Background Knowledge in Ontology Matching: A Survey

Jan Portisch a,b,*, Michael Hladik b and Heiko Paulheim a

a Data and Web Science Group, University of Mannheim, B6 26, 68159 Mannheim, Germany
E-mails: jan@informatik.uni-mannheim.de, heiko@informatik.uni-mannheim.de
b SAP SE, Dietmar-Hopp-Allee 16, 69190 Walldorf, Germany
E-mails: jan.portisch@sap.com, michael.hladik@sap.com

Editor: Jérôme Euzenat, INRIA Grenoble Rhône-Alpes, France
Solicited review: Seven Anonymous Reviewers

Abstract. Ontology matching is an integral part for establishing semantic interoperability. One of the main challenges within the ontology matching operation is semantic heterogeneity, i.e., modeling differences between the two ontologies that are to be integrated. The semantics within most ontologies or schemas are, however, typically incomplete because they are designed within a certain context which is not explicitly modeled. Therefore, external background knowledge plays a major role in the task of (semi-) automated ontology and schema matching.

In this survey, we introduce the reader to the general ontology matching problem. We review the background knowledge sources as well as the approaches applied to make use of external knowledge. Our survey covers all ontology matching systems that have been presented within the years 2004 – 2021 at a well-known ontology matching competition together with systematically selected publications in the research field. We present a classification system for external background knowledge, concept linking strategies, as well as for background knowledge exploitation approaches. We provide extensive examples and classify all ontology matching systems under review in a resource/strategy matrix obtained by coalescing the two classification systems. Lastly, we outline interesting and yet underexplored research directions of applying external knowledge within the ontology matching process.

Keywords: ontology matching, schema matching, background knowledge, data integration, semantic integration, knowledge graphs, ontologies

1. Introduction

Ontology matching is the non-trivial task of finding correspondences between entities of two or more given ontologies or schemas. It is an integral part to ensure semantic interoperability. The matching can be performed manually or through the use of an automated matching system. Ontology matching is a problem for Open Data (e.g. matching publicly available domain ontologies or interlinking concepts in the Linked Open Data Cloud1) as well as for private companies which need to integrate disparate data stores for transactional or analytical purposes.

A major challenge for matching ontologies is the fact that they are typically designed within a given context and deep background knowledge that is not explicitly expressed in the schema definition [1]. In order to automatize the ontology matching process, external background knowledge is therefore required so that the automated matching system can interpret for example

1see https://lod-cloud.net/
textual labels and descriptions of the elements within the schemas that are to be matched.

Current surveys in the ontology matching [2–5] and schema matching [6, 7] domain classify matching systems according to their matching technique (strongly influenced by Euzenat and Shvaiko [8, 9] as well as Rahm and Bernstein [10]) with minor or no emphasis at all on the background knowledge used.

In the area of context-based matching, i.e. matching with intermediate resources, Locoro et al. [11] present an abstract seven-step process for context-based matching together with an experimental evaluation of different parameter configurations. The proposed framework is flexible but experimentally focused on ontologies as background knowledge and a path- and logic-based exploitation approach. The survey at hand takes a broader look at the types of background sources and different exploitation strategies used in research including, for instance, unstructured data and statistical or neural approaches.

A recent survey by Trojahn et al. [12] provides a detailed perspective into foundational ontologies in ontology matching which includes, among other use cases, the exploitation of those for the task of matching domain ontologies. The survey presented here is broader in the sense that foundational ontologies are considered only as one kind of external background knowledge; it is narrower in the sense that it focuses purely on the use case of finding equivalence relations between schemas with additional background knowledge automatically.

Thiéblin et al. [13] review complex matching systems, i.e. systems that are capable of generating correspondences involving multiple entities, transformation functions, and logical constructors. The matching systems covered in their survey use different knowledge representation models (including table-based or document-based schemas, for instance). The systems are characterized based on the correspondence output and the underlying process type which generated the complex alignment. Background knowledge is not discussed and does not play a major role in the current implementations of complex matching systems. The survey at hand is complementary in the sense that it focuses on systems producing simple equivalence correspondences through the use of background knowledge.

This comprehensive survey reviews an extensive set of ontology matching and integration systems published in the last two decades in terms of the background knowledge used and in terms of the strategy that is applied to exploit the external background knowledge. It further covers the approaches used to link schema concepts to background knowledge. Based on the extensive collection of reviewed systems, we provide a comprehensive overview of background knowledge sources and strategies used in the past. Furthermore, this survey reveals a number of blind spots that have not yet been thoroughly explored.

In the following, the selection method for publications used in this survey is presented (Section 2.1). Afterwards, the core theoretical concepts are introduced in Section 3, namely schema matching and ontology matching (OM). In Section 4, background knowledge is defined, its usage in ontology matching system is analyzed, and the most used resources are presented. Thereupon, classification systems for background knowledge sources (Section 5), concept linking approaches (Section 6), and exploitation approaches (Section 7) are presented together with examples. In Section 8, we outline interesting directions for future work in the research field.

2. About this Survey

2.1. Selection of Publications

Search Parameters For this survey, we defined three search parameters: (Q1) “ontology matching”, (Q2) “ontology alignment”, and (Q3) “ontology mapping”. We queried publications via the dblp computer science bibliography (DBLP) without further filters. The search criteria have been intentionally chosen to be very broad since the usage of background knowledge is very often not indicated in the title or abstract of a paper.

We further manually added all matching systems that participated in the schema matching tracks of the ontology alignment evaluation initiative (OAEI, see Section 3.4) from its inception in 2004 until 2021 [14–31].

The number of retrieved papers for each search parameter can be found in Table 1. The bibtext files can be found in the GitHub repository of this survey.

2see https://dblp.org/

3Back then the competition was actually referred to as EON Ontology Alignment Contest.

4see https://github.com/janothan/bk-in-matching-survey/
De-Duplication  The bibtex files of all publications were gathered and loaded via the Zotero\textsuperscript{5} bibliographic management tool. The latter was used to detect duplicate publications based on the metadata of the papers. All scientific artifacts were exported as a CSV file including the metadata (title, authors, publication venue, date, etc.) for manual de-duplication.

The resulting set of papers constitutes the final set of publications used for identifying relevant works for this survey. In total, 1,814 papers were considered in this study.

Selection Process  In order to identify papers which are relevant for this survey, inclusion criteria (IC) and exclusion criteria (EC) were defined. The set of all papers was manually scanned in order to filter out publications not relevant for this survey. The complete list of inclusion and exclusion criteria is shown in Table 2. Every paper that is considered in this survey has to satisfy all inclusion criteria.

Papers considered in this survey had to be written in English language (C1), had to be accessible through the infrastructure of a large German research university (C2), and had not to be a duplicate of another paper (C3). It is important to note that multiple publications on the same topic (such as a matching system) do not qualify as duplicates despite their potentially large content overlap. This is rooted in the observation that there are often multiple versions and papers of a single matching system which evolves over time (for example AML [32] or LogMap [33]); in such cases, we always refer to the specific matching paper we mean in order to be precise rather than referencing the most current or most extensive paper published for the system in question.

We explicitly exclude works limited solely to instance matching or entity linking (C4). We further focus on matching systems that produce simple correspondences rather than complex ones (C5). Lastly, we only cover papers that present an actual system, i.e. a background knowledge-based (C6) ontology matching system implementation (C7) for which an evaluation is presented. The usage of the background knowledge must be appropriately documented (C8). In total, 341 papers fulfilled the inclusion criteria of this survey.

All matching systems were systematically evaluated in terms of (i) the background knowledge sources used, (ii) the strategy deployed to link ontology concepts to the background knowledge source, and (iii) the strategies the matching systems apply to exploit the background knowledge sources.

2.2. Figures and Data

All data points and code used for the quantitative analysis of this survey are available online.\textsuperscript{6} This includes statistical figures which are also available online in a higher resolution; they can further be regenerated with the provided Python code.

3. Schema Matching and Ontology Matching

3.1. The Schema Matching Problem within the Data Integration Process

Data Integration  Data integration (DI) describes the process to obtain uniform access over a set of heterogeneous and autonomous sources of data [34]. The process can be divided in four main parts [35] as depicted in Figure 1: (i) Schema Matching, (ii) Schema Translation, (iii) Record Linkage, and (iv) Data Fusion.

Schema Matching  Schema matching is an important and time consuming part within the data integration process. Out of the actions to carry out in order to integrate two given schemas (depicted in Figure 1), schema matching is the first step. Schema matching describes the process of finding the relations that hold between the elements of the schemas that are to be matched. The most important relation here is the equivalence relation. In this step, structural as well as semantic heterogeneity between the two schemas are bridged.

Schema Translation  Schema translation describes the process of deriving the translation function from one schema to the other schema.

\textsuperscript{5}see https://www.zotero.org/

\textsuperscript{6}see https://github.com/janothan/bk-in-matching-survey/
Criteria | Inclusion Criteria (IC) | Exclusion Criteria (EC)
---|---|---
C1 Language | The paper is written in English. | The paper is not written in English; the paper is written in English but heavily ungrammatical.
C2 Accessibility | The paper can be accessed through the infrastructure of the University of Mannheim without additional payment. | The paper cannot be accessed through the infrastructure of the University of Mannheim without additional payment.
C3 Duplication | Included are papers whose content is unique. This explicitly includes papers on the same matching system; for example, all OAEI LogMap papers are included in this survey rather than only the latest publication in order to carry out a thorough time analysis. | Excluded are papers with identical content such as preprints which are identical in content with their peer-reviewed publications or identical papers published in multiple venues.
C4 Ontology Matching System | The paper presents a matching system, i.e. a system which accepts two ontologies and returns an alignment. The matching system must be able to match ontologies (T-box). Papers which align schema and instances are also included. | The paper does not present a matching system which is able to match ontologies such as pure entity-linking or pure instance matching approaches.
C5 Simple Correspondences | The matching system produces simple correspondences. | The paper presents a matching system for complex matching.
C6 Background Knowledge | The matching system exploits some form of external knowledge. | The matching system presented does not use any external knowledge.
C7 Application/Evaluation | The paper presents a matching system which is evaluated on the task of ontology matching. | The paper merely describes a framework or a theoretical idea but lacks a concrete implementation regarding ontology matching.
C8 Level of Detail | The paper describes the use of background knowledge with an appropriate level of detail. | The usage of background knowledge is mentioned but it is unclear which knowledge source is used or how it is used.

Table 2
Inclusion and exclusion criteria for the papers in this survey.

![Diagram](image_url)  
Figure 1. Process for integrating two schemas, compiled from [35].

**Record Linkage**  
Record linkage describes the process of linking the records of instances of two schemas, i.e. finding equivalent records in disparate datasets.

**Data Fusion**  
Data fusion describes the process of resolving conflicting information concerning individual instances.

### 3.2. Schemas and Ontologies

The focus of this paper is a special case of the first step of the DI process, schema matching. It is important to note that a schema is not bound to a technology stack. It is, for example, possible that the same schema is implemented on different technology stacks such as different database types. Many formalization notations for schemas have evolved over time – for example in the area of (conceptual) entity relationship models Barker’s notation [36], IDEFIX [37] by the National Institute of Standards and Technology, or MERISE [38]. In semantic data modelling, data representation paradigms such as controlled vocabularies, taxonomies, knowledge graphs, among others, are...
used [39], all of which have been subsumed under the umbrella term of ontologies in different publications [40–44]. Hence, we conclude that most of the methods described for ontology matching can be more broadly understood as methods for matching semantic models in general [45].

3.3. The Ontology Matching Problem

**Ontology**  The term ontology has roots in philosophy and describes the study of being. In the computer science domain, an ontology is a “formal, explicit specification of a shared conceptualization”7, i.e. an abstract model of real-world concepts that is represented in a computer-readable way and is shared by a group of stakeholders. The definition is technology-independent; conceptually, even an XML Schema could be interpreted as an ontology [48]. While multiple ontology languages are available, most ontologies are typically defined in the W3C Web Ontology Language (OWL). An OWL ontology consists of different element types: classes/concepts (C), individuals/instances (I), relations (R), data types (DT), and data values (DV). Hence, we define an ontology O as $O = \{C, I, R, DT, DV\}$.

**Ontology Matching**  Given two ontologies $O_1$ and $O_2$, the matching problem describes the task of finding an alignment $A$ between $O_1$ and $O_2$. An alignment is a set of correspondences whereby a correspondence is a triple in the form $(e_1, e_2, r)$ with $e_1 \in O_1$ and $e_2 \in O_2$ being elements of the ontologies to be matched and $r$ being the relation that holds between the two elements. Examples for the relation are equivalence ($\equiv$) or inclusion ($\subseteq$). A correspondence may optionally have an explanation $e$ and a confidence value $c$ assigned to it and is, therefore, sometimes also described as a quintuple in the form $(e_1, e_2, r, c, e)$. Two types of correspondences are distinguished: Simple ones, that link one element from $O_1$ to one element from $O_2$ and complex ones, i.e. correspondences that contain logical constructors or transformation functions [49].

A matching system can be seen as a function $f(O_1, O_2, A', p, b) = A$. Variable $A'$ refers to an existing alignment (which may be empty), $p$ specifies additional parameters for the matching process, and $b$ represents external background knowledge sources used in the matching process. [50] For this survey, it is of particular interest how $b$ is used in $f$.

**Ontology Integration**  Multiple interpretations exist to the terms ontology integration and ontology merging. We follow the proposal from Osman et al. [51] in this survey and regard ontology merging as a special case of ontology integration:

Ontology integration (also referred to as ontology enrichment, ontology inclusion, or ontology extension) describes the process of extending a given target ontology $O_T$ with another (source) ontology $O_S$ given an alignment $A_{S\rightarrow T}$ between $O_S$ and $O_T$: $\text{Integrate}(O_S, O_T, A_{S\rightarrow T}) = O_T$. A special case is ontology merging where given two ontologies $O_1$ and $O_2$, a third ontology $O_3$ is derived given an alignment $A_{1\rightarrow 2}$ between $O_1$ and $O_2$: $\text{Merge}(O_1, O_2, A_{1\rightarrow 2}) = O_3$. According to Osman et al. [51], the ontology integration process can be generally seen as a four step process:

1. **Pre-processing Phase**
2. **Matching Phase**
3. **Merging Phase**
4. **Post-processing Phase**

**Pre-processing** describes preparing the ontology files that are to be matched, e.g. by converting them into the same uniform representation. The **Matching Phase** describes the ontology matching process as outlined in the previous paragraph. The **Merging Phase** describes the execution of the $\text{Integrate}/\text{Merge}$ operator, and the **Post-processing Phase** summarizes various amendments to the resulting ontology to improve its quality, such as resolving cycles, or coherence and conservancy violations. For details, we refer the reader to the comprehensive survey by Osman et al. [51].

In this article, we also cover papers and systems which address the ontology integration problem where background knowledge plays a significant role in the matching phase. In figures and tables, those systems are noted with a subscript $I$ such as MoA/I.

3.4. The Ontology Evaluation Initiative since 2004

**About the OAEI**  Schema matching can be performed manually, through an automated matching system, or in a hybrid environment. For systematically evaluating the latter two cases, the Ontology Alignment Evaluation Initiative (OAEI)8 is running campaigns every year since 2004. Unlike other evaluation cam-

---

7This definition is a merge of previous definitions by Gruber [46] and Borst [47].

8Originally called $r$ but renamed for better clarity here.
paigns where researchers submit datasets as solutions to report their results (such as Kaggle\textsuperscript{10}), the OAEI requires participants to submit a matching system, i.e. an implemented and packaged matching system, which is then executed on-site.\textsuperscript{11} In order to do so, multiple frameworks and platforms for standardized matcher development, packaging, and evaluation have been developed and are used by OAEI participants, namely the Alignment API\textsuperscript{[52]} format and framework, the SEALS\textsuperscript{[53, 54]} and HOBBIT\textsuperscript{[55]} packaging and evaluation platforms as well as MELT\textsuperscript{[56–58]}, a framework for matcher development, packaging, and evaluation which also integrates with the aforementioned frameworks. After the evaluation, the results are publicly reported. The individual matching tasks are referred to as test cases which are bundled in tracks. Originally, the OAEI started with plain ontology matching tasks focused on simple alignments with an equality relation, i.e. a correspondence which contains only one entity from the source ontology and one ontology from the target ontology and where \( r = \text{equivalence} \). More recently, new tracks have been introduced such as the Knowledge Graph Track\textsuperscript{[59, 60]} which combines schema and instance matching tasks. The most transparent way of presenting and benchmarking a new matching system is the participation in an OAEI campaign – however, most datasets are also available for download\textsuperscript{12} and can be used outside of OAEI campaigns to evaluate matching systems.

**OAEI Tracks** Figure 2 summarizes all OAEI schema matching tracks since the inception of the initiative. As visible in the figure, some older tracks have been discontinued\textsuperscript{13} while new tracks have also been introduced. All current schema matching tracks that were evaluated in the OAEI 2020 and 2021 are listed in Table 3 together with a quick description and the best performing system of the corresponding year.

\textsuperscript{10}see https://www.kaggle.com/

\textsuperscript{11}Prior to 2010, participants submitted resulting alignments directly. The submission of packaged tools (at first in the form of URLs of Web services running on the participants’ site) instead of results was started in 2010. Since 2012, the submission of packaged tools is the standard evaluation procedure at the OAEI.

\textsuperscript{12}see https://dwslab.github.io/melt/track-repository

\textsuperscript{13}The discontinuation of tracks is often due to missing track organizers. Reasons may be the high effort connected to evaluating other researchers’ matching systems and writing summarizing reports or a change in the research focus. However, most track data is still available for download and for further usage.

**OAEI Matching Systems** Since 2004, many matching systems have been submitted and evaluated. Figures 3 and 4 list all matching systems that have been evaluated in OAEI schema matching campaigns\textsuperscript{14} since its inception on the y-axis; the x-axis represents a time line and the black bars represent the time frame in which the systems have participated in the campaigns. As visible in the figures, many systems have been evaluated in multiple campaigns. For this survey, all of the listed matching systems that are used for schema matching have been examined in terms of what background knowledge source is used if any, how a connection between the ontologies and the background knowledge source is established, and how the background knowledge source is exploited.

Figure 5 reveals that over the years the number of participating schema matching systems to date has slightly dropped from the peak in the year 2012 albeit the current participation total is still comparatively high compared to the early days of the initiative.\textsuperscript{15}

Table 3 lists all schema matching tracks from 2020 and 2021 together with the best performing system and the background knowledge sources used by those. As visible in the table, all those systems make use of external knowledge datasets. AML, which scores as best performing system in multiple tracks, exploits multiple external knowledge sources.

4. Background Knowledge in Ontology Matching

4.1. Background Knowledge

We define background knowledge in matching as any knowledge source that is external to the matching process and is used to obtain the final alignment. Hence, within the matching process, external knowledge can be used in the form of an existing alignment

\textsuperscript{14}The tracks which were considered are listed in Figure 2. Figures 3 and 4 do not include other evaluation tracks such as team participations in the SemTab\textsuperscript{[76]} track. Due to very high similarity, the following matching systems have been merged in the figure: NLM\textsuperscript{[77]} and AOAS\textsuperscript{[78]}, Agreement Maker and AMExt\textsuperscript{[82]} (both described in [79]), as well as GeRoMe\textsuperscript{[80, 81]} and GeRoMe-SMB\textsuperscript{[82]}. Figure 5 has been compiled from Figures 3 and 4, hence the concrete number of schema matching systems is counted each year excluding pure instance matching systems. The OAEI does not calculate this statistic. In addition, we found that over the years the OAEI counted inconsistently with regards to participation (for example counting participating teams in 2012 but matching systems in 2013 on their results Web page).
<table>
<thead>
<tr>
<th>Track</th>
<th>Track Description</th>
<th>Best Performing System in the OAEI 2020</th>
<th>Best Performing System in the OAEI 2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anatomy [61]</td>
<td>An alignment between the Adult Mouse Anatomy and a part of the NCI Thesaurus is to be found.</td>
<td>AML [62] (Uberon, DOID, MeSH, WordNet, Microsoft Translator, OBO logical definitions)</td>
<td>AML [63] (Uberon, DOID, MeSH, WordNet, Microsoft Translator, OBO logical definitions)</td>
</tr>
<tr>
<td>Conference [64]</td>
<td>16 ontologies from the conference domain have to be matched.</td>
<td>VeeAlign [65] (Google Universal Sentence Encoder)</td>
<td>AML [63] (see above)</td>
</tr>
<tr>
<td>Multifarm [66]</td>
<td>7 conference ontologies translated into 8 languages (+ English) have to be matched.</td>
<td>AML [62] (see above)</td>
<td>AML [63] (see above)</td>
</tr>
<tr>
<td>LargeBio</td>
<td>An alignment between 3 large bio ontologies is to be found.</td>
<td>AML [62] (see above)</td>
<td>AML [63] (see above)</td>
</tr>
<tr>
<td>Phenotype [67]</td>
<td>An alignment between two disease and two phenotype ontologies is to be found.</td>
<td>LogMapBio [68] (Bioportal)</td>
<td>LogMapBio [69] (Bioportal)</td>
</tr>
<tr>
<td>Biodiversity and Ecology [70]</td>
<td>4 matching tasks from the biodiversity and ecology domains.</td>
<td>AML [62] (see above)</td>
<td>AML [63] (see above)</td>
</tr>
<tr>
<td>Knowledge Graph [71]</td>
<td>5 matching tasks consisting of knowledge graphs extracted from fandom.com.</td>
<td>Wiktionary Matcher [72] (Wiktionary/DBnary)</td>
<td>Wiktionary Matcher [73] (Wiktionary/DBnary)</td>
</tr>
<tr>
<td>Common Knowledge Graph [74]</td>
<td>An alignment between the classes of two large, automatically constructed knowledge graphs is to be found.</td>
<td>–</td>
<td>KGMatcher [75] (BERT, Google language model)</td>
</tr>
</tbody>
</table>

Table 3

Depicted are all schema matching tasks of the OAEI 2020 and 2021 together with the best performing systems in terms of $F_1$. For the conference track, the rar2-M3 results have been used to determine the best system. For tracks with multiple tasks that do not name a best performing system (LargeBio, phenotype), the average position in all tasks was chosen as criterion to determine the best performing system here. The Common Knowledge Graph track was first evaluated in 2021.
Figure 2. OAEI schema matching tracks since the inception of the initiative. Explicitly excluded are complex matching tracks and instance matching tracks. The knowledge graph track is not a pure schema matching task but a combined one where schemas and instances have to be matched simultaneously. The library track has been organized multiple times with completely different datasets and by different researchers using the same track name. Therefore, the track streams have been divided in three groups (A, B, C).

Background knowledge can significantly improve the performance of ontology matching systems. This is clearly visible by analyzing different OAEI systems: When comparing LogMap and LogMapBio [69] in the OAEI 2021 campaign, for instance, it can be seen that the latter system scores a significantly higher recall on the OAEI Anatomy dataset. Other examples can be found through a comparison of AML [83] and Gomma in the 2013 campaign: Both systems participated in two configurations – with and without background knowledge. On the Anatomy track, the background knowledge configurations significantly outperformed all other systems in terms of recall and $F_1$. Another indicator for the value of background knowledge is the fact that all best performing schema matching systems of the 2020 and 2021 campaigns use external background knowledge (see Table 3).

In [85], Faria et al. evaluate strategies for matching biomedical ontologies. The experiments show a clear performance increase when background knowledge is used. In terms of exploitation strategies, the authors recommend to use cross-references (if available) over lexical expansion.

While evaluating an approach to build a background knowledge resource for ontology matching, Annane et al. [86] also analyze the performance of the YAM++ matching system with and without background knowledge finding that the matcher configuration which uses background knowledge significantly outperforms the version without additional resources. They report that the better performance is mainly due to a higher recall.

In an extensive survey on the systems participating in the OAEI Anatomy track from 2007 to 2016, Dragisic et al. report that “[f]or the systems that participated with a version using biomedical auxiliary sources and a version not using biomedical auxiliary sources, the F-measure for the one with biomedical auxiliary sources was always higher” [87].

Missing background knowledge was named as one of the 10 challenges for ontology matching in 2008 [88]; this was re-affirmed in 2013 [1] and it is still under active research.

16There is no results paper for the OAEI 2013 participation of Gomma. However, the system is described in the paper of the 2012 campaign [84].
Figure 3. All OAEI schema matching systems (which participated in the tracks listed in Figure 2) and their evaluation time frame since the inception of the OAEI; Part 1 of 2 from 2012 - 2021.
4.2. Background Knowledge Selection in Ontology Matching

As there are often multiple potentially beneficial sources of background knowledge available for ontology matching, some authors propose heuristics to determine the benefit of a background knowledge source in order to select one before performing the match operation. Nasser et al. [89] define four criteria to automatic background knowledge selection:

1. type independence: A selection system should be capable to handle various serialization formats.
2. domain independence: A selection system should be domain-independent and be able to select sources for any domain.
3. multilingualism: A selection system should be language-independent, i.e. support cross-lingual ontology matching.
4. optimality: A selection system should return the best background knowledge source from the corpus.

Based on their universal requirements, they propose an approach which models the selection task as information retrieval problem. Ontologies and background sources are indexed using TF-IDF; the ontologies are
then regarded as query on the background knowledge sources.

In the LogMapBio system, Chen et al. [90] apply a relatively simple lexical algorithm to identify suitable mediating ontologies from BioPortal [91, 92]. In the OAEI 2020 campaign, the system achieved a significantly higher recall and $F_1$ measure than the classic LogMap matching system.

Faria et al. [93] propose a heuristic called $\text{Mapping Gain}$ which is based on the number of additional correspondences found given a baseline alignment. Quix et al. [94] use a keyword-based vector similarity approach to identify suitable background knowledge sources. Similarly, Hartung et al. [95] introduce a metric, called $\text{effectiveness}$, that is based on the mapping overlap between the ontologies to be matched.

4.3. Background Knowledge in Ontology Matching Over Time

Tables 4 to 7 list all background knowledge sources that have been used by the systems evaluated in this survey together with the actual systems that use the corresponding knowledge source. As multiple papers exist for some systems, the first documented usage of the knowledge source by the matching system is referenced. Consequently, there is no guarantee that the latest system still uses the specified sources. WeSeE Match, for example, used the Microsoft Bing search
engine in its 2012 version [96] but switched to the FARO Web Search framework in 2013 [97]. Therefore, different papers are referenced for the system. For each knowledge source, the systems in column Used by System are ordered according to publication year. Since this survey covers a large time period, not all resources used in the past are still available; therefore, column Resource Available indicates whether the resource is still available to researchers. For the frequent usage of WordNet [98], systems that use this source are listed in Tables 8 and 9 which are organized according to the same methodology as Tables 4 to 7. Tables 4 to 9 also include some non-OAEI matching systems (indicated by italics).

Figure 6 shows the cumulative usage of background knowledge sources that have been referenced in at least four different publications. The by far most often used external knowledge resource is WordNet [98]. Further often used resources are the Unified Medical Language System (UMLS) [99] as well as the Microsoft Bing Translation API. When looking at the distribution of the usage counts in Figure 6, a power-law distribution can be recognized: Most systems use the same knowledge source; although many knowledge sources exist, most are used only by very few systems. It is important to note that the long-tail in the distribution is actually much longer as shown in the figure because the latter only lists sources used by at least four different matching system publications.

In Figure 7, background knowledge source usage is plotted over time. As in the figure before, only sources are depicted which are used at least four times by the papers included in this survey. What is visible from the figure (and also from Tables 4, 5, 6, 7, 8, and 9) is that background knowledge has been used from very early on. In the first OAEI in 2004, for example, the OWL-Lite Alignment (OLA) [100] matching system already uses WordNet to retrieve synonym sets. A look at the usage over time (Figure 7) reveals that only few sources have been used in the early days of ontology matching. With a progression of time, more and more resources are evaluated. However, only few sources show a consistently high application, in particular WordNet, the Microsoft Translation API, UBERON, and UMLS. We can also observe spikes of usage, i.e., a resource has been used within a short timeframe in multiple papers but not afterwards: Examples here are Swoogle [101], a Semantic Web search engine17, or the Google Search API.

4.4. Most Used Background Knowledge Resources

In the following, the ten most used external resources in ontology matching (see Figure 6) are shortly introduced.

**WordNet** WordNet is a database of English words grouped in sets which represent a particular meaning, so called *synsets*; further semantic relationships, such as *hypernymy*18 and *hyponymy*19, also exist in the database. The resource is publicly available. In fact, WordNet is so heavily used that there exists a dedicated survey paper titled “A survey of exploiting WordNet in ontology matching” [361]. The resource is under a permissive license can also be used for commercial purposes.

**Bing/Microsoft Translation API** The Microsoft Translation API20, formerly known as Bing Translation API, allows, among other functions such as language detection, for translating a text string from a source language to a target language. The cloud API can be accessed through any programming language. Since the service is provided in a cloud infrastructure, the translation service is continuously improved. These changes impede reproducibility of matching systems using the API. The service is not free, but as of 2021, 2 million characters of translation/detection per month are not charged.

**UMLS** The Unified Medical Language System (UMLS) is a manually-built compendium of vocabularies in the biomedical domain. The UMLS is maintained by the United States National Library of Medicine (NLM). UMLS can be used without charge but a download21 requires a registration at the NLM.

---

17 The search engine is not online anymore.

18 A *hypernym* or *hypernym* is a concept which is superordinate to another one. In computer science, it is often represented as an IS-A relationship. For example, *animal* is a hypernym of *cat* [360].

19 A *hyponym* is a concept which is subordinate to another one. In computer science, it is often represented as an IS-A relationship. For example, *cat* is a hyponym of *animal* [360].

20 see https://wordnet.princeton.edu/download

21 see https://wordnet.princeton.edu/license-and-commercial-use

22 see http://www.microsoft.com/translator


24 see https://www.nlm.nih.gov/research/umls/index.html
<table>
<thead>
<tr>
<th>Knowledge Source</th>
<th>Source Description</th>
<th>Resource Available</th>
<th>Used by System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apertium [102]</td>
<td>A free open-source platform for machine translation.</td>
<td>yes</td>
<td>Bella et al. (2017) [103]</td>
</tr>
<tr>
<td>BabelNet [104]</td>
<td>Multilingual, large knowledge graph derived through the integration of multiple knowledge sources such as WordNet and Wikipedia.</td>
<td>yes</td>
<td>LYAM++ (2015) [105], Helou et al. (2016) [106], Biniz et al. (2017) [107], EVOCROS (2018) [108], Kolyvakis et al. (2018) [109]</td>
</tr>
<tr>
<td>BERT [110]</td>
<td>A transformer-based language model.</td>
<td>yes</td>
<td>Neutel et al. (2021) [111], KGMatcher (2021) [75], Fine-TOM (2021) [112], TOM (2021) [113]</td>
</tr>
<tr>
<td>Big Huge Thesaurus</td>
<td>Web API for synonyms and antonyms.</td>
<td>yes</td>
<td>SOCOM++ (2012) [114], HotMatch (2012) [115]</td>
</tr>
<tr>
<td>Bing Search Engine API</td>
<td>Cloud API for the Microsoft Bing Web search engine.</td>
<td>yes</td>
<td>Fu et al. (2011) [116], WeSeE Match (2012) [96], SOCOM++ (2012) [114], SYNTHESIS (2013) [117]</td>
</tr>
<tr>
<td>Bing Translator / Microsoft Translator</td>
<td>Cloud API for the Microsoft Bing translation service.</td>
<td>yes</td>
<td>SOCOM (2010) [118], Spoehr et al. (2011) [119], WeSeE Match (2012) [96], YAM++ (2012) [120], Koukourikos et al. (2013) [121], AML (2014) [122], XMap (2014) [123], Kachroodi et al. (2014) [124], LogMap (2015) [125], CLONA (2015) [126], KEPLER (2017) [127], Kachroodi &amp; Yahia (2018) [128]</td>
</tr>
<tr>
<td>BioBERT [129]</td>
<td>A language model pre-trained on medical text.</td>
<td>yes</td>
<td>MEDTO (2021) [130]</td>
</tr>
<tr>
<td>ConceptNet [135]</td>
<td>A freely-available word graph collected from multiple sources.</td>
<td>yes</td>
<td>Kolyvakis et al. (2018) [109]</td>
</tr>
<tr>
<td>Cooking Dictionary</td>
<td>A collection of term definitions in the cooking domain.</td>
<td>yes</td>
<td>van Hage et al. (2005) [136]</td>
</tr>
<tr>
<td>DBpedia [137]</td>
<td>A knowledge graph extracted from Wikipedia info boxes.</td>
<td>yes</td>
<td>BLOOMS (2010) [138], LDOA (2011) [139], Grütze et al. (2012) [140]</td>
</tr>
<tr>
<td>DOLCE [143]</td>
<td>The descriptive ontology for linguistic and cognitive engineering (DOLCE) is an upper ontology.</td>
<td>yes</td>
<td>Mascardi et al. (2010) [144], Davarpanah et al. (2015) [145]</td>
</tr>
<tr>
<td>FAROO Web Search</td>
<td>A framework for Web search.</td>
<td>yes</td>
<td>WeSeE Match (2013) [97]</td>
</tr>
<tr>
<td>fastText model</td>
<td>A model trained with facebook’s AI research (FAIR) fastText [146] framework.</td>
<td>yes</td>
<td>OntoConnect (2020) [147], Neutel et al. (2021) [111]</td>
</tr>
</tbody>
</table>

Table 4: Knowledge sources and matching systems that use them part 1 of 4. Referenced is the first documented usage by the matching system. Systems that did not participate in the OAEI are italicized. Named systems are referred to using their system name.
<table>
<thead>
<tr>
<th>Knowledge Source</th>
<th>Source Description</th>
<th>Resource Available</th>
<th>Used by System</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grofl et al. (2011) [149]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GOMMA (2012) [84]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Petrov et al. (2013) [150]</td>
</tr>
<tr>
<td>FMA</td>
<td>The Foundational Model of Anatomy (FMA).</td>
<td>yes</td>
<td>AOAS (2007) [78]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Grofl et al. (2011) [149]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GOMMA (2012) [84]</td>
</tr>
<tr>
<td>Google NNLM</td>
<td>A neural text embedding model available through TensorFlow Hub by Google.</td>
<td>yes</td>
<td>KGMatcher (2021) [75]</td>
</tr>
<tr>
<td>Freelang</td>
<td>A translation API (available as offline and as online version).</td>
<td>yes</td>
<td>Medley (2012) [151]</td>
</tr>
<tr>
<td>Google Search API</td>
<td>Cloud API for the Google Web search engine.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Google Translation API</td>
<td>A translation Web API by Google.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Google Universal Sentence Encoder [163, 164]</td>
<td>Pre-trained encoder by Google (monolingual [163] and multilingual [164]).</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Google Word2Vec Vectors</td>
<td>Word2vec models by Google.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>HowNet [167]</td>
<td>An online sememe knowledge base in Chinese and English.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>ImageNet</td>
<td>A large database of images.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>iTranslate4</td>
<td>API for machine translation.</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>KGvec2go [171]</td>
<td>Pre-trained RDF2Vec embeddings.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Lanes API</td>
<td>Language Analysis Essentials (LANES) API. Does not seem to be online anymore.</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Medical Subject Headings (MeSH) [173]</td>
<td>The Medical Subject Headings (MeSH) are a controlled vocabulary thesaurus.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Medline</td>
<td>Bibliographic database of the National Library of Medicine. Medline is a subset of PubMed.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>MyMemory API</td>
<td>A translation REST API provided by translated.com.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Ontology Lookup Service (OLS)</td>
<td>Repository and Web APIs for biomedical ontologies.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>OpenCyc [178]</td>
<td>Open-source version of the Cyc knowledge base by Cycorp. No longer available.</td>
<td>no</td>
<td></td>
</tr>
<tr>
<td>Paraphrase DB (PPDB) [179]</td>
<td>A very large collection of paraphrases.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>PubMed</td>
<td>Bibliographic database maintained by the National Library of Medicine.</td>
<td>yes</td>
<td></td>
</tr>
</tbody>
</table>

Table 5
Knowledge sources and matching systems that use them from part 2 of 4. Referenced is the first documented usage by the matching system. Systems that did not participate in the OAEI are italicized. Named systems are referred to using their system name.
<table>
<thead>
<tr>
<th>Knowledge Source</th>
<th>Source Description</th>
<th>Resource Available</th>
<th>Used by System</th>
</tr>
</thead>
<tbody>
<tr>
<td>RadLex</td>
<td>A radiology lexicon</td>
<td>yes</td>
<td>Groß et al. (2011) [149]</td>
</tr>
<tr>
<td>SAP Term</td>
<td>Definitions of terms in SAP software</td>
<td>not publicly</td>
<td>DESKMatcher (2020) [148]</td>
</tr>
<tr>
<td>SBERT [183]</td>
<td>A BERT modification so that similarity can be determined via cosine distance</td>
<td>yes</td>
<td>MEDTO (2021) [130]</td>
</tr>
<tr>
<td>SDL FreeTranslation</td>
<td>An online translation service</td>
<td>no</td>
<td>SOCOM (2010) [160]</td>
</tr>
<tr>
<td>SPECIALIST Lexicon</td>
<td>Contains common English words as well as biomedical vocabulary.</td>
<td>yes</td>
<td>LogMap (2018) [185]</td>
</tr>
<tr>
<td>SUMO [186]</td>
<td>The suggested upper merged ontology (SUMO), an upper ontology.</td>
<td>yes</td>
<td>Mascardi et al. (2010) [144]</td>
</tr>
<tr>
<td>synonyms-fr.com</td>
<td>A Web service to retrieve French synonyms and antonyms.</td>
<td>yes</td>
<td>Fu et al. (2011) [116]</td>
</tr>
<tr>
<td>UMLS [99]</td>
<td>The unified medical language system is a compendium of vocabularies in the biomedical domain.</td>
<td>yes</td>
<td>GOMMA (2012) [84]</td>
</tr>
<tr>
<td>Universal Knowledge Core (UKC)</td>
<td>A multilingual lexical resource.</td>
<td>yes</td>
<td>NuSM (2017) [103]</td>
</tr>
<tr>
<td>WebIsALOD [205, 206]</td>
<td>Web-extracted hypernymy relations provided as an RDF knowledge graph.</td>
<td>yes</td>
<td>ALOD2Vec Matcher (2018) [207]</td>
</tr>
<tr>
<td>Webtranslator API</td>
<td>A Java translation API</td>
<td>yes</td>
<td>AUTOMS (2012) [208]</td>
</tr>
<tr>
<td>Wikipedia Corpus</td>
<td>Text corpus of the online encyclopedia Wikipedia.</td>
<td>yes</td>
<td>CIDER-CL (2013) [209]</td>
</tr>
</tbody>
</table>

Table 6

Knowledge sources and matching systems that use them from part 3 of 4. Referenced is the first documented usage by the matching system. Systems that did not participate in the OAEI are italicized. Named systems are referred to using their system name.
<table>
<thead>
<tr>
<th>Knowledge Source</th>
<th>Source Description</th>
<th>Resource Available</th>
<th>Used by System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia MediaWiki API</td>
<td>Web API of the online encyclopedia Wikipedia.</td>
<td>yes</td>
<td>BLOOMS (2010) [138, 212]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SOCOM (2010) [118]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Fu et al. (2011) [116]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>WikiMatch (2012) [213]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>OntoEmma (2018) [176]</td>
</tr>
<tr>
<td>Wikisynonyms</td>
<td>Semantic lexicon built from Wikipedia redirects.</td>
<td>yes</td>
<td>Kolyvakis et al. (2018) [109]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DeepAlignment (2018) [180]</td>
</tr>
<tr>
<td>Wiktionary</td>
<td>A community-built dictionary; an RDF version [214]</td>
<td>yes</td>
<td>Lin &amp; Krizhanovsky (2011) [215]</td>
</tr>
<tr>
<td></td>
<td>is also available.</td>
<td></td>
<td>Wiktionary Matcher (2019) [216]</td>
</tr>
<tr>
<td>WordNet [98]</td>
<td>A well-known database of English synsets.</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WordsAPI</td>
<td>A Web API for (English) word definitions, multiple word relations, and more.</td>
<td>yes</td>
<td>Hnatkowska et al. (2021) [217]</td>
</tr>
<tr>
<td>YAGO [218]</td>
<td>A large knowledge base extracted from multiple sources.</td>
<td>yes</td>
<td>Todorov et al. (2014) [211]</td>
</tr>
<tr>
<td>Yahoo Image Search</td>
<td>A search engine for images on the Web.</td>
<td>yes</td>
<td>Doulavakias et al. (2015) [170]</td>
</tr>
<tr>
<td>Yahoo Search</td>
<td>A search engine for the Web.</td>
<td>yes</td>
<td>Vazquez &amp; Swoboda (2007) [189]</td>
</tr>
<tr>
<td>Yandex Translation API</td>
<td>A translation Web API by the Yandex search engine.</td>
<td>yes</td>
<td>CroLOM (2016) [219]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>SimCat (2016) [220]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Ibrahim et al. (2020) [221]</td>
</tr>
</tbody>
</table>

Table 7
Knowledge sources and matching systems that use them from part 4 of 4. Referenced is the first documented usage by the matching system. Systems that did not participate in the OAEI are italicized. Named systems are referred to using their system name.

UBERON  In the anatomy domain, the Uber-anatomy ontology (UBERON) [191, 192] is an ontology for multiple species comprising of more than 13,000 classes (as of 2021). Since UBERON defines a canonical model, it can be used as a “hub ontology” to solve various integration problems in the anatomy domain. The ontology can be used on its own but also in combination with other anatomical ontologies such as the Foundational Model of Anatomy (FMA). Particularly the bridging ontologies which connect UBERON to other ontologies (such as UBERON to FMA) make the resource interesting for the task of ontology matching in this domain. UBERON is publicly available and can be directly downloaded\(^{25}\) without any registration.

Google Translation API  The Google Translation API\(^{26}\) is very similar to the Microsoft Translation API: It is also a continuously improved cloud service. The

---

\(^{25}\) see [http://uberon.org](http://uberon.org)

\(^{26}\) see [https://cloud.google.com/translate](https://cloud.google.com/translate)
<table>
<thead>
<tr>
<th>Knowledge Source</th>
<th>Used by System</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLA (2004) [100]</td>
<td>Cardoso et al. (2008) [222]</td>
</tr>
<tr>
<td>MoAy (2005) [227]</td>
<td>Fatemi et al. (2008) [228]</td>
</tr>
<tr>
<td>Mongiello &amp; Totaro (2005) [232]</td>
<td>Lera et al. (2008) [233]</td>
</tr>
<tr>
<td>DSSim (2006) [239]</td>
<td>Xia et al. (2009) [240]</td>
</tr>
<tr>
<td>Alekovski et al. (2006) [243, 244]</td>
<td>Eff2Match (2010) [245]</td>
</tr>
<tr>
<td>Nagy et al. (2006) [261]</td>
<td>LogMap (2011) [261]</td>
</tr>
<tr>
<td>Trojahn et al. (2006) [265]</td>
<td>OMRReasoner (2011) [266]</td>
</tr>
<tr>
<td>Wang et al. (2006) [267]</td>
<td>Optima (2011) [268]</td>
</tr>
<tr>
<td>Kim et al. (2006) [269]</td>
<td>YAM++ (2011) [270]</td>
</tr>
<tr>
<td>Wang et al. (2006) [271]</td>
<td>Liu &amp; Krizhanovsky (2011) [271]</td>
</tr>
<tr>
<td>ASMOV (2007) [198]</td>
<td>Sadaqat et al. (2011) [272]</td>
</tr>
<tr>
<td>Tan &amp; Lambrix (2007) [278]</td>
<td>Liu et al. (2012) [279]</td>
</tr>
<tr>
<td>Trojahn et al. (2007) [280]</td>
<td>Acampora et al. (2012) [281]</td>
</tr>
<tr>
<td>Jin et al. (2007) [283]</td>
<td>Fernández et al. (2012) [283]</td>
</tr>
<tr>
<td>UFOMe (2007) [288]</td>
<td>Song et al. (2012) [289]</td>
</tr>
<tr>
<td>Alasoud et al. (2008) [292]</td>
<td>Galic et al. (2013) [293]</td>
</tr>
<tr>
<td>Jeong-Woo et al. (2008) [294]</td>
<td>MAPiOM (2013) [295]</td>
</tr>
<tr>
<td>Ichise (2008) [304]</td>
<td>SPHeRe (2013) [305]</td>
</tr>
</tbody>
</table>

Table 8

Matching systems using WordNet; Part 1 of 2. Referenced is the first documented usage by the matching system. Systems that did not participate in the OAEI at some point in time are italicized. Ontology integration systems are indicated by a subscript I. Named systems are referred to using their system name.
Matching systems using WordNet; Part 2 of 2. Referenced is the first documented usage by the matching system. Systems that did not participate in the OAEI at some point in time are italicized. Ontology integration systems are indicated by a subscript \textsuperscript{I}. Named systems are referred to using their system name.

<table>
<thead>
<tr>
<th>Knowledge Source</th>
<th>Used by System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Petrov et al. (2013)</td>
<td>Bulygin et al. (2018) [165]</td>
</tr>
<tr>
<td>Fang et al. (2013)</td>
<td>ONTMAT\textsuperscript{I} (2019) [313]</td>
</tr>
<tr>
<td>UFOM (2014) [314]</td>
<td>Lily (2020) [197]</td>
</tr>
<tr>
<td>Todorov et al. (2014)</td>
<td>WeGO++ (2019) [315]</td>
</tr>
<tr>
<td>Xue et al. (2014)</td>
<td>Bulygin &amp; Stupnikov (2019) [166]</td>
</tr>
<tr>
<td>Jadboonlou et al. (2014)</td>
<td>Biniz &amp; Fakir (2019) [320]</td>
</tr>
<tr>
<td>InsMT/InsMTL (2014)</td>
<td>WeGO++ (2019) [315]</td>
</tr>
<tr>
<td>Chaker et al. (2014)</td>
<td>Yang (2019) [325]</td>
</tr>
<tr>
<td>Schull &amp; Roos (2014)</td>
<td>Ibrahim et al. (2020) [221]</td>
</tr>
<tr>
<td>ServOMBI (2015) [327]</td>
<td>Real et al. (2020) [174]</td>
</tr>
<tr>
<td>DKP-AOM (2015) [328]</td>
<td>Xue &amp; Chen (2020) [329]</td>
</tr>
<tr>
<td>Kiren &amp; Shoaib (2015)</td>
<td>Lv et al. (2021) [331]</td>
</tr>
<tr>
<td>Wang (2015) [334]</td>
<td>Xue et al. (2021) [335]</td>
</tr>
<tr>
<td>Xue et al. (2015) [336–340]</td>
<td></td>
</tr>
<tr>
<td>Benzissa et al. (2015)</td>
<td></td>
</tr>
<tr>
<td>Schull &amp; Roos (2015)</td>
<td></td>
</tr>
<tr>
<td>ALIN (2016) [343]</td>
<td></td>
</tr>
<tr>
<td>CroLOM (2016) [219]</td>
<td></td>
</tr>
<tr>
<td>CroMatcher (2016) [195]</td>
<td></td>
</tr>
<tr>
<td>OMI-DL (2016) [344]</td>
<td></td>
</tr>
<tr>
<td>Anam et al. (2016) [345]</td>
<td></td>
</tr>
<tr>
<td>Xie et al. (2016) [346]</td>
<td></td>
</tr>
<tr>
<td>Mountasser et al. (2016)</td>
<td>[347]</td>
</tr>
<tr>
<td>Idoudi et al. (2016)</td>
<td></td>
</tr>
<tr>
<td>Xue et al. (2016) [349]</td>
<td></td>
</tr>
<tr>
<td>ALINSyn (2017) [349]</td>
<td></td>
</tr>
<tr>
<td>Liu et al. (2016) [350]</td>
<td></td>
</tr>
<tr>
<td>HSOMap (2016) [351]</td>
<td></td>
</tr>
<tr>
<td>FCA-Map (2016) [184]</td>
<td></td>
</tr>
<tr>
<td>KEPLER (2017) [127]</td>
<td></td>
</tr>
<tr>
<td>ONTMAT (2017) [352]</td>
<td></td>
</tr>
<tr>
<td>Xue et al. (2017) [353–355]</td>
<td></td>
</tr>
<tr>
<td>He et al. (2017) [356]</td>
<td></td>
</tr>
<tr>
<td>OIM-SIM\textsuperscript{I} (2017) [357]</td>
<td></td>
</tr>
<tr>
<td>SANOM (2018) [358]</td>
<td></td>
</tr>
<tr>
<td>EVOCROS (2018) [108]</td>
<td></td>
</tr>
<tr>
<td>FCA-MapX (2018) [204]</td>
<td></td>
</tr>
<tr>
<td>Ochieng &amp; Kyanda (2018)</td>
<td>[142]</td>
</tr>
<tr>
<td>Roussille et al. (2018)</td>
<td>[359]</td>
</tr>
</tbody>
</table>

Table 9
Google Translation API is not free, but as of 2021, a translation of 500,000 characters per month are free of charge.  

**BioPortal** The National Center for Biomedical Ontology (NCBO) developed and maintains BioPortal\[28\] [91, 92], a Web repository of interlinked biomedical ontologies. The portal grants access to biomedical ontologies and terminologies developed in various Semantic Web formats. Via REST services, users can query (among other things) for ontologies, their metadata, and also for individual ontology terms. Registered users can also submit ontology mappings. This allows for community-created integration content. Particularly interesting in the area of ontology matching are the mapping services provided: Mappings can be easily obtained for a term or for a given ontology. The BioPortal services and data can be used free of charge.

**DOID** The Human Disease Ontology (DO, very often also abbreviated with DOID) [141] contains, as of 2021, more than 10,800 human diseases which are described through an ontology; its identifiers start with the prefix DOID. The resource is built by a community of experts. The disease ontology contains mappings to other vocabularies such as MeSH (see below), ICD\[29\], or SNOMED-CT\[30\] concepts. It is publicly available\[31\] under a very permissive license (CC0).

**Google Search API** The Google Search API\[32\] allows to perform Web searches programmatically. Like the Google Translation API, it is not free, but as of 2021, 100 search queries per day are free of charge.

**Medical Subject Headings (MeSH)** The Medical Subject Headings (MeSH) [173] form the controlled vocabulary thesaurus which is used to index medical articles. It is built by experts and maintained by the US National Library of Medicine (NLM). The data is freely available online for download in multiple formats (including RDF).\[33\] The dataset is available under a permissive license.

**BabelNet** BabelNet\[34\] [104] is a large multilingual knowledge graph that integrates (originally) Wikipedia and WordNet. Later, additional resources such as Wiktionary were added. The integration between the resources is performed in an automated manner. The dataset does not just contain lemma-based knowledge but also instance data (named entities) such as the singer and songwriter Trent Reznor. For BabelNet 3.6, an RDF version exists [362]. The dataset can be queried via a UI, SPARQL, and an HTTP API (a Java and a Python client are also available). The dataset is under a restrictive license and the number of free queries is limited. However, researchers can request access to the indices for non-commercial research projects.

5. Categorization of Background Knowledge in Ontology Matching

5.1. Classification System

Multiple approaches for categorizing general matching techniques have been proposed [8–10]. The matching techniques further studied in this survey can be broadly categorized as context-based approaches according to Euzenat and Shvaiko [8, 9] or as schema-only based approaches according to Rahm and Bernstein [10].\[35\] Rahm et al. do not group background knowledge sources while Euzenat et al. distinguish formal resources, i.e. those on which reasoning can be applied, and informal resources, i.e. those on which reasoning cannot be applied. The latter authors further name the dimensions breadth, formality, and status [363]. In this survey, we propose a more fine-grained categorization with a clear distinction between the background knowledge source that is used and the strategy that is applied to exploit the given knowledge source.

**Target Domain** Background knowledge sources for matching can be grouped by their target domain or target purpose. Here, it can be differentiated between domain-specific assets and general-purpose assets. While general-purpose background knowledge is intended to improve the overall matching quality on any

---

27[see https://cloud.google.com/translate/pricing
28[see https://bioportal.bioontology.org/
29ICD stands for “International Classification of Diseases”.
30SNOMED-CT stands for “Systematized Nomenclature of Medicine Clinical Terms”.
31[see https://disease-ontology.org/
32[see https://developers.google.com/custom-search/v1/overview
33[see https://www.nlm.nih.gov/databases/download/mesh.html
34[see https://babelnet.org/
35This is naturally not precise. WordNet and other lexical resources, for example, are not classified as formal/informal resource-based but instead as language-based according to Euzenat and Shvaiko.
task, domain-specific background knowledge is intended to improve the matching performance within a specific domain or even for a specific matching task. An example for a widely used general-purpose knowledge source is WordNet; a point in case for a popular domain-specific knowledge source is the Unified Medical Language System (UMLS). The distinction between domain-specific and domain-independent (lexical and grammatical) sources is also made by Real et al. [174] who show in a recent publication that the inclusion of domain specific lexical- and grammatical knowledge can significantly improve matching systems in domain-specific tasks. In Figure 8, the aggregated usage of background knowledge in schema matching systems is plotted per year. It is visible that – up to date – general-purpose knowledge sources are used more often than domain-specific knowledge sources. This finding is intuitive, since general-purpose datasets are easier to find and their application makes sense for any matcher whereas domain-specific datasets may be harder to find (depending on the matching task) and require a concrete, domain-bound matching problem. It is also visible that the research community initially started with general-purpose background knowledge and explored domain-specific sources at a later stage. Most publications using external background knowledge sources (general and domain-specific) were published in 2018. It is important to note that this survey does not cover the full year of 2021.

Structuredness Independent of the domain, the knowledge sources can be split in structured sources and unstructured sources. Structured data is organized according to a known data schema whereas unstructured data is not. An example for a structured external data source in ontology matching is WordNet; an example for a general-purpose unstructured data source in ontology matching is the entirety of Wikipedia texts whereas SAP Term, a set of definitions of terms in SAP software, is an example of a domain-specific unstructured resource. Unstructured external resources are rarely used in ontology matching. We, therefore, only classify into textual and non-textual unstructured resources whereby we did observe merely one publication [170] using non-textual, unstructured sources (i.e., images).

Structured sources appear in different variations (type): (i) Lexical and taxonomic resources, (ii) factual databases, (iii) Semantic Web datasets, and (iv) pre-trained neural models. Lexical and taxonomic resources as well as pre-trained neural models can again be subdivided into monolingual and multilingual resources. Semantic Web datasets can be subdivided into single datasets and interlinked datasets.

An overview of the proposed classification system is presented in Figure 9; in Table 10, all resources covered in this survey are categorized according to the presented classification system. In the following, we will

36 Theoretically, the other structured resources can also be monolingual – however, the focus of the knowledge provided there is rather factual and the language is typically not the core property of the knowledge resource. Therefore, we decided against a subdivision here in favor of clarity.
Background Knowledge

Figure 9. Classification of background knowledge sources that are used for matching.

further define each structured resource and provide examples for all fine-grained categories.

Lexical and Taxonomical Knowledge Lexical and taxonomical knowledge is the most exploited external type of knowledge in ontology matching. The most commonly used resource in this class in our study is WordNet. The resource is monolingual, this means it is available in only one language, i.e. English. Similar resources exist in other languages such as the German thesaurus GermaNet [364] – however, since most ontology matching benchmark datasets are provided in English, our study is consequently also skewed towards English resources. Concerning multilingual lexical knowledge, dictionaries and dictionary-like resources, such as APIs, are heavily used for multilingual ontology matching. In our study, we found substantial usage of the Microsoft Bing Translation API but also of other general-purpose translation APIs. Although not appearing in the tables, domain-specific multilingual resources exist, for example the Fachwörterbuch Versicherungswirtschaft und -recht37 [365].

Factual Databases A factual database provides (non-lexical) facts that can be included into the matching process. An example here might be a database of postal codes and cities. We did not find any significant usage of such a resource despite imaginable use case scenarios. An example for a domain-specific database would be MEDLINE, the bibliographic database of the National Library of Medicine which is used by the DisMatch [175] and OntoEmma [176] matching systems.

Semantic Web Dataset A Semantic Web (SW) dataset is a knowledge base developed with technologies from the Semantic Web technology stack, such as RDF or OWL files. The category includes knowledge graphs with or without instance data where we define a knowledge graph slightly broader than in its original sense [366] and also count domain-specific graphs. We also consider SPARQL endpoints as SW datasets in this survey as well as plain ontologies.

We further differentiate between (i) single and (ii) linked SW datasets. A single dataset is in this case an individual knowledge graph or ontology.

An example for a general-purpose single SW dataset would be DBpedia [137] (used e.g. by LDOA [139]), WebsALOD [205, 206] (used e.g. by ALOD2Vec Matcher [207]), or Wikidata. An example for a domain-specific single SW dataset would be the Financial Industry Business Ontology (FIBO) used for instance in [148].

An example for domain-specific linked SW dataset in this sense would be some or all BioPortal [92] ontologies together with their mappings while an example for general-purpose linked SW dataset would be any two linked general-purpose knowledge graphs.

Pre-trained Neural Models A recent development is the application of deep learning in a multitude of applications. A pre-trained neural model in this classification system may be an API exposing latent representations of concepts, such as KGvec2go38 [171], or a pre-trained model such as the Google Universal Sentence Encoder39 [163, 164] used by VeeAlign [65].

5.2. Further Relevant Properties

Further properties of background knowledge sources that are not used here for the proposed classification are (i) resource size, (ii) task dependence, (iii) license permissions, and (iv) authoring level. Those properties are important in particular when it comes to the strategies that are applied to exploit the background knowledge.

37German book title, translates to dictionary of insurance and insurance law.
38see http://www.kgvec2go.org/
39see https://tfhub.dev/google/universal-sentence-encoder-large/
The resource size may limit the utility provided by the source – a small general knowledge thesaurus, for example, may only be of limited use – but may at the same time also limit the exploitation strategy that can be used: the RDF2Vec [367] embedding approach (a comparatively scalable embedding approach) is very hard to apply to the BabelNet (RDF) knowledge graph [362] due to its sheer size. Surprisingly, the most used general-purpose background knowledge source, WordNet, is relatively small compared to community-built resources such as BabelNet, Wiktionary, or Wikidata.

The task-dependency also limits the options to exploit the source (see Section 7). A very specific Web-API providing only a very specific service may limit the strategy to the simple call of the service.

While license permissions are not of utmost concern to the research community, they are very important in the enterprise world when it comes to the actual application of matching systems in the real world for commercial purposes.

The level of authoring or trust of a knowledge source is affecting the exploitation strategy as well. Generally, four main categories can be observed: (1) expert-built resources such as WordNet, (2) community-built resources such as Wiktionary, (3) semi-automatically built resources such as BabelNet, and (4) automatically built resources such as WebIsALOD. It can be assumed that the amount of trust decreases from (1) to (4): A deeply reviewed, expert built dictionary such as WordNet may be used with less caution than a community built online dictionary like Wiktionary or a heuristically extracted dataset such as WebIsALOD. The quality of the matching results is likely not in every case proportional to the level of trust since it depends on the exploitation strategy used and the concrete resource. Automatically-trained neural language models, for instance, have a low authoring level but may produce very good results.

6. Categorization of Linking Approaches

In order to exploit an external knowledge source, the concepts in one or both of the ontologies to be matched need to be linked to the knowledge source. The linking process is also known as anchoring or contextualization [363]. For example, to determine whether the classes http://mouse.owl#MA_0002390 and http://human.owl#NCI_C33743 of the OAEI Anatomy track [61] are similar using Wiktionary, the URIs have to be first linked to one or more Wiktionary entries. In this case, the label of the first can be used to link it to the entry of “temporalis” and the label of the latter can be used to link it to the entry of “temporal muscle”. Within the knowledge source, we can then find a synonymy relation between the two entries and derive a degree of similarity.

While many publications address the concrete application of a background source for ontology matching, few discuss the actual linking problem. However, since linking is the first step in exploiting a knowledge source, it significantly determines the quality of the outcome. In a visionary paper by Sabou et al. [188], online ontologies obtained with a Semantic Web search engine have been used for ontology matching. Out of the 1,000 correspondences checked manually, 217 false ones have been identified. The authors find that out of those, 53% are due to anchoring errors. This emphasizes the need for a solid anchoring strategy.

The linking process is typically dependent on the knowledge source used and can be as simple as forwarding a label (e.g. when using the Google search API) or as complicated as the ontology matching problem itself (e.g. when another knowledge graph shall be used).

For linking, we distinguish two goals: (i) finding at most one link for each concept in an ontology and (ii) finding up to many links for each concept in an ontology. Multiple links can be sensible in the case of partial linking; for example, a concept with label “derivatives exchange” may be linked to “derivatives” and “exchange” in cases where there is no match for the complete concept. Other reasons for multi-linking are datasets with homonyms or knowledge sources that explicitly provide multiple senses for strings. For the latter two cases, a Word Sense Disambiguation (WSD) approach may help to decide on a smaller set of links.

In terms of classifying linking approaches, we propose a classification system consisting of four categories: (i) given links, (ii) direct label linking, (iii) fuzzy linking, (iv) Word Sense Disambiguation (WSD). The proposed classification system is summarized in Figure 10. In the following, we will introduce Homonyms are words that have the same writing (homographs) or the same pronunciation (homophones) but different senses [368]. An example would be the word “bank” in two different contexts: It may refer to the financial institution in one case and to a seating accommodation in the other case. To be precise, for the linking problem at hand only homographs are challenging.
<table>
<thead>
<tr>
<th>Background Knowledge Type</th>
<th>Background Knowledge Source</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structured</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Domain-specific</strong></td>
<td></td>
</tr>
<tr>
<td>Lexical and Taxonomical</td>
<td>RadLex</td>
</tr>
<tr>
<td></td>
<td>SPECIALIST Lexicon</td>
</tr>
<tr>
<td>Factual Database</td>
<td>Medline</td>
</tr>
<tr>
<td></td>
<td>PubMed</td>
</tr>
<tr>
<td>Semantic Web Dataset</td>
<td>DOID</td>
</tr>
<tr>
<td></td>
<td>FMA</td>
</tr>
<tr>
<td></td>
<td>FIBO</td>
</tr>
<tr>
<td></td>
<td>Medical Subject Headings (MeSH)</td>
</tr>
<tr>
<td></td>
<td>UBERON</td>
</tr>
<tr>
<td>Pre-trained Neural Model</td>
<td>BioPortal</td>
</tr>
<tr>
<td></td>
<td>Ontology Lookup Service (OLS)</td>
</tr>
<tr>
<td></td>
<td>UMLS</td>
</tr>
<tr>
<td>Textual</td>
<td>Cooking Dictionary</td>
</tr>
<tr>
<td></td>
<td>SAP Term</td>
</tr>
<tr>
<td><strong>Unstructured</strong></td>
<td></td>
</tr>
<tr>
<td>Lexical and Taxonomical</td>
<td>Big Huge Thesaurus</td>
</tr>
<tr>
<td></td>
<td>Paraphrase DB (PPDB)</td>
</tr>
<tr>
<td></td>
<td>synonyms-fr.com</td>
</tr>
<tr>
<td></td>
<td>Universal Knowledge Core (UKC)</td>
</tr>
<tr>
<td></td>
<td>Wikipedia MediaWiki API (non-text search)</td>
</tr>
<tr>
<td></td>
<td>Wikisynonyms</td>
</tr>
<tr>
<td></td>
<td>WordNet</td>
</tr>
<tr>
<td></td>
<td>WordsAPI</td>
</tr>
<tr>
<td>Semantic Web Dataset</td>
<td>Apertium</td>
</tr>
<tr>
<td></td>
<td>Bing/Microsoft Translator</td>
</tr>
<tr>
<td></td>
<td>Freelantr</td>
</tr>
<tr>
<td></td>
<td>Google Translation API</td>
</tr>
<tr>
<td></td>
<td>HowNet</td>
</tr>
<tr>
<td></td>
<td>iTranslate4</td>
</tr>
<tr>
<td></td>
<td>Lanes API</td>
</tr>
<tr>
<td></td>
<td>MyMemory API</td>
</tr>
<tr>
<td></td>
<td>SDL FreeTranslation</td>
</tr>
<tr>
<td></td>
<td>Webtranslator API</td>
</tr>
<tr>
<td></td>
<td>Yandex Translation API</td>
</tr>
<tr>
<td>Factual Database</td>
<td>BabelNet</td>
</tr>
<tr>
<td></td>
<td>DBNary</td>
</tr>
<tr>
<td></td>
<td>DBpedia</td>
</tr>
<tr>
<td></td>
<td>ConceptNet</td>
</tr>
<tr>
<td></td>
<td>DOLCE</td>
</tr>
<tr>
<td></td>
<td>OpenCyc</td>
</tr>
<tr>
<td></td>
<td>SUMO</td>
</tr>
<tr>
<td></td>
<td>Swoogle</td>
</tr>
<tr>
<td></td>
<td>WebIsALOD</td>
</tr>
<tr>
<td></td>
<td>YAGO</td>
</tr>
<tr>
<td><strong>General-Purpose</strong></td>
<td></td>
</tr>
<tr>
<td>Structured</td>
<td>BERT</td>
</tr>
<tr>
<td>Pre-trained Neural Model</td>
<td>fastText model</td>
</tr>
<tr>
<td></td>
<td>Google Word2Vec Vectors</td>
</tr>
<tr>
<td></td>
<td>KGvec2go</td>
</tr>
<tr>
<td></td>
<td>SBERT</td>
</tr>
<tr>
<td>Textual</td>
<td>Google Universal Sentence Encoder</td>
</tr>
<tr>
<td></td>
<td>Bing Search Engine API</td>
</tr>
<tr>
<td></td>
<td>FARO Web Search</td>
</tr>
<tr>
<td></td>
<td>Google Search API</td>
</tr>
<tr>
<td></td>
<td>Wikipedia Corpus</td>
</tr>
<tr>
<td></td>
<td>Wikipedia MediaWiki API (for text search)</td>
</tr>
<tr>
<td></td>
<td>Yahoo Search</td>
</tr>
<tr>
<td>Non-Textual</td>
<td>ImageNet</td>
</tr>
<tr>
<td></td>
<td>Yahoo Image Search</td>
</tr>
</tbody>
</table>

Table 10

Background knowledge sources sorted according to their type.
Linking Strategies

![Diagram of Linking Strategies]

Figure 10. Categorization of Linking Approaches

each category in detail and provide examples. It is important to note that not every linking strategy can be applied on each dataset; WSD, for instance, can only be applied if there are multiple senses available in the background dataset.

**Given Links** In few cases, linking can be omitted if the external knowledge source already contains links, e.g., in the form of \texttt{owl:sameAs} or \texttt{owl:equivalentClass} statements. A case in point is Wikidata where multiple identifiers are typically specified; the concept \texttt{pneumonia} (Q12192)\footnote{\url{https://wiki.datso.com/wiki/Q12192}} for instance, lists more than 30 identifiers for other datasets – among them IDs for MeSH, BabelNet, the Disease Ontology, Freebase, or UMLS.

**Direct Label Linking** Given the sparse information provided in publications concerning the linking strategy, it can be assumed that in most cases linking is performed by directly looking up a potentially normalized label. This works particularly well if the external dataset has a very large coverage of concepts or even provides synonyms such as lexical and large taxonomical background knowledge datasets. Recent matching systems that apply this kind of linking are for example FCA-MapX\footnote{\url{https://www.ontomat.org}}, \texttt{ONTMAT1}\footnote{\url{https://wwwulg.github.io/ontomat-impl/}}, or \texttt{Wiktionary Matcher}\footnote{\url{https://github.com/squid-lab/wiktionary-matcher}}.

**Fuzzy Linking** The linking process can also be based on only parts of a label, n-grams within a label, or expanded labels. Such linking approaches fall under the fuzzy linking category. The underlying goal of this strategy is to find more links than through direct label linking. Naturally, this strategy is attractive if the background dataset is small and/or the concepts in it are described by a single label (without stating alternative names, abbreviations, synonyms etc.). Mascarelli et al.\footnote{\url{https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3313888}}, for instance, match two ontologies to an upper ontology and then use the obtained two alignments to derive a final alignment; they perform an involved

\begin{itemize}
\item **Word Sense Disambiguation (WSD)** We did not find matching systems that try to actually disambiguate the sense of a label through Word Sense Disambiguation (i.e. which try to settle with one correct sense) – despite the heavy usage of WordNet (which is built around senses).\footnote{Some authors consider WordNet metrics such as the Resnik word similarity \cite{Resnik1999} or WuPalmer \cite{WuPalmer1994} as WSD (e.g. \cite{107}) – however, we regard averaging synset similarity scores or picking the maximum score across multiple synset comparisons not as real Word Sense Disambiguation; the obtained similarity through such approaches is a word similarity rather than a disambiguated sense similarity.}
\item Instead, similarity approaches that can handle multiple senses are typically used. The \texttt{NBILM}\footnote{\url{https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3313888}} matching system narrows down the number of WordNet synsets – but only to reduce the computational complexity.
\end{itemize}

7. Categorization of Background Knowledge Exploitation Approaches

In Section 5, the background knowledge resources used in ontology matching have been presented and categorized. The second main dimension of this survey is the exploitation strategy of the background resource. In many cases, there are multiple options to beneficially use an external knowledge source.

We classify exploitation strategies into four groups: (i) factual queries, (ii) structure-based approaches, (iii) statistical/neural approaches, and (iv) logic-based approaches. A factual query is the request for one or more data records contained in the background resource. Structure-based approaches exploit structural elements in the background knowledge source. Statistical or neural approaches apply statistics or deep learning on the background knowledge source or consume an existing pre-trained model. Lastly, logic based approaches employ reasoning with the externally provided resource. In the following, the categories are further described and extensive examples are provided. An overview of the proposed classification system is provided in Figure 11.

**Factual Queries** A factual query is the extraction of an existing record from the knowledge source. This type of exploitation strategy is the most common one and used since the early days of (semi-) au-
Background Knowledge Exploitation Strategies

- factual query
- structure-based
- statistical/neural
- logic-based

Figure 11. Overview of the types of background knowledge exploitation strategies.

There is in some cases no clear boundary between structure-based and statistical approaches since structure-based approaches typically apply statistics. We classify an approach as structure-based if the focus is the exploitation of the structure of the knowledge source.

Context-based matching. An example for retrieving factual information would be retrieving synonyms from WordNet (applied by many matching systems e.g. RiMom [371], AgreementMaker [79], or FCA-Map [204]) or from DBnary [214] (e.g. by Wiktionary Matcher [72, 216]).

Structure-based Approaches Structure-based methods require a structural dimension in the background resource such as a tree or graph structure. Elements to be compared are typically projected into the background source and the structure is used to derive a new fact between the projected elements such as equivalence or subsumption. Structure-based approaches are often applied on WordNet to determine similarity such as the path-based approaches by Wu and Palmer [370] or Jian and Conrath [372] (both used for example by the YAM++ matching system [270]) or the information-based approach proposed by Lin [373] (used for example by the RiMom [374] matching system).33 Many more WordNet-based approaches that fall into the structure-based category of this survey paper have been proposed and used in ontology matching; we direct the interested reader to the survey by Lin et al. [361]. Structure-based approaches have not only been used together with WordNet but have also been applied on other datasets such as overlap-based metrics based on WeIsALOD [375]. A structural approach on Wikipedia categories is applied by BLOOMS [138] where concepts are linked into the Wikipedia taxonomy and an overlap measure on taxonomy sub-trees is defined to determine similarity. Given a repository of ontologies together with correspondences, Annane et al. [132] apply a structure-based strategy, where they first form a so-called global mapping graph. Source and target ontology are linked into the latter and a path-based strategy is applied so that the correspondences with the highest confidence can be extracted.

Due to their nature, structure-based approaches are not (obviously) applicable to factual databases, or pre-trained neural models.

Statistical/Neural Approaches Statistical approaches apply a statistical process on the data derived from the external knowledge source. The WeSeE-Match system [96, 97], for instance, builds virtual documents from search engine results and derives a similarity estimate by applying a strategy that is based on the term frequency-inverse document frequency (TF-IDF) vectors of the documents.

Neural approaches employ artificial neural networks either directly on the background knowledge source or re-use existing pre-trained models. For example, the background knowledge source may be transformed into a vector space [207] or the background knowledge source is already a vector space that may be used directly to link the schemas to be matched [65] in a vector space. We also count neural APIs into this category; ALOD2Vec Matcher [172], for example, uses in its most recent version the API of KGvec2go [171] to obtain vectors for concepts. While this could be seen as a factual query, we still consider this strategy to be a neural one due to the nature of the approach. It is important to note that we focus only on strategies applied to the background knowledge – a matching system that uses neural networks to configure weights of various features (e.g. the 2011 version of CIDER [376]) does not fall in this category and neither does a matching system that applies a neural model to the ontologies that are to be matched such as DOME [377]; the reason for this decision is that the latter two system types do not actually use external background knowledge for their matching strategy. Systems that apply statistical approaches are not novel – however, systems that apply neural methods are relatively recent (the oldest ones of this survey being from 2018, e.g. [207]), not plentiful in numbers, and achieve mixed results. This is most likely due to the novelty of this exploitation strategy. Notable in this category is the VeeAlign [65] matching system which uses a sentence encoder as external knowledge and achieved the best results on the Conference [64] track in the OAEI 2020.

Logic-based Approaches Logic-based approaches apply reasoning on or together with the external resources. This class of approach is also referred to as context-based matching [11] or indirect match-
The term **indirect matching** may also refer to structure-based approaches such as the works by Annane et al. [86, 132]. This is due to the fact that in this survey, we differentiate in structure-based approaches (such as a path-based algorithm) and logic-based approaches—a distinction that other authors do not make.

The steps are namely: (i) ontology arrangement, (ii) contextualization, (iii) ontology selection, (iv) local inference, (v) global inference, (vi) composition, and (vii) aggregation.

**SUMO** stands for “suggested upper merge ontology”, **DOLCE** stands for “descriptive ontology for linguistic and cognitive engineering”, and OpenCyc is a subset of the Cyc knowledge base by Cycorp that is not available anymore.

Typical external resources are upper ontologies, domain-ontologies, knowledge graphs, or linked data. We differentiate reasoning from the factual queries in that a reasoning operation goes beyond querying a graph with an **ASK** query for equivalence or any other relation between two concepts. Logic-based approaches are already envisioned in the earlier days of ontology matching. An archetypal setup of such an approach is presented in Figure 12 which was first presented by Sabou et al. [187] and slightly adapted for this survey: Elements of the ontologies to be matched are linked to the external ontology (Sabou et al. call this step **anchoring**, Euzenat et al. refer to this step as **contextualization**, see Section 6) and reasoning is applied to derive correspondences. It is important to note that reasoning can also be applied across multiple ontologies: Locoro et al. [11] generalize and significantly extend the approach by Sabou et al.; they perform reasoning also across more than one intermediate ontology. Their proposed generalized framework consisting of seven logical steps is particularly applicable for logic-based approaches.

However, we did not find broad usage of logic-based exploitation approaches in past and current (OAEI and non-OAEI) ontology matching systems that go beyond singled out experiments. Approaches that fall into this category are Sabou et al. who use Swoogle to retrieve ontologies from the Web. **BLOOMS+** [212] does not strictly reason on the external resource but applies a context similarity measure which is based on overlap of superclasses which could be seen as such. Mascardi et al. [144] perform experiments on multiple upper ontologies (**DOLCE** [143], **SUMO** [186], **OpenCyc** [178]) following a similar approach of exploiting the transitivity of equivalence relations. Strictly speaking, Mascardi et al. are also not performing a real reasoning operation as defined in the beginning of this paragraph. Despite the clear vision of the latter two publications, upper ontology approaches that exploit actual reasoning have not gained traction so far.

### 8. Directions for Future Work

In Section 5, we proposed a classification system for background knowledge sources and in Section 7 we presented a classification system for exploitation approaches. In this section, we will overlap those to a matrix and will position the systems evaluated in this survey in there. We will use this matrix as a starting point for discussions of white-spots in the area of background knowledge-based ontology matching. We further outline interesting observations, shortfalls and biases found in the ontology matching domain.

#### 8.1. White Spots

Tables 11 (domain knowledge) and 12 (general knowledge) present the systems evaluated in this study in a source/strategy matrix. The exploitation strategy (columns) in the table follows the proposed classification which is summarized in Figure 11. The rows represent the background knowledge type and follow the proposed classification which is summarized in Figure 9. Irrelevant combinations of source and strategy are grayed out in the tables. Empty or rarely filled white cells hint at yet underexplored and potentially interesting research directions in the area of background knowledge-based ontology matching.
<table>
<thead>
<tr>
<th>Background Knowledge (Domain-Specific)</th>
<th>Strategy</th>
<th>Factual Queries</th>
<th>Structure-based</th>
<th>Logic-based</th>
<th>Statistical/Neural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monolingual</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical and Taxonomical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factual Database</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic Web Dataset</td>
<td>Single</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linked</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domain-Specific</td>
<td>Pre-trained Neural Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstructured</td>
<td>Textual</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Non-Textual</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Systems in the background knowledge type / exploitation method type matrix (domain-specific background knowledge).
### Strategies

<table>
<thead>
<tr>
<th>Factual Queries</th>
<th>Structure-based</th>
<th>Logic-based</th>
<th>Statistical/Neural</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOCOM + (2012)</td>
<td>[114]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S awe WordNet system, see Tables 8 and 9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mao et al. (2011)</td>
<td>[137]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hashimoto et al. (2021)</td>
<td>[217]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logic-based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Statistical/Neural</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Systems

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Monolingual</th>
<th>Multilingual</th>
<th>General-Purpose</th>
<th>Factual Queries</th>
<th>Structure-based</th>
<th>Logic-based</th>
<th>Statistical/Neural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unstructured</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Note

Table 12: Systems in the background knowledge-type / exploitation method type matrix (general-purpose background knowledge).
From the tables we see that general purpose background knowledge is used more often than domain-specific background knowledge.\(^{47}\) The most often used background knowledge type are lexical and taxonomical resources with WordNet being the clear winner. Clearly not often used are unstructured, non-textual data, pre-trained neural models, and general-purpose Semantic Web datasets.\(^{48}\) It is important to note that the heavy usage of linked data in Table 11 is mainly due to UMLS falling in that category – almost all systems listed use this single resource. Hence, the general application of linked data is not yet common, too. Interestingly, the application of general-purpose textual data has been explored in multiple publications whereas there is merely a single application of domain-specific free text.

It is quickly visible that factual queries are most often used regarding the strategy. When it comes to yet underexplored research directions of background knowledge usage, we see that in terms of the approaches used, logic-based and neural-based strategies are an interesting and promising research direction. Pre-trained embedding-models and architectures, for instance, are up to 2020 rarely used but may be very promising given breakthroughs in other scientific communities. An increase in publications in 2021 in this category may indicate that scientific interest is already moving in this direction. Structural approaches are almost completely limited to the English WordNet. The exploration of structural methods on multilingual datasets as well as on Semantic Web datasets may yield interesting results given good results on the English WordNet and given that this class of approaches is typically intuitive to understand and can be comprehended by humans (unlike neural models).

8.2. It’s a Biomedical World

If we take a closer look at the domain-specific knowledge sources used, it is striking that almost all datasets are from the biomedical domain. This may be due to a particularly prolific bioinformatics community that holds open standards and open data high – however, the skewness of ontology matching publications towards the biomedical domain must be pointed out. In Figure 6 (cumulative background knowledge usage), it is striking that all domain-specific datasets are from the biomedical domain. This domain-focus also visible when looking at OAEI tracks where almost all domain-specific problems are from this domain. This fact is likely self-enforcing: New researchers use existing evaluation datasets and existing background knowledge and quickly find themselves in this domain area.

Nonetheless, ontology matching is a problem in all domains that are concerned with data management which makes it ubiquitous. Enterprise schema matching and integration challenges in the business world, for example, are not reflected at all in OAEI tracks.\(^{49}\) In addition, there are indications that top-performing OAEI schema matching systems perform comparatively bad on real world business integration tasks [384]. More insights on the generalization of current matching methods, properties of matching problems in other domains, or further well-performing domain-specific or general-purpose datasets are desirable.

An interesting research direction is, therefore, also to broaden the domain-focus of the ontology matching problem and to evaluate which background datasets and exploitation strategies are applicable in other domains. Therefore, new and publicly available benchmark datasets from more domains are required to support research efforts in this area. New challenges may come to light such as missing domain-specific knowledge sources not being broadly available [385]. The provisioning of further evaluation datasets in other domains is a clear desideratum.

8.3. Multilinguality

A further bias besides a domain-focus is the focus on monolingual ontology matching. At the OAEI, there is currently only one multilingual matching task with few participants. The techniques currently applied are purely lookup-based despite advances in machine translation.

Multilingual ontology matching requires the addition of external resources; hence, we can find many

\(^{47}\)Note that systems that use WordNet (see Tables 8 and 9) are not explicitly listed for better clarity in Table 12.

\(^{48}\)The low usage of factual databases may be due to the fact that the community prefers knowledge presented in a graph.

\(^{49}\)In the years 2016 and 2017, there was a Process Model Matching Track at the OAEI. While the topic of process model matching is relevant for the industry, the dataset was limited to the domains of university admissions in 2016 and additionally birth registrations in 2017. At the OAEI, the overall participation in the track was rather low with only four systems in two years: AML [378, 379], DKP [380], LogMap [381, 382], and I-Match [383].
multilingual background sources in Tables 4 to 7. However, when we compare the resource/strategy matrix in Tables 11 and 12, we quickly see that there are many systems that use general-purpose multilingual resources but there is not a single system that uses domain-specific multilingual resources. This may be due to the fact that there are at the moment no benchmark datasets for more advanced multilingual matching tasks available – despite this being a relevant problem in the real world. The current multilingual evaluation datasets are all from the conference domain with a rather low level of domain-complexity.

It could be further observed that, although many diverse multilingual resources such as Wikidata or EuroVoc50 exist, most multi-lingual matchers use translation APIs with a simple factual query strategy. This setup limits reproducibility and transparency.

Interesting research directions are the exploration of new multilingual matching methods and datasets as well as the exploration of multilingual matching challenges in domain-specific settings. The provisioning of further evaluation datasets is also for the aspect of multilinguality a desideratum. Given well-performing and publicly available deep-learning models from the NLP domain, their application should also be considered for the ontology matching task.

8.4. The English Bias

Another language-based bias is the focus on aligning schemas that are semantically described in the English language. The research community currently mainly solves English-English alignment problems.51 This bias can already be seen when reviewing the most common evaluation datasets – but this bias is also found in the background knowledge used: The majority of background knowledge sources listed in Tables 4 to 7 are available in English as main language (with the exception of some translation-oriented datasets such as translation APIs). It is unlikely that this setting reflects the real-world situation.

An interesting research direction is, therefore, the exploration of non-English rooted ontology matching problems with non-English background knowledge sources. As with the multilingual background knowledge sources. As with the multilingual background knowledge, the community would greatly benefit from the provisioning of more evaluation datasets.

8.5. Manual Background Knowledge Selection

While multiple automatic background knowledge selection approaches have been proposed (see Subsection 4.2), we did not find significant usage of documented automated selection processes in the publications reviewed for this survey. Up to the date, the majority of background knowledge sources in ontology matching is either bound to one predefined source or uses few hand-picked resources. With the exception of LogMapBio, most matching systems which apply an automated selection approach are presented in the context of background knowledge selection. Hence, self-configuring matching systems that select their own background resources based on a particular matching problem are still an interesting area of research. Very recent approaches, such as the usage of pre-trained language models that are fine-tuned on the matching task, do not solve this task (but instead emphasize the importance since the pre-trained model also needs to be selected).

8.6. Linking

Our analysis on how concepts are linked into the background knowledge source revealed that most matching systems do not perform elaborated linking approaches but use a direct string lookup. While this may be sufficient for some background datasets, there is indication that in some cases linking is a significant component in the performance of background knowledge-based matching systems [188, 190].

A reason for the negligence when it comes to linking might be that Word Sense Disambiguation is perceived as too hard. Another reason might be due to the fact that schemas to be integrated are often derived from the same domain which significantly reduces the amount of concept and definiens and concept mismatches [386] induced by homonyms since words will often refer to the same senses. For example, when two ontologies from the financial services domain use the term “bank”, they likely both refer to the sense of a financial institution – an elaborated WSD approach would not provide any value here. Existing evaluation datasets are all more or less from the same domain and do not reflect this problem appropriately.

51It has to be mentioned here that this survey only considers publications published in English (see C1 in Table 2) which may skew the observations. However, given that English is the lingua franca in the ontology matching community, we assume that this skew is small.
However, when large external knowledge bases are to be matched or when the schemas to be matched are large and diverse such as in the case of knowledge graph matching, WSD may significantly improve the results obtained with external background knowledge. This finding is in line with a recent publication on knowledge graph matching by Hertling and Paulheim [60] who show that state-of-the-art matching systems perform badly when it comes to matching non-related or weakly-related knowledge graphs due to non-disambiguated homonyms.

An interesting research direction is consequently the development, evaluation, and comparison of multiple linking approaches and their effect on the performance of automated matching systems. We also see a need for the provisioning of additional matching gold standards in the area of knowledge graph matching as well as matching of weakly related schemas.

9. Conclusion

Since the early 2000’s, the understanding of the (automated) ontology matching problem as well as the development of advanced matching systems have greatly improved. Nonetheless, the ontology matching problem is not solved and will stay an interesting research area for the years to come. One key to coming closer to the solution is the deeper integration of background knowledge within the ontology matching process.

In this survey, we reviewed all ontology matching systems that participated in the OAEI from 2004 until today, as well as systematically selected ontology matching systems in terms of what background knowledge sources they use, which linking approach they employ, and how they use the external knowledge. We classify background knowledge in multiple structured and unstructured classes according to their purpose (domain-specific or general-purpose). The main structured knowledge source types are (i) lexical and taxonomical resources, (ii) factual databases, (iii) Semantic Web datasets, and (iv) pre-trained neural models. The main unstructured resource types are (i) textual and (ii) non-textual. In our review we found that mostly general-purpose structured knowledge is used in ontology matching. Most systems to date make use of simple lexical and taxonomical sources. Yet under-explored sources of background knowledge are unstructured resources, pre-trained neural models, general purpose knowledge graphs, and linked data.

We further presented a classification system for linking strategies consisting of four categories: (i) given links, (ii) direct linking, (iii) fuzzy linking, and (iv) Word Sense Disambiguation. Although linking is important when it comes to exploiting external knowledge sources, we found that most systems use direct label linking.

Concerning the strategy that is used to exploit knowledge sources, we presented a classification system consisting of four categories: (i) factual queries, (ii) structure-based approaches, (iii) logic-based approaches, and (iv) statistical/neural approaches. We found that a look-up strategy of facts is most commonly used. Structure-based strategies are almost exclusively applied on WordNet. Despite a clear vision, logic-based approaches did not gain much traction in recent years. A novel research area in terms of exploitation strategies are neural approaches which are currently barely used but showed very good results in other domains.

In our survey, we found multiple biases when it comes to ontology matching with background knowledge: (i) A focus on biomedical matching tasks, (ii) a focus on monolingual matching, and (iii) a focus on matching schemas rooted in the English language. In particular the business world where integration problems are plentiful and multi-faceted, is hardly considered in current research efforts. Although the focus of this survey is the usage of external knowledge within the ontology matching process, we consider the identified biases to be generally applicable.

Acknowledgement

The publication of this article was funded by the University of Mannheim.

References

J. Portisch et al. / Background Knowledge in Ontology Matching: A Survey


[23] P. Shvaiko, J. Euzenat, E. Jiménez-Ruiz, M. Cheatham and O. Hassanzadeh (eds), Proceedings of the 12th International Workshop on Ontology Matching co-located with the 16th


J. Portisch et al. / Background Knowledge in Ontology Matching: A Survey

Background Knowledge in Ontology Matching: A Survey


[267] Z. Wang, Y. Wang, S. Zhang, G. Shen and T. Du, Effective Large Scale Ontology Mapping, in: Knowledge Science, Engineering and Management, First International Conference, KSEM 2006, Guilin, China, August 5-8, 2006, Pro-


S. Gherbi and M.T. Khadir, ONTMAT1: Results for OAEI


A. Khiat and M. Benaissa, InsMT / InsMTL results for OAEI 2014 instance matching, in: Proceedings of the 9th International Workshop on Ontology Matching collocated with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Trentino, Italy, October 20,


computing, V ol. 536, 2016, pp. 157–165. doi:10.1007/978-3-319-
2016
[351] J. Euzenat and P. Shvaiko, RDF2Vec: RDF graph embeddings and their applica-
tions, in: Proceedings of the Ninth
International Conference on Language Resources and Evalu-
ation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014,
N. Calzolari, K. Choukri, T. Declerck, H. Lofsson, B. Mae-
gaard, J. Mariani, A. Moreno, J. Odijk and S. Peveridis, eds,
European Language Resources Association (ELRA), 2014,
lrec2014/summaries/810.html.
[352] M. Ehrmann, F. Cecconi, D. Vannella, J.P. McCrae, P. Cimi-
ano and R. Navigli, Representing Multilingual Data as Linked
Data: the Case of BabelNet 2.0, in: Proceedings of the Ninth
International Conference on Language Resources and Eval-
uation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014,
N. Calzolari, K. Choukri, T. Declerck, H. Lofsson, B. Mae-
gaard, J. Mariani, A. Moreno, J. Odijk and S. Peveridis, eds,
European Language Resources Association (ELRA), 2014,
lrec2014/summaries/810.html.
[353] J. Euzenat and P. Shvaiko, Ontology Matching, 2nd edn,
978-3-662-38720-3.
[354] B. Hamp and H. Feldweg, Germanet-a lexical-semantic net for
german, in: Automatic information extraction and building
of lexical semantic resources for NLP applications, 1997.
[355] K. Purvis, Fachwörterbuch Versicherungswirtschaft und
Recht: Englisch / Deutsch - Deutsch / Englisch, C.H.Beck &
Vahlen, Heidelberg, Germany, 2021. ISBN 978-3-406-72353-
7.
[356] H. Paulheim, Knowledge graph refinement: A survey of
approaches and evaluation methods, Semantic Web 8(3) (2017),
heim, RDF2Vec: RDF graph embeddings and their applica-
tions, Semantic Web 10(4) (2019), 721–752. doi:10.3233/SW-
180317.


