic approaches to uncertainty, approximation and any-time behaviour in large ontologies.

mail address: Vrije Universiteit Amsterdam, Department of Computer Science, De Boelelaan 1081a, 1081HV Amsterdam, The Netherlands
phone: +31 (0)20 598 7678
fax: +31 (0)20 598 7653
email: schlobac@few.vu.nl
web: http://www.few.vu.nl/∼schlobac