Instance-based Semantic Interoperability in the Cultural Heritage

Editor(s): Dimitrios A. Koutsomitropoulos, University of Patras, Greece; Eero Hyvönen, Aalto University, Finland; and Theodore S. Papatheodorou, University of Patras, Greece
Solicited review(s): Giorgos Stoilos, Department of Computer Science, University of Oxford, UK; Christophe Dupriez, DESTIN inc. SSEB, Belgium; and Achille Felicetti, PIN, University of Florence, Italy

Shenghui Wang [a,*], Antoine Isaac [b,a], Stefan Schlobach [a], Lourens van der Meij [a] and Balthasar Schopman [a]
[a] *Vrije Universiteit Amsterdam, De Boelelaan 1081a, 1081HV Amsterdam, NL*
E-mail: {swang,aisaac,schlobac,lourens,baschopm}@few.vu.nl
[b] *Koninklijke Bibliotheek, Prins Willem-Alexanderhof 5, 2595BE Den Haag, NL*

Abstract. This paper gives a comprehensive overview over the problem of Semantic Interoperability in the Cultural Heritage domain, with a particular focus on solutions centered around extensional, *i.e.*, instance-based, ontology matching methods. It presents three typical scenarios requiring interoperability, one with homogeneous collections, one with heterogeneous collections, and one with multi-lingual collections. It discusses two different ways to evaluate potential alignments, one based on the application of re-indexing, one using a reference alignment. To these scenarios we apply extensional matching with different similarity measures, which gives interesting insight into the applicability of this matching approach.

Finally, we firmly position our work in the Cultural Heritage context through an extensive discussion of the relevance for, and issues related to this specific field. The findings are as unspectacular as expected, but nevertheless important: the provided methods can really improve interoperability in a number of important cases, but they are not universal solutions to all related problems.

This paper provides a solid foundation for any future work on Semantic Interoperability in the Cultural Heritage domain, in particular for anybody intending to apply extensional methods.

Keywords: semantic interoperability, ontology matching, instance-based methods, cultural heritage

1. Introduction

With more and more data being published, the need grows for interlinking, reusing and increasing accessibility of data in non-purpose built applications. This is commonly referred to as the Semantic Interoperability problem, which has become a core topic of Semantic Web research. The dynamic community that works on ontology matching is proof of this.

Ontology matching in general is the task of linking resources from different knowledge organisation schemes. The Cultural Heritage (CH) domain always had a special role in ontology matching, which is mainly due to two facts: the multitude of annotated artefacts and shallow vocabularies. More concretely,

1. CH collections usually come with **huge amounts of semantically annotated resources**, such as books, multi-media objects (movies, music, images, *etc.*) and other tangible objects such as paintings, sculptures, *etc*. Common to all of those is that they are systematically catalogued and formally described by means of controlled vocabularies.

*Corresponding author: swang@few.vu.nl

2. Those vocabularies[1] are often historically developed over decades with the purpose of storing, finding and accessing a specific collection of objects. In fact, those thesauri are purposely built to describe the objects, and not the world these objects refer to. As a consequence, the **semantics of thesauri differs** from the semantics of the ontologies usually considered in ontology matching, and therefore, so do the required matching methods.

These facts make the task of Semantic Interoperability in CH a particular subfield of ontology matching with very specific properties to which this paper is devoted.

Semantic Interoperability in Cultural Heritage Controlled knowledge organisation systems, such as thesauri or subject heading lists (SHLs), are often used to describe collection objects. These vocabularies, specified at the semantic level using dedicated relations—typically *broader*, *narrower* and *related*— can be of help when accessing collections, *e.g.*, for guiding a user through a hierarchy of subjects, or performing automatic query reformulation to retrieve more results for a given query.

However, nearly every CH institution uses its own *subject indexing system*, in its own natural language. It is therefore impossible to exploit the semantically rich information of multiple controlled vocabularies simultaneously. This greatly hinders access to, and usability of the content of CH organisations such as The European Library (TEL)[2], which aims at providing unified access to European national libraries. A solution to this specific issue is the semantic linking (or *matching*) of the concepts in vocabularies. Individual links (or *mappings*) between concepts generally specify equivalence at the semantic level and can, *e.g.*, be used to *reformulate* a query from one language to the other. For example, an equivalence link between `Sprinting`, `Course de vitesse` and `Kurzstreckenlauf` will allow to transform a query for sprints, that would only give results in the British Library catalogue, into equivalent queries that have matching results in the French and German collections, respectively.

While large efforts are taken to create *alignments* (set of mappings) manually, a crucial problem in this approach is its enormous cost. Some experience reports mention that around 90 terms may be matched per day [54] by a skilled information professional dealing with concepts in a same language. The vocabularies in CH often contain hundreds of thousands of concepts. Clearly, automatic methods are needed.

For the purpose of this paper, Semantic Interoperability problems in CH can be classified into three classes, homogeneous, heterogeneous and heterogeneous multi-lingual. For each of the three classes we discuss a particular problem that we encountered within our research.

1. Collections of books that are annotated with two thesauri that are to be matched.
2. A collection of books and a "multi-media" collection, both of which have been annotated with their own thesauri.
3. The above mentioned problem of integrating collections of books from libraries of different European nations. Each library annotates the books in their collections with different thesauri in different languages, and with different metadata schemas.

The first application scenario stems from the Dutch National Library (KB) which requires mappings between two thesauri both used to annotate two *homogeneous* book collections. The second scenario is related to supporting integrated online access of parts of the collections of the Dutch institute of Sound and Vision (B&G) and the KB, *i.e.*, a mapping between two thesauri that are used to describe *heterogeneous* collections. The final scenario stems from our involvement in The European Library project, *a multi-lingual case*.

Extensional methods for Semantic Interoperability Extensional methods for Semantic Interoperability are based on the assumption that similarity of the objects annotated by two concepts determines the similarity between those concepts. Technically, those methods usually apply measures based on the overlap of instances of two concepts.

Extensional methods have a number of important benefits. As opposed to *structure-based* methods, extensional methods do not depend on a rich ontology structure; this is important in the case of thesauri, which often have a weak, and sometimes even almost flat structure. Contrarily to *lexical* methods, they do not depend on the concept labels, which is particularly important when the ontologies or thesauri are written in multiple languages. Extensional techniques are thus likely to provide a useful complement to lexical tech-

[1]For the sake of simplicity, we use the term **thesaurus** to refer to (different types of) such knowledge organisation schemes, including subject heading lists and classification systems.

[2]`http://www.theeuropeanlibrary.org`

niques, by focusing on an element of the semantics of concepts: the way they are used in actual descriptions. For example, finding the equivalence between `Kurzstreckenlauf` and `Course de vitesse` by using lexical methods would require a sophisticated translation service and/or a domain-specific lexical base. Such resource may not be available or be hard to re-use for a specific matching task. A significant amount of books that are all annotated with both Concept A and Concept B provides a hint of a semantic relation between the concepts A and B.

However, measuring the common extension of concepts requires the existence of a sufficient amount of shared instances, something which is often not the case. Furthermore, it only uses part of the available information, *i.e.*, ignores similarity between instances that have not been doubly annotated. Similarity on the instance-level is often ignored. In this paper we also apply a more general *similarity-based extension comparison*, deriving concept mappings from similarity of their instances.

STITCH In order to stimulate the use of advanced information and communication technologies in CH, the Dutch government has been funding a large research program called CATCH (Continuous Access to Cultural Heritage)[3] in which one of the first projects (STITCH, 2006-2010) was dedicated to Sematic Interoperability. In STITCH, we investigated the use of ontology matching technology in CH and established the exceptional position of **extensional methods** (also called **instance-based**) for concept matching in CH. Since 2007 we have published technical papers studying these extensional methods in detail: in [21] a number of extensional similarity measures were compared and evaluated. [34] extended this approach to collections without joint extensions by matching instances first to create sets of "fake" joint instances. Both these approaches make **direct** use of the individual instance annotated by a concept. In [50,49], we investigated the use of **aggregation** of instances, and applying Machine Learning methods to select the most important features of these aggregations for determining concept similarity. In two other lines of research we studied the representational issues of vocabularies in CH in ontology languages such as SKOS [20], and finally, how to evaluate such Semantic Interoperability. Our findings in [22] showed how much the meaning of these semantic links is determined by the usage scenarios, and how this knowledge should be used in evaluating matching results. An important contribution throughout the first effort was to help make available thesauri from the CH to the Linked Open Data[4] collection, such as Rameau, LCSH and others. Today, this knowledge from CH has become an integral part of the Semantic Web. Linking those data sources becomes an even more crucial issue. What is still missing is a general overview over methods for achieving Semantic Interoperability specific to the CH domain.

Given our experience with extensional methods, this paper attempts to close this gap for the rich set of methods based on instance data. Our insights from several years of (mostly technical) research in Semantic Interoperability in the CH domain indicates that extensional methods can (1) generate mappings that are complementary to mappings that can be generated using traditional, lexical matching methods and can (2) viably contribute to the linking and combining access to CH collections within and across institutions. We give an overview of approaches that can have a direct practical impact to practitioners. We therefore focus on rather simple methods directly using instance information and will not go into aggregation based methods.

Evaluation methods and results We have run the implemented extensional methods in the three above-mentioned scenarios. Each of these three scenarios are evaluated in two ways: using expert mappings as gold-standards, and using a re-indexing scenario that better reflects the typical usage of thesauri in CH.

The results from the three scenarios suggest that extensional methods can provide quality mappings which are complementary to those of lexical methods. Particularly, when evaluated in the re-indexing scenario, some measures outperform the lexical method. This shows that extensional methods can be a valid complement to intensional methods, such as lexical-based or structure-based ones. It also confirms our claim that designing mapping methods and evaluating mapping results should take the domain knowledge and application scenarios into account. However, extensional methods have limitations, which we discuss in detail in Section 7.

Contribution and structure of this paper This paper provides an extensive overview over extensional methods to address the Semantic Interoperability problem in CH. It builds on previous work (particularly

[3] `http://www.nwo.nl/catch`

[4] `http://linkeddata.org/`

[34,21,22]) which it combines into one coherent story, and extends in several ways:

1. describes a general framework for applying extensional methods in diverse CH interoperability scenarios,
2. extends and conducts more systematic assessment of the core methods with new measures and better evaluation, and
3. discusses comprehensively the specific properties of Semantic Interoperability in CH and the role of extensional methods.

The paper is structured as follows: after this general introduction, the main body of the work related to instance based matching in CH is discussed in Section 2. Section 3 describes the three different scenarios and their respective datasets. Sections 4 and 5 describe the applied matching methods, and the proposed evaluation framework. Section 6 gives an overview over a number of experiments, Section 7 describes the problems that are specific to interoperability in the CH sector. Finally Section 8 concludes the paper.

2. Related Work

Concept matching in Cultural Heritage This paper focuses on Semantic Interoperability in the Cultural Heritage (CH) domain. There are several justifications for this restriction: 1) this domain presents distinct challenges for ontology matching, 2) for that reason, evaluation requires distinct approaches, 3) its properties make it particularly suitable for instance-based methods and finally 4) our experience in the CH domain allows us share some useful insights in the problem and proposed solutions. Therefore we discuss related work in the context of CH first.

Manual matching by domain experts remains the most common approaches to ontology matching in the CH domain, such as in the MACS[5] [24], Renardus [6] and CrissCross[6] [3] projects. MACS, in particular, is building an extensive set of manual links between three Subject Heading Lists used at the English, French and German national libraries, namely LCSH, Rameau and SWD respectively. KoMoHe [31] is another relevant project that presents an application scenario (search over heterogeneous information systems) for which many manual mappings were made, some of which cross-language.

The STITCH project particularly focused on automated methods for Semantic Interoperability in CH. The main contributions of STITCH with respect to Semantic Interoperability in CH were extensive studies on extensional methods [21,34,50] and the introduction of specific evaluation scenarios, that both provided specific requirements on mapping techniques and evaluation methods. Within the TELplus project,[7] extensional methods were applied to the international, multi-lingual context.

Outside these two projects the specific problems of Semantic Interoperability in CH have attracted far less attention in the ontology matching community. In the three challenges organised by STITCH on matching thesauri within the community-run Ontology Alignment Evaluation Initiative [12,4,11], relatively few systems participated, as many of them were not geared towards the specific properties of the datasets.

The MultimediaN E-Culture project [35] is not just an excellent illustration of the potential of Semantic Interoperability, but has also triggered some relevant technical work in ontology matching on CH thesauri [43,42]. Other recent projects have applied generic methods on datasets from the CH domain, such as [30], but correctly identify the limitation of techniques that do not take specificities of the domain into account, as well as the need for using more appropriate contextual knowledge. The other direction is even more frequent: the application of the "traditional" (*i.e.*, manual) mapping in combination with simple lexical or adhoc matching techniques that are useful in practice, but often not well understood and almost never scientifically reported on. Many linked datasets from the library domain at `http://ckan.net/tag/lld`, or other examples such as the Linked Movie Database[8] contain automatically derived links, but their precise provenance is often unknown to all but the creator of the links. The only major exception is to our knowledge OCLC's work on instance-based mapping [38], and more recently [48].

Instance-based ontology matching The general literature on ontology matching is huge,[9] and we refer to some overviews with many references [8,23,13] by

[5] `http://macs.cenl.org`
[6] `http://linux2.fbi.fh-koeln.de/crisscross/`
[7] `http://www.theeuropeanlibrary.org/telplus/`
[8] `http://www.linkedmdb.org/`
[9] See, *e.g.*, `http://www.ontologymatching.org/publications.html`

Doan, Kalfoglou, and Euzenat *et al*. In this paper we focus on instance-based methods, and only discuss the literature related to instance-based matching.

A common way of judging whether two concepts from different ontologies are semantically linked is to observe their *extensional* information, that is, the instances they classify. A first and straightforward way is to measure the *common extension* of the concepts—the set of objects that are simultaneously classified by both concepts [16,21]. In a survey in 2006 [5] Choi *et al*. reported that 4 out of 9 systems they studied used instance-based methods, namely LSD [7], GLUE [10], MAFRA [28] and FCA-Merge [37]. Many modern systems, such as RiMOM [26], apply combinations of mapping techniques, and often include an instance-based component. This even holds for approaches in rather expressive representation languages [14].

There is, however, to the best of our knowledge, no systematic overview particularly over those extensional methods apart from our preliminary report in [21], neither for the general case of ontology matching, nor for the more specific case of matching in the CH where this approach seems most promising and relevant.

Instance-based methods: variants and extensions The most common approach to extensional matching is using Jaccard-like similarity measures, such as in [25]. Udrea *et al*. [45] uses such measures as a basis, which is later extended with logical inference. Other variants use the DICE similarity [39], or the Jensen-Shannon distance [53].

Zaiss [56] presents two other instance-based matchings methods, one of which is based on aggregation of both the properties and the instances of the concepts that are to be mapped. A similar idea was exploited earlier by Wang *et al*. in [50] where a classifier was trained to classify pairs of source and target concepts into matches and non-matches. Todorov *et al*. in [40] use Support Vector Machines for weighting features of similarities between classes of instances, in [41] they extend this method to the heterogeneous case. Finally, Li uses Neural Networks [27] to similar ends.

Common to these aggregation based methods is that they even work in the absence of jointly annotated instances, which used to be one of the weak characteristics of extensional methods. Schopman *et al*. [34] presents another approach, which first calculates similarity between instances, in order to construct a "fake" doubly annotated corpus.

An alternative use of instance data is given by Avesani *et al*. [2] in the evaluation of mappings that might have been derived with other methods.

Work on schema matching using instance data The database community has put a lot of efforts into schema matching, which corresponds to ontology matching in the Semantic Web context. An overview over schema-based methods is provided in [33]. However, there are significant differences between the two types of problems which make the studies and approaches difficult to compare: databases schemas are usually much smaller than the thesauri we consider (with several thousands of concepts), and instances in CH are usually very well described. This explains why instance-based methods in schema matching have attracted less attention than in concept matching, and will not be discussed further in this paper.

3. Matching scenarios in Cultural Heritage

While the different Cultural Heritage (CH) institutions are opening up their collections in order to achieve better interoperability, matching different thesauri that are used to annotate these collections, within or across CH institutions, is a promising solution. In the following, we present three different application scenarios where instance-based matching methods can play a role.

3.1. Homogeneous collections with multiple thesauri

3.1.1. General description of the scenario

When two collections share the same metadata schema, that is, the objects are described using the same structure, we call them *homogeneous* collections. Collections with objects of a same type in a single CH institution are often homogeneous collections, for example, different book collections in a library. These homogeneous collections are often annotated using different thesauri due to differences in purpose or usage of the collections. Sometimes collections share some features and overlap partially. For end users, access to different collections *via* a unique entry point is more time-efficient than going through each collection separately. In order to improve the interoperability between different collections and allow users to use different thesauri to access collections simultaneously, these thesauri need to be aligned first.

3.1.2. Specific scenario

The National Library of the Netherlands (KB) maintains a large number of book collections. Two of them are the *Deposit Collection*, containing all the Dutch printed publications (one million items), and the *Scientific Collection* mainly about the history, language and culture of the Netherlands. These books have the same metadata structure, which can be partly represented using the Dublin Core metadata standard.[10] The Scientific Collection is described using the GTT,[11] a huge vocabulary containing 35,000 general terms ranging from `Wolkenkrabbers` (Skyscrapers) to `Verzorging` (Care), while the books in the Deposit Collection are mainly indexed against the *Brinkman* thesaurus that contains more than 5000 headings. Both thesauri have similar coverage but differ in granularity. The Deposit and Scientific collections share 215K books. These 215K books have both GTT and Brinkman annotations, compared to 307K books annotated with GTT concepts only and 490K books with Brinkman concepts only. In order to improve the interoperability between these two collections, for example, allowing the end users to use GTT or Brinkman concepts only to access two collections, matching GTT and Brinkman is the first step.

3.2. Heterogeneous collections with multiple thesauri

3.2.1. General description of the scenario

Different CH institutions often have *heterogeneous collections*, *i.e.*, the CH objects are described using different metadata schemas that are chosen to fit the data. In many cases, heterogeneous collections do not overlap, for example, a collection of paintings and a collection of TV programmes do not share objects. Similar to the homogeneous case, the different collections are indexed with different thesauri. When applications need to access different collections simultaneously to provide a multimedia access to various kinds of CH objects, mappings between these thesauri are required. Differently from the previous scenario, differences of the metadata schemas should be taken into consideration in the matching process.

3.2.2. Specific scenario

This task is to connect the multimedia collection in the Netherlands Institute for Sound and Vision (BG) to the book collections of the KB. BG archives all radio and TV programmes broadcasted by the Dutch public broadcasting companies. Besides over 700,000 hours of material, the B&G also houses 2,000,000 still images and the largest music library of the Netherlands. Each object in the BG collection is annotated by concepts from the GTAA thesaurus that contains 160K concepts, among which more than 3000 concepts indicate the subject of the object. This part of GTAA can be matched to the GTT and Brinkman thesauri.

Matching GTAA to the KB thesauri is interesting from a CH perspective. For example, one would be interested to search for some broadcasts in BG, which are about the author of the book he is reading from the KB. Different from the homogeneous collections case, the metadata of books in KB and multimedia objects in BG are very different, and there are no common instances shared by both collections even though they might be semantically related. Therefore, instance-based methods that actually use the metadata of the instances are more important to provide potential mappings.

3.3. Multi-lingual heterogeneous collections with multiple thesauri

3.3.1. General description of the scenario

When interoperability across collections of different nations is desired, for example, accessing all paintings in multiple national galleries plus related books, matching multi-lingual thesauri used to annotate different CH collections is necessary. In this scenario, the collections are heterogeneous, and described in their native languages, including the thesauri which annotate them and all the other metadata available. Common instances are rare across *multi-lingual heterogeneous* collections, although not completely impossible (*e.g.*, in the case described below). One necessary step is to translate the thesauri and metadata into the same language before the matching process.

3.3.2. Specific scenario

While providing access to the resources of different European national libraries, one crucial issue is the fact that collections—and the associated metadata—come in different languages. This hampers the access to several collections at a same time, for example, using one search term in one language to search multiple collections. Such interoperability requires mappings between multi-lingual thesauri. Within the TELplus project, the LVAT prototype uses alignments between

[10]These include properties like *creator*, *title*, *publisher*, *coverage*... See `http://dublincore.org/specifications/`.

[11]`http://goo.kb.nl/`

three thesauri (LCSH,[12] Rameau[13] and SWD[14]) to provide multi-lingual subject-based access to books from British, French and German national libraries [17]. It uses mappings from the already mentioned MACS project, which has a long history of manually aligning these three thesauri. In such context, we investigate instance-based matching techniques, mainly reporting the task of matching LCSH and Rameau.

4. Evaluation methods

There are often two ways of evaluating alignments: one is to evaluate against a reference alignment, and the other is to evaluate the alignment in a real application, *i.e.*, "end-to-end" evaluation [47,19].

4.1. Reference alignment based evaluation

A traditional evaluation method is to (manually) build a reference alignment and measure the *precision* and *recall* of the generated mappings. The precision is the proportion of the correct mappings over all generated mappings, and the recall is the proportion of the correct mappings over all possible correct mappings. Since it is very labor-intensive and time-consuming to build a complete reference alignment, we cannot compute the true precision and recall. Although methods have been developed to approximate the recall measure, an ideal solution does not exist. Depending on application scenarios, a full recall is sometimes not really required. Therefore, in this paper we will not investigate this measure for evaluation based on reference alignments.

Manually evaluating a generated alignment is also not feasible when the two thesauri both contain tens or hundreds of thousands of concepts and the alignment contains a similar amount of mappings. Luckily, the generated mappings normally come with certain confidence values based on which they can be ranked. In practice mappings with high confidence values are more interesting to evaluate. One can measure the *precision-at-rank-n*, which is the average precision up to rank n. This gives an idea of the quality of mappings decreases as the confidence values decreases. Alternatively, for a more global view of all mappings, *sample-based manual evaluation* is often applied in practice. Using a decreasing sample rate, a set of sample mappings are selected and domain experts are asked to manually evaluate whether these samples are correct mappings or not. In this way, the precision-at-rank-n can be approximated.

In this paper, we take the following sample strategy. For each measure, if the generated mappings can be ranked by their confidence values (cf), we take every 10 mapping in the top 1000 mappings, every 100 in the top 10,000 mappings and every 1000 in the top 100,000 mappings, which gives 280 sample mappings per measure for manual evaluation. For measures whose mappings cannot be ranked (*e.g.*, Jensen-Shannon Divergence) or have arbitrary confidence values (*e.g.*, lexical mappings), 200 mappings in total or 50 mappings per confidence value are randomly selected. All the sample mappings from different measures are put together. After removing the redundant sample mappings, they are presented to the evaluator one by one. During evaluation, the evaluator can explore the thesaurus hierarchy or instances (CH objects) associated with the two concepts before they decide whether a mapping is correct or not. In the end, the precision-at-rank-n is calculated with an interval estimation using Wilson's 95% score interval [55].

4.2. End-to-end evaluation

While investigating the matching problem, more and more attention is paid to the scenarios where mappings are actually used [47,22]. The performance in different application scenarios appears to be different from traditional precision and recall measures. They reflect more underlying properties of mappings. Therefore, we take the so-called "end-to-end" evaluation as a second method for evaluation.

Automated re-indexing evaluation In the CH domain, re-indexing is a common scenario in which mappings are used. In scenarios such as data migration, objects in one collection which were originally annotated with one thesaurus need to be incorporated into another collection which is annotated with a different thesaurus. In other words, the objects need to be re-indexed (re-annotated) using the other thesaurus. If an alignment between these two thesauri is available, such re-indexing can be automated, *i.e.*, automatically assigning annotations to an object that are mapped to the original annotations of that object. When there are instances which are already dually annotated with both

[12] http://id.loc.gov
[13] http://rameau.bnf.fr
[14] http://www.d-nb.de/standardisierung/normdateien/swd.htm

thesauri, these common instances can be used to evaluate the quality of the mappings. That is, if the automatically assigned new annotation is consistent with the real annotation, then the mappings are considered to be good or useful in this scenario.

Evaluation measures We take the example of re-indexing books with GTT concepts with Brinkman concepts to introduce the measures in this evaluation. The quality of an alignment is assessed in terms of, for each book, the quality of its newly assigned Brinkman index. We measure the correctness and completeness of the re-indexing as follows: we define precision (P_a) as the average proportion, for the books provided with a Brinkman re-indexing, of the new indices that also belong to a reference (gold standard) set of Brinkman indices. Recall (R_a) is the average proportion, for all books, of the reference indices that were also found using the alignment. The Jaccard similarity (J_a)—the overlap measure of candidate indices and reference ones—provides a combination of precision and recall, which fits well the scenario at hand.[15]

Note that in these three measures, results are counted on an annotation basis. This reflects the importance of different mappings: a mapping for a frequently used concept is more important for this application than a mapping for a rarely used concept.

5. Matching methods

5.1. *Instance-based matching*

Instance-based ontology matching techniques determine the similarity between concepts of different ontologies by examining the extensional information of concepts [27,9], that is, the set of instances they classify. The idea behind such instance-based matching techniques is that similarity between the extensions of two concepts reflects the semantic similarity of these concepts.

Methods based on the overlap of common instances If two ontologies share instances, then the larger overlap of instances of two concepts, the more related these two concepts are. Figure 1 depicts an example, where two concepts `surfsport` and `plankzeilen` (windsurfing) are matched because they have six common instances. In real world scenarios we also have to deal with incorrectly classified instances, data sparseness and ambiguous concepts, so that basic statistical measures of co-occurrence, such as the Jaccard measure, might be inappropriate if applied in a naive way.

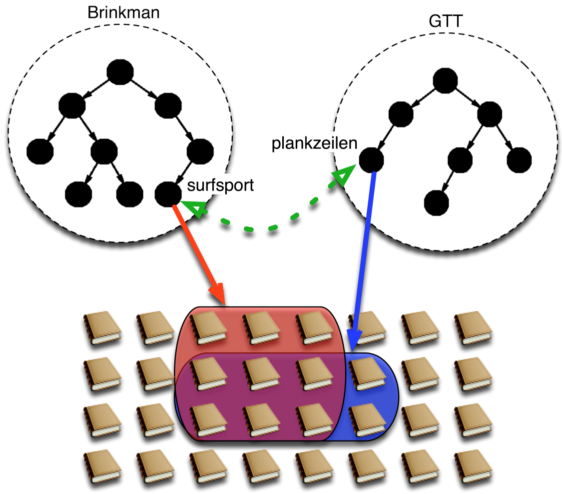


Fig. 1. Two concepts are matched using the Jaccard coefficient.

Extended from our previous work in [21], we deal with the above-mentioned problems by applying a few measures for calculating relatedness of sets based on their elements, such as Pointwise Mutual Information (PMI), Log-Likelihood ratio (LLR) or Jensen-Shannon divergence (JSD), which have been developed in information theory and statistics. In the meantime, we also consider statistical thresholds, particularly for the standard Jaccard and Pointwise Mutual Information measures, which explicitly exclude statistically unreliable information.

5.1.1. *Measures*

In the following we will call the set of instances annotated by a concept C its extension, abbreviated by C^i. As usual, the cardinality of a set S is denoted by $|S|$.

– Jaccard similarity:

$$Jacc(C_1, C_2) = \frac{|{C_1}^i \cap {C_2}^i|}{|{C_1}^i \cup {C_2}^i|}$$

– Corrected Jaccard similarity:

$$Jacc\_corr(C_1, C_2) = \frac{\sqrt{|{C_1}^i \cap {C_2}^i| \times (|{C_1}^i \cap {C_2}^i| - 0.8)}}{|{C_1}^i \cup {C_2}^i|}$$

[15]Please refer our previous work [19,22] for the details.

- Pointwise Mutual Information:

$$PMI(C_1, C_2) = log_2 \frac{|{C_1}^i \cap {C_2}^i| \times N}{|{C_1}^i| \times |{C_2}^i|}$$

where N is the number of annotated instances.
- Corrected PMI[16]

$$PMI\_corr(C_1, C_2) = \frac{|{C_1}^i \cap {C_2}^i|}{|{C_2}^i|} - 0.46 \times \frac{2}{|{C_2}^i| + 1}$$

- Log Likelihood ratio: let $k_1 = |{C_1}^i \cap {C_2}^i|$, $k_2 = |{C_1}^i| - |{C_1}^i \cap {C_2}^i|$, $n_1 = |{C_2}^i|$, $n_2 = N - |{C_2}^i|$, $p_1 = k_1/n_1$, $p_2 = k_2/n_2$, $p_0 = |{C_1}^i|/N$, then

$$\begin{aligned} &LLR(C_1, C_2) = \\ &-2[\log L(p_0, k_1, n_1) + \log L(p_0, k_2, n_2) \\ &\quad - \log L(p_1, k_1, n_1) - \log L(p_2, k_2, n_2)] \end{aligned}$$

where $\log L(p_i, k, n) = k \log p_i + (n-k)\log(1-p_i)$, $i \in \{1, 2, 3\}$.
- Jensen- Shannon divergence (JSD)—distance between co-occurrence distributions [53]. Each concept is represented by its co-occurrence probabilities with all other concepts. The dissimilarity between concepts is calculated by applying the Jensen-Shannon divergence on these representations.

For each measure, we first calculate the similarity of each pair of concepts from two thesauri, we then rank these pairs of concepts based on their similarity measurements which reflects the confidence that a pair of concepts is a true mapping, *i.e.*, the confidence value cf. It is often the case that one concept is mapped to different concepts with different confidence values. We keep them all for the later evaluation as our previous work [22] have shown that one-to-many mappings are also useful in certain application scenarios. Unfortunately, the JSD distance is not comparable across concept pairs due to the way of calculation (less frequently used concepts always have very high distances to other concepts, even if one of those mappings is true). We only take the closest concept as a mapping candidate for the JSD measure.

[16]The detailed explanation of this correction is stated in [18]

5.2. *Instance-based ontology matching by instance enrichment (IBOMbIE)*

Measuring the common extension of concepts requires the existence of a sufficient amount of shared instances, which is often not the case. Therefore, one possible solution is to enrich one ontology by instances from the other ontology which it is mapped to and vice versa. Such enrichment is carried out through mappings between instances, that is, similar instances should be classified to the same or similar concepts.

Take book collections as an example. The instance matching method first matches books from both collections. For each book from Collection A, i_a, there is a most similar book from Collection B, i_b. We then consider i_a shares the same annotation as i_b does. In other words, i_a is now an instance of all concepts which i_b is annotated with. This matching procedure is carried out in both directions. In this way, we can again apply measures on common extensions of the concepts, even if the extensions have been enriched artificially.

There are different ways to match instances. The simplest way is to consider each instance as a document with all its metadata as its description, and apply information retrieval techniques to retrieve the similar instances (documents). We can use the tf-idf weighting scheme which is often used in the vector space model for information retrieval and text mining, or some existing search engine, such as Lucene.[17] Obviously, the quality of the instance matching has an important impact on the concept mappings later.

Multi-lingual cases In order to apply the IBOMbIE method in a multi-lingual context, good automated translation is important. The quality of translation clearly plays an important role in providing reliable instance matchings. One can take a naive approach, *e.g.*, using the Google translate service,[18] or more powerful translation tools to translate both book metadata and concept labels into the same language before applying the IBOMbIE method. In our paper, we use the Google translation service, and will investigate other services such as Inter-Active Terminology for Europe,[19] Wikipedia translations or other online dictionaries in the future.

[17]http://lucene.apache.org/
[18]http://translate.google.com/
[19]http://iate.europa.eu

5.3. SKOS lexical matcher: a baseline

In this paper, we also use a lexical matcher as a baseline to compare the performance of the other methods.

Many lexical matchers are only dedicated to English. We use a matcher–first developed in [29]—that also works with Dutch and French. It is mostly based on the CELEX[20] morphology databases, which allows the recognition of lexicographic variants (based on their *lemmas*) and morphological components of a word form.

We use this matcher to produce *equivalence* mappings between concepts. The different comparison methods it uses give rise to different confidence values (cf): using exact string equivalence is more reliable than using lemma equivalence. Also, the matcher considers the status of the lexical features it compares. The vocabulary conversions that we made describe concepts according to the SKOS model [20], which means they have *preferred* and *alternative* labels. For two concepts, a comparison based on alternative labels is considered less reliable than a comparison based on preferred labels. The combination of these two factors—different comparison techniques and different features compared—results in a grading of the produced mappings, which can be used as a confidence value.

This lexical matcher only works considering one language at a time. To apply it in a multi-lingual case, we processed with a simple translation of the vocabularies before applying it. For each vocabulary pair, we translate each vocabulary by adding new labels that result from translating the original labels with the Google translate service.[21] We then run the matcher twice. The translation of Rameau to English is matched (in English) to the original LCSH version, and the translation of LCSH in French is matched (in French) to the original Rameau version. The results of both runs are then merged into a single set of mappings. It is important to notice that we just select the *exact* (equivalence) links provided by the lexical matcher.

6. Experiments and results

In this section, we describe our experiments in three representative thesaurus matching cases from the Cultural Heritage domain, that is, based on homogeneous, heterogeneous and heterogeneous multi-lingual instances.

Evaluation setup As mentioned in Section 4, instances shared by both collections can be used not only to generate mappings, but also to evaluate the quality of the mappings. Therefore, when joint instances are available, such as in the first and third cases, we separate one third of joint instances with which we evaluate the quality of the mappings generated from the rest of the instance data.

In the evaluation step, the generated mappings are first sampled and manually evaluated, as introduced in Section 4.1. If applicable, an automated re-indexing evaluation applying all mappings on the one third unused joint instances will be carried out, as introduced in Section 4.2.

6.1. Matching thesauri in homogeneous collections

We first match the GTT and Brinkman thesauri, which contain 35K and 5K concepts respectively. As said, these two thesauri are used to annotate two book collections in KB, which actually share 215K books. This gives us the opportunity to apply simple instance-based matching methods (using the different measures mentioned in Section 5.1) on two thirds of the 215K joint instances, and to apply the IBOMbIE method which takes both two thirds of dually and all singly annotated books into account.

The SKOS lexical matcher is also applied to get lexical alignments. It produces 3,043 mappings with a confidence value (cf) equal to 0.95, 544 mappings with $cf = 0.9$ and 50 with $cf = 0.85$. The JSD technique produces 4,905 mappings with the same confidence value. The other measures all produced more than 100K mappings, which can be ranked by their confidence values.

Manual evaluation on samples After sampling (see Section 4), in total 1914 mappings generated from eight methods are manually evaluated. Table 1 presents an overview of the evaluation results.

As listed in Table 1, Jaccard, its variation and IBOMbIE score very high, slightly lower than the lexical mappings, especially among the top 10,000 mappings. In Table 2, we can see that among the top 10,000 mappings, there is less than 30% mappings shared by the instance-based mappings and the lexical mappings. Therefore, the similar quality shown in Table 1 suggests that the instance-based methods can produce new

[20] http://celex.mpi.nl/
[21] http://translate.google.fr/

Measures	up to 1,000	up to 10,000	up to 100,000
Jacc	0.84±0.06	0.37±0.09	0.08±0.05
Jacc_corr	0.81±0.07	0.34±0.09	0.11±0.06
PMI	0.45±0.09	0.27±0.08	0.07±0.05
PMI_corr	0.55±0.09	0.24±0.08	0.08±0.05
LLR	0.65±0.08	0.37±0.09	0.12±0.06
IBOMbIE	0.86±0.06	0.35±0.09	0.10±0.05
JSD	0.42 ±0.07		
Lexical	$cf = 0.95$	$cf = 0.9$	$cf = 0.85$
	0.91 ± 0.07	0.41 ±0.13	0.57 ± 0.13

Table 1

Manual evaluation – GTT *vs*. Brinkman

Jacc	3,088		
Jacc_corr	3,094	8,663	
IBOMbIE	2,750	4,120	4,259
	Lexical ($cf \geq 0.9$)	Jacc	Jacc_corr

Table 2

The number of mappings shared by measures among top 10,000 mappings

quality mappings which are complementary to the lexical mappings.

Fig. 2 gives detailed information on the precision-at-rank-n. The straight line and stair-shaped line for the JSD and lexical mappings show the average precision in general or at certain confidence levels. As we can see, the quality of the mappings from the most of measures, including the lexical mappings, deteriorates rapidly after rank 5000, approximately. This suggests a cutting point below which instance-based mappings may be considered generally valid—or at least, valid enough for scenarios where a human operator is in the loop.

Re-Indexing evaluation We now measure the performance of the mappings in the re-indexing scenario. Again, the mappings are ranked in a descending order of their confidence level. We calculated the three measures P_a, R_a and J_a (see Section 4.2) at every 1,000 mappings and plot them in Fig. 3. We can observe the typical tradeoff between precision and recall. It is somehow not surprising that, compared to the original PMI measure, the directed PMI_corr has the best performance in terms of J_a, as the re-indexing process translate GTT concepts into Brinkman concepts which is the consistent with the calculation of PMI_corr . The two Jaccard measures perform similarly. It is crucial here to notice that they both reach their J_a peak much ahead of PMI_corr —considering the logarithmic scale of the graphs. Both Jaccard measures are better at giving higher ranks to the correct mappings that are most useful for re-indexing—*i.e.*, the ones between concepts that are used most often. PMI_corr seems a bit more conservative, and catches up only after having returned a large number of mappings.

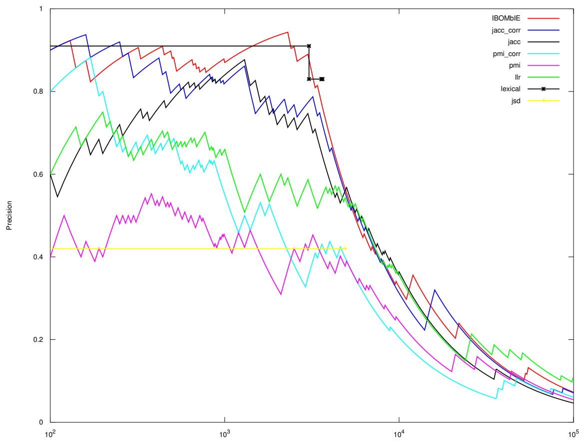


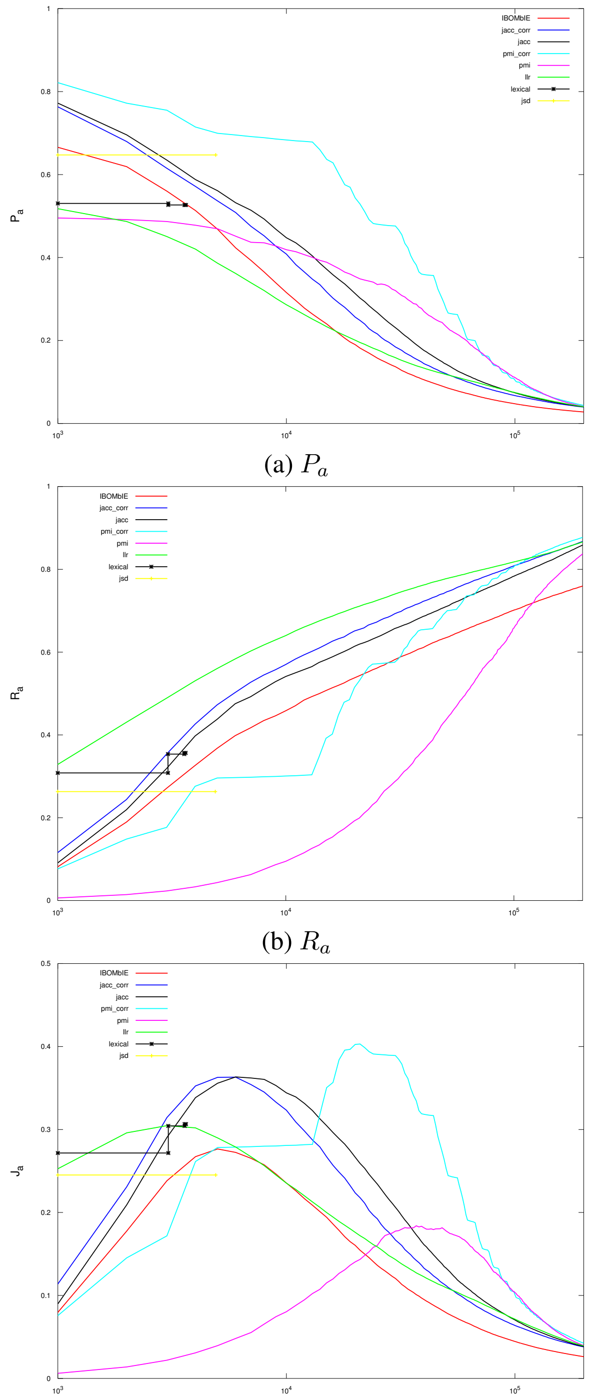
Fig. 2. Manual evaluation – GTT *vs*. Brinkman, where the x-axis is the rank of the mappings and the y-axis is the precision-at-rank-n

Another fact worth noticing is that the lexical alignment performed worse than the above mentioned three instance-based measures, while it has much stable and high quality when evaluated by humans. This further confirmed that end-to-end evaluation is sometimes more useful in certain application scenarios.

6.2. Matching in heterogeneous collections

We now match thesauri which are used to annotate heterogeneous collections, more specifically, matching the GTT/Brinkman thesauri from KB to the GTAA thesaurus of BG. The collections from these two CH institutes have their own metadata schema, and do not have any shared instances. This prevents us to apply simple instance-based matching methods which rely on the existence of common instances. We can only apply the lexical matcher and the IBOMbIE method to match GTT/Brinkman to GTAA. Because of the absence of joint instances, the re-indexing evaluation is not applicable either. Therefore, we only report the manual evaluation results.

The BG collection contains nearly 60K instances. The KB collections for GTT and for Brinkman contain more than 500K and 300K books, respectively. The GTAA thesaurus has many branches, among which we are interested to map the “Subject” one to GTT and Brinkman. This “Subject” branch contains 3,869 concepts.



(a) P_a

(b) R_a

(c) J_a

Fig. 3. Re-Indexing evaluation – GTT *vs*. Brinkman

	$cf = 0.95$	$cf = 0.9$	$cf = 0.85$
GTT-GTAA	2355	794	113
Brinkman-GTAA	1252	305	40

Table 3

Lexical mappings between GTT/Brinkman and GTAA

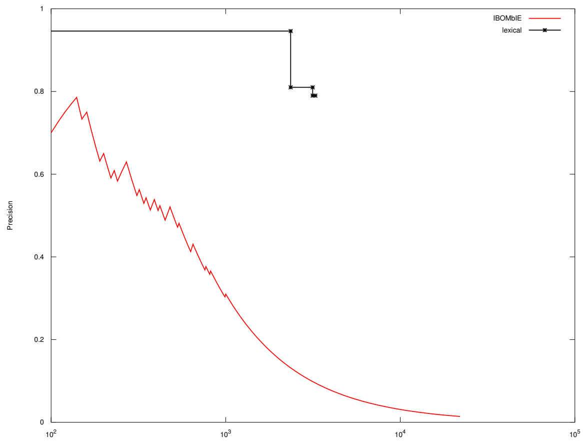
Table 3 gives the results for lexical mappings between GTT/Brinkman and GTAA. Compared to the other two cases, the IBOMbIE mappings share much less with the lexical mappings. Among the top 1,000 Brinkman-GTAA and GTT-GTAA mappings, 265 and 328 mappings involve lexically equivalent concepts. There are very few (13 for Brinkman-GTAA and 30 for GTT-GTAA) more lexical mappings found after rank 1,000.

After the manual evaluation, the precisions up to 1,000 mappings are 0.32 ± 0.08 for GTT-GTAA and 0.22 ± 0.07 for GTAA-Brinkman. Among the sampled mappings, there are actually very few common mappings between the lexical alignment and the IBOMbIE one. This again suggests that the instance-based matching method can produce new true mappings complimentary to the lexical method. Its precision is however very low: the KB collections are probably too different from the BG one. Either the usage of semantically equivalent concepts for indexing varies too much across collections, or IBOMbIE fails at capturing similarity between instances that have similar topics.

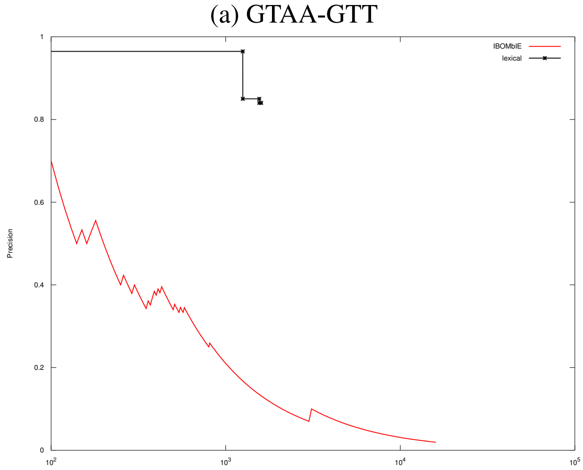
6.3. *Matching in multi-lingual collections*

We now match the English-language subject heading list LCSH and the French subject heading list Rameau. As mentioned before, LCSH contain nearly 340K concepts that are used to annotate books in the British National Library. Rameau contains more than 150K concepts used to annotate books in the French national library. Although both national libraries mainly contain the books in their own languages and formats, they do share a small amount of books (more than 180K books, 4.9% of the both collections), identified by the same ISBN numbers. This allows us to apply all matching methods we discuss in this paper.

As indicated previously, one third of the joint books are separated to be used in the re-indexing evaluation, and the rest of books are used to generate mappings. The lexical matcher generated 32,223 mappings with $cf = 0.95$, 536 mappings with $cf = 0.9$ and 47 mappings with $cf = 0.85$. The JSD method generated



(a) GTAA-GTT



(b) GTAA-Brinkman

Fig. 4. Manual evaluation

Measures	1,000	10,000	100,000
Jacc	0.44 ± 0.09	0.50 ± 0.10	0.27 ± 0.08
Jacc_corr	0.93 ± 0.04	0.62 ± 0.09	0.20 ± 0.08
PMI	0.53 ± 0.09	0.47 ± 0.10	0.18 ± 0.07
PMI_corr	0.19± 0.07	0.28 ± 0.08	0.12 ± 0.06
LLR	0.77± 0.07	0.52 ± 0.10	0.21 ± 0.08
IBOMbIE	0.67 ± 0.08	0.55 ± 0.09	0.16 ± 0.07
JSD	0.25 ± 0.06		
Lexical	$cf = 0.95$	$cf = 0.9$	$cf = 0.85$
	0.89 ± 0.08	0.57 ± 0.13	0.61 ± 0.13

Table 4

Manual evaluation – LCSH *vs*. Rameau

Jacc_corr	4,277		
LLR	3,690	2,339	
IBOMbIE	3,570	1,344	756
	Lexical ($cf \geq 0.9$)	Jacc_corr	LLR

Table 5

The number of mappings shared by measures among top 10,000 mappings

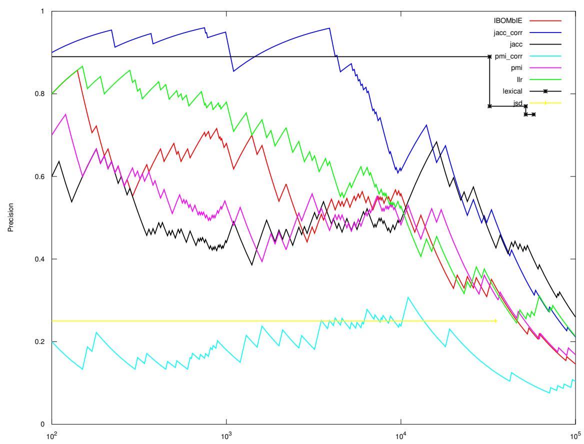


Fig. 5. Manual evaluation – LCSH *vs*. Rameau

34,839 mappings with equal confidence value. The other measures similarly generated more than 150K mappings.

Manual evaluation on samples One English-French bilingual evaluator was asked to evaluate 1377 mappings sampled from the mappings generated by 8 measures. Table 4 gives an overview of the evaluation results.

As seen in Fig. 5, the corrected Jaccard measure slightly outperforms the lexical technique for the first thousand mappings. From Table 5, we can see there are nearly 43% mappings generated by Jacc_corr are lexical mappings. The higher quality indicates that instance-based method do provide high quality mappings which are missed by the lexical mapper. The very small amount of shared mappings between LLR and IBOMbIE and the similar quality after evaluation suggests that these two methods focus on different parts of the mapping space. Compared to the GTT/Brinkman case, it is obvious that IBOMbIE does suffer from the quality of automated translation.

Re-Indexing evaluation Similar to the GTT/Brinkman case, the performance of different methods in the re-indexing scenario is different from the manual evaluation. The worst measure according to the human evaluator, PMI_corr, gives the best J_a again, and its peak occurs way after the other measures have reached their peaks, leading to the same interpretation as previously. The two Jaccard measures are similarly good. Here too, the lexical mappings do not demonstrate much advantage for the re-indexing task.

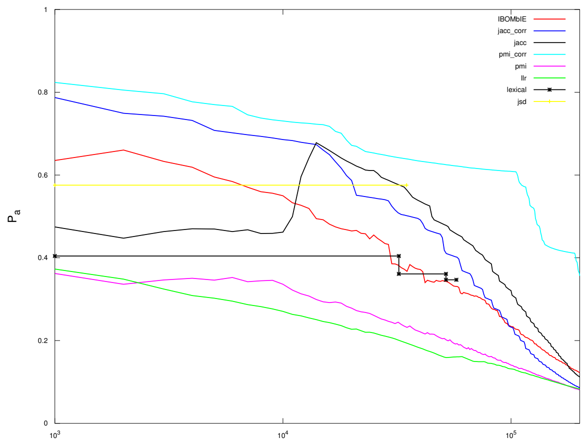
In both cases, LLR is the best the measure in terms of R_a, and almost always the worst in terms of P_a. It seems that the LLR measure privileges correspondences that provide more re-indexing suggestions over the ones that lead to only more precise suggestions. This might still be useful for more interactive re-indexing processes, where the human re-indexer can invest time picking up the correct suggestions and is reluctant to miss good ones.

7. Relevance and issues of instance-based matching for Cultural Heritage problems

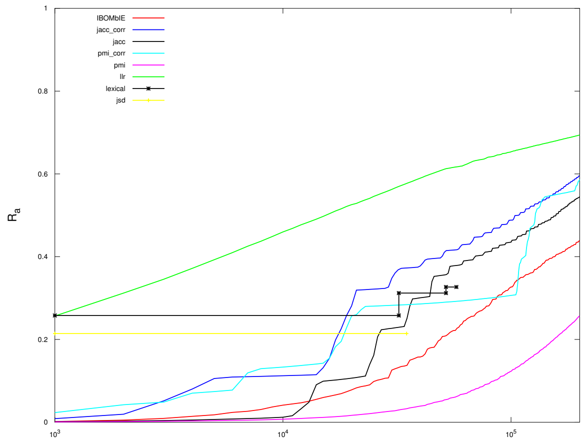
7.1. Instance-based matching is a promising technique...

Fitting available data Instance-based techniques have the advantage that they do not suffer under some weaknesses of the vocabularies used in the cultural sector. First, thesauri, subject heading lists, *etc*. feature structural relations between their elements. But while some classification systems are entirely structured as trees, the networks of semantic relationship are generally poor—as testified by the vocabularies of our example cases. Their quality is quite unpredictable: some parts of a vocabulary can receive more attention than others, depending on vocabulary maintainers' resources and interest. Also, the ontological correctness of the links is often debatable. Hierarchical links, for example, can be employed for (among others) part-whole, class-subclass or set-member relationships within a same vocabulary. Varied interpretations of these relations across different parts of a vocabulary can lead to surprising findings [36]. This makes comparison of structural similarity across vocabularies, as performed by common structure-based ontology matching tools, unreliable for the Cultural Heritage case.

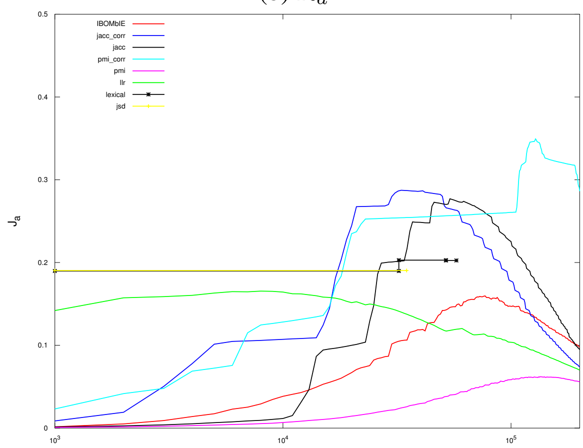
Further, instance-based techniques are not impacted by lexical issues that undermine the results of many lexical matching techniques, which assume (near)-synonymy between all labels associated with a given resource. Vocabularies do usually include synonyms or near-synonyms for many of the preferred labels of concepts. However there are coverage problems: some vocabularies do not feature much appropriate lexical data. This of course especially applies when vocabularies have to be mapped across languages. There are also precision issues caused by many vocabularies using



(a) P_a



(b) R_a



(c) J_a

Fig. 6. Re-Indexing evaluation – LCSH *vs*. Rameau

upward posting, a practice intended to streamline vocabulary management and usage by attaching specific terms to more general concepts. For example, "Spanish flu" is directly attached to the concept of "flu" in GTT.

On the other hand, as noted in the introduction, available instance data is relatively abundant in cultural collections. Numerous books and other documents have been described using the vocabularies at hand. Additionally to the subject annotation that links to concepts, these descriptions typically include the title of the document, its creators, the dates associated with it, sometimes a short summary, *etc*. This enables:

- compensating the above mentioned "structural semantics" weaknesses fitting extensional semantics in the matching process.
- compensating the potential shortfall of lexical data at the concept level by considering lexical data at the instance level.

Finding mappings that are difficult to find by humans Manual matching is labor-intensive, and automatic matching can help a lot to assist this process. An interesting feature of instance-based techniques is that they are able to detect mappings that are hard to detect manually. When human operators align vocabularies, they will first focus on the easiest hints for semantic relations, namely lexical similarities. For example, the usual starting point, when one has to find an equivalent for a given concept in another vocabulary, is to search in that vocabulary for concepts having the same or similar label. Of course this becomes much more difficult when labels are in different languages, equivalent concepts do not have the same label or application scenario dictates taking into account the usage of concepts in collections. Getting a precise idea on how well large sets of documents are related requires lots of time and translation efforts for human operators, even trained ones. As a result, the extensional dimension of matching is likely to be neglected.

This has been confirmed when comparing the results of instance-based matching between LCSH and Rameau to the manual mappings created in the MACS project [51]. Extensional techniques produce results that overlap with, but are non-identical to the MACS mappings. Around 50% of the first 50,000 instance-based mappings are not "judgeable" considering the MACS results: *i.e.*, there is no mapping in MACS that allows to assess them as right or wrong. On the other hand, 86% of the mappings found by the simplest lexical technique were found to be already present in the MACS dataset. Exploiting usage of concepts have in fact been deemed relevant for MACS [24]. But it is a difficult task. Future initiatives with development resources specifically dedicated to helping projects such as MACS, should investigate how the results of experiments such as ours can be effectively included in their workflow. For example, a person matching two vocabularies could be prompted with a number of suggestions, which she could use as a complement to her own intellectual efforts. The task of manually matching vocabularies would thus become one of rather validating the results of automatic techniques, and finding mappings, which are not found by these techniques. One could imagine adapting the annotation suggestion interface we have developed for the re-indexing scenario at KB [18], which prompts a cataloger with a number of concept suggestions for indexing a given book, while giving her some hints on why these suggestions are made, such as the matching technique employed and a confidence measure.

Finding mappings that are relevant for concrete applications In previous work, we have argued for properly taking the application scenarios into account when matching vocabularies [22]. Instance-based techniques do fit that vision, since they are based on the actual use of concepts in collections. This is especially precious for applications that are tightly connected with such use, *e.g.*, using mappings to perform query reformulation over two collections indexed with different vocabularies, as suggested by MACS.

In STITCH we have thoroughly experimented with re-indexing, namely, taking documents that have been described with one vocabulary, and enriching them with annotations using a second vocabulary. Here, instance-based techniques prove to be useful for "reproducing" enrichment patterns from an existing base of dually described documents.[22] A typical example, Rameau's `Cavitation` can be mapped to LCSH's `Hydraulic system`. Allthough these two concepts are not equivalent in principle, they are used for describing the same books. This is a case where subject indexing strategies differ in the respective collections: one library's practice may very well dictate a different focus from the one of another library, for the same books. The ability to overcome heterogene-

[22] Assuming such dually-indexed set is not unrealistic: if re-indexing is of any relevance, librarians may have already been performing it in the past, as in the KB case. Or they could be ready to work on a bootstrap dataset, if it can be exploited by automatic techniques that will later assist them.

ity of practices have been found a very interesting feature by librarians confronted with instance-based alignments [18].

In fact, instance-based matching techniques can be useful to detect mappings which are not strict equivalence links, such as the `skos:broadMatch` and `skos:relatedMatch` relations from the SKOS model, which are derived from thesaurus standards. In the evaluations we carried out for this paper, we found for example links between `spoorwegen` (railways) and `treinverkeer` (train traffic) or between `volkscultuur` (popular culture) and `volkskunde` (cultural anthropology). Alternative techniques may detect such relations, but they would require exploiting relevant background sources, either dictionaries that bring extra lexical knowledge for the concepts at hand, or entire structured vocabularies which can be used as "oracles" providing with semantic paths that are missing in the two initial vocabularies [1,15].

Finally, instance-based techniques will perform best for the concepts that are most relevant to applications concerned with actual vocabulary usage. The more two concepts are used, the more the semantic equivalence measure obtained on the basis of their (co)-occurrence can be trusted. An alignment that fails to detect many equivalences could still have a quite high "perceived" quality for a given application, if it gets correctly the mappings for the concepts most often used in that application. This in fact explains many of the artifacts observed in the re-indexing evaluations from the previous section.

7.2. ... but it does not solve every problem yet

Applicability of instance-based techniques In this paper we have applied straightforward instance-based techniques, which allowed us to deploy and apply a number of them relatively quickly. The most complex method is clearly IBOMbIE, which requires setting an instance matching process prior to the concept matching step. The reader should be aware that we also tested more complex techniques, including machine-learning approaches [50]. Applying these more complex techniques requires more effort and expertise than the techniques we evaluate in this paper, while the simpler techniques have been proven to work relatively well.

One crucial issue remains: availability of suitable data. Simple co-occurrence based methods will not work in cases where dually annotated instances are not available. Such cases will dictate the use of methods like IBOMbIE, which has a much broader application range but greater complexity and lower precision. Further, even when dually annotated instances are available, one may miss the critical mass of documents necessary to obtain reliable results for entire vocabularies. Concept usage statistics show very long tails, with lots of concepts being used to annotate only a couple of documents. Many concepts from the vocabularies at hand were not even used in the collections we had, as shown in [51]. Finally, the instance data present in the cultural heritage institution may sometime not lend themselves to instance-based matching. The LCSH-Rameau case exemplifies this: available book descriptions use the labels of concepts in their subject fields, rather than well-defined identifier references. This required specific pre-processing steps to handle syntax errors, updates of the vocabularies that were not reflected in the instance data, *etc*. Such efforts higher the barrier for application of instance-based techniques, and are error-prone.

Operationalization of results While they fit many applications better, instance-based techniques are not an *exact* fit for all applications. Often, applying these techniques as such will suffer from the same drawbacks as applying out-of-the-box tools based on other techniques. In particular, for the kind of application scenarios we envisioned (query re-formulation across collections, re-indexing) the results of simple techniques do not directly meet the requirement for mapping *groups* of concepts.

When books are annotated in libraries, it is indeed possible to assign several concepts to one book, each of them reflecting a facet of that book's subject. This raises important matching issues when the granularity of vocabularies differ, or when indexing practices have different foci. A concept in one vocabulary may not be suitable to match with one single equivalent concept in the second vocabulary; it may correspond to a combination of concepts. This means that, *e.g.*, for the re-indexing case, each occurrence of the first concept should lead to a group occurrence of the second combination of concepts. This is especially valid for vocabularies like LCSH or Rameau that provide rules for *pre-coordination*, *i.e.*, constructing complex "strings" from multiple simple concepts, as in `France-History-13th century`. This also applies to cases where no such established rules are available. Experiments at KB with GTT and another thesaurus [18] illustrated the need for rules associating, *e.g.*, on the one hand the concept `Spanje ; reisgidsen` (Spain; travel guides) and on the other hand the two concepts `Reisgidsen` (travel guides) and `Spanje` (Spain).

When applying simple techniques that produce one-to-one mappings, one may have to post-process an alignment to operationalize it so that it can be consumed by the application at hand. In [52] we experimented with various strategies that exploit similarity measures to create concept group associations. It is also possible to adapt and compute similarity measures for groups of concepts, as we did in [18]. This however raises the computational complexity of matching. It also may lower the chances of re-using the resulting mappings in other alignments.

Genericity and interoperability This leads us to a third problem of instance-based techniques: the possibly limited scope of their results. The more fine-tuned to a given application an alignment is, the more difficult it is to adapt it to other applications. As said, an interesting feature of instance-based techniques is their ability to easily integrate into the matching process the usage of concepts in a given application. While this is valuable for all applications that have a same concept usage "profile," it may prove harmful when re-using an alignment outside its original production context.

Consider the linked data context: alignments are crucial there to "follow one's nose" from one dataset to the other, even when these datasets are not directly aligned together. It would suffice to have a "hub" that is directly aligned to each of them, which can be used to derive indirect mappings.[23] That vision is at risk when one "step" of such chains introduces a bias.

One basic solution is to make sure that appropriate context information is published to orient data consumers. Specific types of mapping relations can be used, such as the `skos:closeMatch` property introduced by SKOS as a complement to `skos:exactMatch`—the latter being used when "two concepts have equivalent meaning, and the link can be exploited across a wider range of applications and schemes" [20]. A complementary, more complete solution, acknowledges that in an open world, and for various applications, different matching techniques will complement each other. It implies a full contextualization of alignments, keeping track of the technique that produced an alignment, who produced it, and possibly, the class of applications it is mainly intended for. Vocabulary and alignment services [32,44,46] provide appropriate environments for managing and delivering of such data. It is up to matching tool developers to ensure that their tools would fit such an infrastructure.

[23]The archetypal "hub" is the DBpedia dataset, which is the target of many of the alignments published so far on the Linked Data "cloud", cf. `http://linkeddata.org`.

8. Conclusion

Ontology matching is crucial to Semantic Interoperability problem in the CH domain as it allows museums, libraries, archives and other CH institutions to organise, describe, share and publish the objects in their care in novel, integrated ways. Of course, the Web plays an important role in this process, but the Interoperability problem can even be found within institutions and in local applications.

The Semantic Web community can play an important role in helping those CH institutions in these efforts by providing methods for interoperability. The large body of work in ontology matching is an ideal starting point. However, experience shows that applicability of generic ontology matching technology is limited to environments with a tradition of semantic annotations and with rather inexpressive knowledge organisation schemes containing up to hundreds of thousands of concepts.

This rich semantic resource requires special methods for interoperability, and extensional methods lends themselves exceptionally well for the problems at hand. This has to do with the intended semantics of the vocabularies (intended for organising objects), the frequent existence of rather huge sets of semantically annotated resources and the rather flat and non-expressive type of schemas.

This paper provides a comprehensive discussion of several aspects of Semantic Interoperability in CH, with a particular focus on extensional matching methods: we give a comprehensive overview of related work, discuss three different matching methods, and evaluate them in three typical matching scenarios we encountered in our long-standing collaboration with a number of European National libraries and other CH institutions.

With these experiments we provide a systematic description of how to use extensional matching methods to heterogeneous and even multi-lingual collections, scenarios that typically do not to lend themselves to ontology matching methods. The results of the experiments are positive throughout, showing surprisingly high precision at a level of reasonable recall (or vice versa). As important as those results, however, should be the observation that conducting those experiments require extensive knowledge of the CH domain and specific scenarios.

This paper aims at giving a comprehensive overview over most relevant aspects of Semantic Interoperability in CH through extensional methods. But even though we believe the reported results to be very promising, and the insights we gained and described to be significant, the problems are far from solved. On the contrary, all the experience shows that the work we report on is just an important first step towards true Semantic Interoperability.

Acknowledgements

This work is funded by the NWO CATCH and EU eContentPlus programmes (STITCH and TELplus projects). We are grateful to the following colleagues from the STITCH project for their constant help and advice: Henk Matthezing, Claus Zinn, Frank van Harmelen and Paul Doorenbosch. Patrice Landry, Jeroen Hoppenbrouwers and Geneviève Clavel provided us with MACS data. Various people at Library of Congress, DNB, BnF, TEL Office, Beeld en Geluid, VU and KB have provided or helped us with SHL and collection data, including Barbara Tillett, Anke Meyer, Claudia Werner, Françoise Bourdon, Michel Minguam, Sjoerd Siebinga, Johan Oomen, Véronique Malaisé, Dirk Kramer. Special thanks to the National Library of the Netherlands (KB) for hosting us for 4 years.

References

[1] Z. Aleksovski, M. C. A. Klein, W. ten Kate, and F. van Harmelen. Matching unstructured vocabularies using a background ontology. In S. Staab and V. Svátek, editors, *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management, EKAW 2006*, volume 4248 of *Lecture Notes in Computer Science*, pages 182–197. Springer, 2006.

[2] P. Avesani, F. Giunchiglia, and M. Yatskevich. A large scale taxonomy mapping evaluation. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, *International Semantic Web Conference*, volume 3729 of *Lecture Notes in Computer Science*, pages 67–81. Springer, 2005.

[3] F. Boterham and J. Hubrich. Towards a comprehensive international knowledge organisation system. In *7th European Networked Knowledge Organization Systems (NKOS) Workshop at the 12th ECDL Conference*, Aarhus, Denmark, 2008.

[4] C. Caracciolo, J. Euzenat, L. Hollink, et al. Results of the ontology alignment evaluation initiative 2008. In *Proceedings of the 3rd International Workshop on Ontology Matching, collocated with the 7th International Semantic Web Conference (ISWC 2008)*, Karlsruhe, Germany, 2008.

[5] N. Choi, I.-Y. Song, and H. Han. A survey on ontology mapping. *SIGMOD Record*, 35(3):34–41, 2006.

[6] M. Day, T. Koch, and H. Neuroth. Searching and browsing multiple subject gateways in the renardus service. In *Proceedings of the Sixth International Conference on Social Science Methodology*, Amsterdam, The Netherlands, 2005.

[7] A. Doan, P. Domingos, and A. Halevy. Learning to match the schemas of data sources: a multistrategy approach. *Machine Learning*, 50(3):279–301, 2003.

[8] A. Doan and A. Halevy. Semantic integration research in the database community: a brief survey. *AI Magazine*, 26(1), 2005.

[9] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. In *Proceedings of the 11th international conference on World Wide Web*, pages 662–673, 2002.

[10] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Ontology matching: a machine learning approach. In *Handbook on Ontologies in Information Systems*, pages 397–416. Springer, 2004.

[11] J. Euzenat, A. Ferrara, L. Hollink, et al. Results of the ontology alignment evaluation initiative 2009. In *Proceedings of the 4th International Workshop on Ontology Matching, collocated with the 8th International Semantic Web Conference (ISWC 2009)*, Washington, DC, 2009.

[12] J. Euzenat, A. Isaac, C. Meilicke, et al. Results of the ontology alignment evaluation initiative 2007. In *Proceedings of the Second International Workshop on Ontology Matching, collocated with the 6th International Semantic Web Conference (ISWC2007)*, Busan, Korea, 2007.

[13] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Berlin Heidelberg (DE), 2007.

[14] J. Euzenat and P. Valtchev. Similarity-based ontology alignment in OWL-lite. In R. L. de Mántaras and L. Saitta, editors, *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)*, pages 333–337. IOS Press, 2004.

[15] F. Giunchiglia, M. Yatskevich, and P. Shvaiko. Semantic matching: algorithms and implementation. *J. Data Semantics*, 9:1–38, 2007.

[16] R. Ichise, H. Takeda, and S. Honiden. Integrating multiple internet directories by instance-based learning. In *Proceedings of the eighteenth International Joint Conference on Artificial Intelligence*, 2003.

[17] A. Isaac, S. Chambers, A. Gos, W. Vermeer, G. Clavel, and P. Landry. Prototype integrating MACS initial data and new alignments into The European Library framework. Technical report, TELplus project, January 2010. Available at `http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/documents/TELplus_D3.4_04012010.pdf` [2011.05.09].

[18] A. Isaac, D. Kramer, L. van der Meij, S. Wang, S. Schlobach, and J. Stapel. Vocabulary matching for book indexing suggestion in linked libraries – a prototype implementation & evaluation. In *Proceedings of the the 8th International Semantic Web Conference (ISWC 2009)*, Washington, DC, October 25-29 2009.

[19] A. Isaac, H. Matthezing, L. van der Meij, S. Schlobach, S. Wang, and C. Zinn. Putting ontology alignment in context: usage scenarios, deployment and evaluation in a library case. In S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, editors, *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*, volume 5021 of *Lecture Notes in Computer Science*, pages 402–417. Springer, 2008.

[20] A. Isaac and E. Summers. SKOS Primer. W3C Group Note, August 18 2009. Latest version available at `http://www.w3.org/TR/skos-primer/`.

[21] A. Isaac, L. van der Meij, S. Schlobach, and S. Wang. An empirical study of instance-based ontology matching. In K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, editors, *Proceedings of the the 6th International Semantic Web Conference (ISWC 2007)*, volume 4825 of *Lecture Notes in Computer Science*, pages 253–266. Springer, 2007.

[22] A. Isaac, S. Wang, C. Zinn, H. Matthezing, L. van der Meij, and S. Schlobach. Evaluating thesaurus alignments for semantic interoperability in the library domain. *IEEE Intelligent Systems, Special Issue on AI and Cultural Heritage*, Mar/Apr 2009.

[23] Y. Kalfoglou and W. M. Schorlemmer. Ontology mapping: the state of the art. In Y. Kalfoglou, W. M. Schorlemmer, A. P. Sheth, S. Staab, and M. Uschold, editors, *Semantic interoperability and integration*, volume 04391 of *Dagstuhl Seminar Proceedings*. IBFI, Schloss Dagstuhl, Germany, 2005.

[24] P. Landry. Multilingualism and subject heading languages: how the MACS project is providing multilingual subject access in Europe. *Catalogue & Index: Periodical of the Chartered Institute of Library & Information Professionals (CILIP) Cataloguing & Indexing Group*, 157, 2009.

[25] L. A. P. P. Leme, M. A. Casanova, K. K. Breitman, and A. L. Furtado. Instance-based OWL schema matching. In J. Filipe and J. Cordeiro, editors, *ICEIS*, volume 24 of *Lecture Notes in Business Information Processing*, pages 14–26. Springer, 2009.

[26] J. Li, J. Tang, Y. Li, and Q. Luo. Rimom: A dynamic multistrategy ontology alignment framework. *IEEE Trans. Knowl. Data Eng.*, 21(8):1218–1232, 2009.

[27] W.-S. Li, C. Clifton, and S.-Y. Liu. Database integration using neural networks: implementation and experiences. *Knowledge and Information Systems*, 2:73–96, 2000.

[28] A. Maedche, B. Motik, N. Silva, and R. Volz. MAFRA - A MApping FRAmework for distributed ontologies. In A. Gómez-Pérez and V. R. Benjamins, editors, *EKAW*, volume 2473 of *Lecture Notes in Computer Science*, pages 235–250. Springer, 2002.

[29] V. Malaisé, A. Isaac, L. Gazendam, and H. Brugman. Anchoring Dutch Cultural Heritage Thesauri to WordNet: two case studies. In *ACL 2007 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, Prague, Czech Republic, June 28 2007.

[30] V. Maltese, F. Giunchiglia, and A. Autayeu. Save up to 99% of your time in mapping validation. In R. Meersman, T. S. Dillon, and P. Herrero, editors, *OTM Conferences*, volume 6427 of *Lecture Notes in Computer Science*, pages 1044–1060. Springer, 2010.

[31] P. Mayr and V. Petras. Building a terminology network for search: the KoMoHe project. *CoRR*, abs/0808.0518, 2008.

[32] N. F. Noy, N. Griffith, and M. A. Musen. Collecting community-based mappings in an ontology repository. In *Proceedings of the 7th International Semantic Web Conference (ISWC 2008)*, Karlsruhe, Germany, 2008.

[33] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[34] B. A. C. Schopman, S. Wang, and S. Schlobach. Deriving concept mappings through instance mappings. In *Proceeding of the 3rd Asian Semantic Web Conference*, pages 122–136, Bangkok, Thailand, December 2008.

[35] G. Schreiber, A. Amin, L. Aroyo, M. van Assem, V. de Boer, L. Hardman, M. Hildebrand, B. Omelayenko, J. van Osenbruggen, A. Tordai, J. Wielemaker, and B. Wielinga. Semantic annotation and search of cultural-heritage collections: the MultimediaN E-Culture demonstrator. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):243 – 249, 2008.

[36] S. Spero. LCSH is to thesaurus as doorbell is to mammal: visualizing structural problems in the Library of Congress subject headings. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, pages 203–203. Dublin Core Metadata Initiative, 2008.

[37] G. Stumme and A. Maedche. Fca-merge: bottom-up merging of ontologies. In *Proceedings of the 17th International Conference on Artificial Intelligence (IJCAI '01)*, pages 225–230, Seattle, WA, USA, 2001.

[38] R. Thompson, K. Shafer, and D. Vizine-Goetz. Evaluating dewey concepts as a knowledge base for automatic subject assignment. In *ACM DL*, pages 37–46. ACM, 1997.

[39] A. Thor, T. Kirsten, and E. Rahm. Instance-based matching of hierarchical ontologies. In A. Kemper, H. Schöning, T. Rose, M. Jarke, T. Seidl, C. Quix, and C. Brochhaus, editors, *BTW*, volume 103 of *LNI*, pages 436–448. GI, 2007.

[40] K. Todorov and P. Geibel. Variable selection as an instance-based ontology mapping strategy. In H. R. Arabnia and A. Marsh, editors, *SWWS*, pages 3–9. CSREA Press, 2009.

[41] K. Todorov, P. Geibel, and K.-U. Kühnberger. Mining concept similarities for heterogeneous ontologies. In P. Perner, editor, *ICDM*, volume 6171 of *Lecture Notes in Computer Science*, pages 86–100. Springer, 2010.

[42] A. Tordai, J. van Ossenbruggen, and G. Schreiber. Combining vocabulary alignment techniques. In Y. Gil and N. F. Noy, editors, *K-CAP*, pages 25–32. ACM, 2009.

[43] A. Tordai, J. van Ossenbruggen, G. Schreiber, and B. J. Wielinga. Aligning large SKOS-like vocabularies: two case studies. In L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, editors, *Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010)*, volume 6088 of *Lecture Notes in Computer Science*, pages 198–212. Springer, 2010.

[44] J. Tuominen, M. Frosterus, K. Viljanen, and E. Hyvönen. ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, volume 5554, pages 768–780. Springer, May 31 - June 4 2009.

[45] O. Udrea, L. Getoor, and R. J. Miller. Leveraging data and structure in ontology integration. In C. Y. Chan, B. C. Ooi, and A. Zhou, editors, *SIGMOD Conference*, pages 449–460. ACM, 2007.

[46] L. van der Meij, A. Isaac, and C. Zinn. A web-based repository service for vocabularies and alignments in the cultural heritage domain. In L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, editors, *Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010)*, volume 6088 of *Lecture Notes in Computer Science*, pages 394–409. Springer, 2010.

[47] W. R. van Hage, A. Isaac, and Z. Aleksovski. Sample evaluation of ontology-matching systems. In R. Garcia-Castro, D. Vrandecic, A. Gómez-Pérez, Y. Sure, and Z. Huang, editors, *EON*, volume 329 of *CEUR Workshop Proceedings*, pages 41–50. CEUR-WS.org, 2007.

[48] D. Vizine-Goetz, C. Hickey, A. Houghton, and R. Thompson. Vocabulary mapping for terminology services. *J. Digit. Inf.*, 4(4), 2004.

[49] S. Wang, G. Englebienne, C. Gueret, S. Schlobach, A. Isaac, and M. Schut. Similarity features, and their role in concept alignment learning. In *Proceedings of the Fourth International Conference on Advances in Semantic Processing (SEMAPRO2010)*, 2010.

[50] S. Wang, G. Englebienne, and S. Schlobach. Learning concept mappings from instance similarity. In *Proceedings of the 7th International Semantic Web Conference (ISWC 2008)*, Karlsruhe, Germany, 2008.

[51] S. Wang, A. Isaac, B. Schopman, S. Schlobach, and L. van der Meij. Matching multi-lingual subject vocabularies. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009)*, Corfu, Greece, Sept. 27 - Oct. 2 2009.

[52] S. Wang, A. Isaac, L. van der Meij, and S. Schlobach. Multi-concept alignment and evaluation. In *Proceedings of the Second International Workshop on Ontology Matching, collocated with the 6th International Semantic Web Conference (ISWC 2007)*, Busan, Korea, November 11 2007.

[53] C. Wartena and R. Brussee. Instanced-based mapping between thesauri and folksonomies. In *International Semantic Web Conference*, volume 5318 of *Lecture Notes in Computer Science*, pages 356–370. Springer, 2008.

[54] L. Will. Costs of vocabulary mapping. High-level Thesaurus (HILT) Workshop, June 19 2001.

[55] E. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212, 1927.

[56] K. S. Zaiss. *Instance-based ontology matching and the evaluation of matching systems*. PhD thesis, Heinrich Heine Universität Düsseldorf, 2010.