Emerald Journal of Documentation

A Systematic Review of Methods for Aligning, Mapping, Merging Taxonomies in Information Sciences

Journal:	*Journal of Documentation*
Manuscript ID	JD-01-2023-0003.R1
Manuscript Type:	Article
Keywords:	Taxonomy, Taxonomy alignment, Knowledge Organization Systems, Taxonomy mapping, Taxonomy merging, Ontology
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
main-revised.tex	

A Systematic Review of Methods for Aligning, Mapping, Merging Taxonomies in Information Sciences

Yi-Yun Cheng [1], Yilin Xia [2]

[1] Department of Library and Information Science, School of Communication and Information, Rutgers, the State University of New Jersey; [2] School of Information Sciences, University of Illinois at Urbana-Champaign

Author Note

Correspondence concerning this article should be addressed to Yi-Yun Cheng, School of Communication and Information, Rutgers, the State University of New Jersey, 4 Huntington St., New Brunswick, NJ 08901. E-mail: yiyun.cheng@rutgers.edu

Abstract

Purpose: The purpose of this study is to provide a systematic literature review on taxonomy alignment methods in Information Science to explore the common research pipeline and characteristics.

Design/methodology/approach: We implement a five-step systematic literature review process relating to taxonomy alignment. We take on a knowledge organization systems (KOS) perspective, and specifically examining the level of KOS on “taxonomies”.

Findings: We synthesize the matching dimensions of 28 taxonomy alignment studies in terms of the taxonomy input, approach, and output. In the input dimension, we develop three characteristics: tree shapes, variable names, and symmetry; for approach: methodology, unit of matching, comparison type, and relation type; for output: the number of merged solutions, and whether original taxonomies are preserved in the solutions.

Originality: There is no existing comprehensive review on the alignment of “taxonomies”. Further, no other mapping survey research has discussed the comparison from a KOS perspective. Using a KOS lens is critical in understanding the broader picture of what other similar systems of organizations are, and enable us to define taxonomies more precisely.

Research implications: The main research implications of this study are three-fold: (1) to enhance the understanding of the characteristics of a taxonomy alignment work; (2) to provide a novel categorization of taxonomy alignment approaches into natural Language processing approach, logic-based approach, and heuristic-based approach; (3) to provide a methodological guideline on the must-include characteristics for future taxonomy alignment research.

Keywords: Taxonomy, taxonomy alignment, Knowledge Organization Systems

A Systematic Review of Methods for Aligning, Mapping, Merging Taxonomies in Information Sciences

Introduction

A *taxonomy* is a hierarchical, parent-child structure used to organize information. Taxonomies, being an easy and intuitive way to classify vocabularies, are prevalently used in quotidian categorization tasks and e-commerce website's product organization (Hedden, 2016; Hlava, 2014). *Taxonomy alignment* is the process of comparing two or more taxonomies to reach agreements on how to merge the taxonomies. The need for reconciliation of taxonomies often arises when multiple taxonomies about the same topic disagree with each other; or when one single taxonomy disagree with itself over time (Franz, Chen, et al., 2016; Franz et al., 2015).

In fact, alignment is not a task exclusive to taxonomies. Interoperability problems between and among knowledge organization systems (KOS) have long been studied in Information Science and extra-information science fields (Baca, 2016; Chan & Zeng, 2006; Otero-Cerdeira et al., 2015; Shvaiko & Euzenat, 2005; Zeng & Chan, 2006). Generic models such as derivation and expansion have been discussed to enhance metadata interoperability (Chan & Zeng, 2006; Zeng & Chan, 2006). Specific metadata interoperability solutions such as metadata crosswalks have also been explored to map one metadata standard to another – examples manifested in the Getty Institute's crosswalk of the CDWA standard to 14 other standards (Baca, 2016), or one of the many crosswalks from MARC to other standards by the Library of Congress (Library of Congress, 2008). Further, mapping methods for a particular type of KOS like the ontology matching methods have been thoroughly reviewed with regards to the different automating mechanisms for matching (Otero-Cerdeira et al., 2015; Shvaiko & Euzenat, 2011). Similar to ontology mapping research, there are also extensive reviews about database schema matching process to analyze how one schema can be transferrable to another (Rahm & Bernstein, 2001; Shvaiko & Euzenat, 2005).

However, while there are many studies working with taxonomy alignment, a comprehensive review surveying the mapping, merging, alignment of *taxonomies* seems to have been missing from the conversations. Many reasons may have caused a gap in taxonomy alignment reviews, some conjectures include: One, there are many diverse application areas that taxonomies are used in, and each taxonomy alignment method can be drastically different and uniquely developed for an application. Second, taxonomies may be embedded in other information systems or databases, and some may have equated taxonomy alignment with the overall schema mapping tasks. Third, taxonomy alignment studies can have many different characteristics, and this makes it difficult to summarize the commonalities among these studies.

Realizing that diverse taxonomy alignment methods have been proposed but there are no existing surveys, the goal of this paper is to provide a systematic review of taxonomy alignment research in Information Science and related fields to explore the commonalities and differences across these studies. As such, we ask the following research questions:

- RQ1: What are the characteristics of a taxonomy alignment process?
- RQ1: How are the characteristics in taxonomy alignment research alike or differ?

Specifically, we examine 28 research relating to taxonomy mapping, matching, or alignment methods, garnered through a five-step systematic literature process (Erfani & Abedin, 2018; Kelly & Sugimoto, 2013; Siqueira & Martins, 2022). We take on an angle from the knowledge organization systems (KOS) framework (Hodge, 2000; Zeng, 2008), and specifically examining the level of KOS on "taxonomies".

Background

Knowledge Organization Systems (KOS)

Knowledge organization systems (KOS) are all kinds of knowledge structures used to organize information and concepts (Hodge, 2000; Mazzocchi, 2018; Zeng, 2008). Hodge

(2000) categorized types of knowledge organization systems into term lists, classifications and categories, and relationship lists; Zeng (2008) added “metadata-like models” and further denoted the different dimensions of each KOS. “Taxonomies”, in Zeng (2008)’s model, were categorized in the “classification and categorization” group as classification schemes and subject headings that arrange groups of items in hierarchical relationships. In Soergel (2009), taxonomies are not only hierarchical representation, but a kind of KOS that focus primarily on concepts rather than terms.

As Mazzocchi (2018) mentioned, many authors have only a “partial agreement” on how to classify KOSs, and the confusion arises from a lack of precise definition for each of the KOSs. The author further suggested that different communities may have been using the same KOS, but calling it by another name. Very much like Soergel (1999) pointed out how “ontologies”, popularized by the computer science domain, could be “reinventing” the ideas of classification schemes, and many may have equated ontologies with taxonomies. However, it is clear by many authors’ categorizations of KOSs that taxonomies and ontologies have many differences structurally and functionally (Hodge, 2000; Mazzocchi, 2018; Zeng, 2008).

The key difference between a taxonomy and an ontology is the way in which *relationships* are represented. Zeng (2008)’s KOS framework stated that taxonomies establish only *hierarchical relationships*, whereas ontologies establish both *hierarchical* and *associative relationships* while representing other *properties*. Mazzocchi (2018) reviewed many authors’ KOS framework and demonstrated that taxonomies and ontologies are usually at different KOS levels where the former are loosely defined hierarchies, and the latter are formal vocabularies that provide many other properties.

However, the reason why the use of the term “taxonomies” are sometimes interchangeable with “ontologies”, is how both KOSs are used to classify natural languages into controlled vocabularies; and the techniques to classify concepts in a taxonomy are similar to the automatic indexing techniques used in ontologies (Gilchrist, 2003; Pieterse & Kourie, 2014). Nevertheless, the two kinds of KOSs have clear distinctions, taxonomies as a hierarchy, or tree model; and ontologies as a linked semantic network with its concepts having complex relations and properties (Hodge, 2000; Mazzocchi, 2018; Pieterse & Kourie, 2014; Smiraglia, 2014; Soergel, 2009; Zeng, 2008).

Myriads of Alignment Methods

The desire to make taxonomies, or knowledge organization systems, interoperable with each other, is a common practice in organizing information (Chan & Zeng, 2006; de Andrade & de Lara, 2016; Zeng, 2019; Zeng & Chan, 2006). To date, there are myriads of mapping, matching, alignment methods in the wild. The process of aligning (mapping, matching) has been studied in many contexts to solve interoperability issues in information systems. In various fields in the Computer Science community, there is a significant amount of work discussing database *schema* matching, *schema* merging, large *schema* mapping (Rahm & Bernstein, 2001; Shvaiko & Euzenat, 2005). In discussing the Semantic Web technologies, extent research on *ontology* mapping, *ontology* matching, *ontology* alignment (Euzenat, Shvaiko, et al., 2007; Shvaiko & Euzenat, 2011), or the more recent *knowledge graph alignment* (Wang et al., 2018) have been emerging since Berners-Lee, Hendler, and Lassila's disruptive concept of the semantic web (Berners-Lee et al., 2001). In Information Sciences, various interoperability approaches have been discussed extensively since the beginning of bibliographic classification systems and metadata standards (Chan & Zeng, 2006; Zeng & Chan, 2006). For instance, utilizing metadata crosswalks are one of the major representations to map one metadata standard to another, such as Getty Institute's crosswalk of the CDWA standard to 14 other standards[1] (Baca, 2016), or one of the many crosswalks from MARC to other standards by the Library of Congress (Library of Congress, 2008). Spanning across different fields, *taxonomy alignment* (mapping,

[1] Getty crosswalk: http://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html

matching, merging) work flourishes in biodiversity (Franz, Chen, et al., 2016), biology (Kanehisa et al., 2022), e-commerce (Aanen et al., 2012), information systems and computer science (Raunich & Rahm, 2014), and many others.

Much like the insider joke in the metadata community about the proliferation of metadata standards: nobody wants to use other people's toothbrushes (existing standards), so they just use their own toothbrush (invent a new standard) – the same phenomenon goes to alignment methods. Different fields have different ways to express mapping, matching, alignment work, and even within the same community, new matchers or methods are constantly developed.

Taxonomies and Taxonomy Alignment

The usage of taxonomies is ubiquitous (Cheng, 2022). Many have attempted to define what a taxonomy is, but the definition of the term often varies based on the contexts that taxonomies are used in. Gilchrist (2003), as cited by Mazzocchi (2018), gave five different contexts that use taxonomies: (1) web directories; (2)taxonomies to support automatic indexing; (3) taxonomies created by automatic categorization; (4) front end filters; (5)corporate taxonomies. Briefly stated, a taxonomy is a hierarchical representation used to organize vocabularies.

Even though there are many definitions and contexts of what constitutes a taxonomy or what a taxonomy does, the common practices of creating and using taxonomies suggest that there is at least a shared understanding of what taxonomies are. In almost every field, taxonomies are used to categorize, group, or classify objects into a tree model, or a hierarchical structure (Hedden, 2016; Pieterse & Kourie, 2014; Van Hooland & Verborgh, 2014).

From a Knowledge Organization perspective, the levels of "what is being aligned" is an important distinction to make. Many who attempted alignment projects could be operating on the KOS level of "ontologies", but they used the term "taxonomies" to be more generic. Given the aforementioned differences between taxonomies and ontologies, it is also necessary to understand that the process of "ontology mapping, ontology matching, ontology alignment" is the union of not only the concepts in the ontologies, but also all the hierarchical, associative relationships, as well as the object and data properties between two or more ontologies. Sometimes, this also means that the axioms and rules associated with the ontologies are expected to be harmonized into one set of axioms and rules (Dou et al., 2005). On the other hand, if the alignment process is on the level of taxonomies, the merge would be a simpler union of concepts from the taxonomies, and the taxonomies' is-a relationships.

Recognizing there are many approaches attempt to align KOSs, we believe it is crucial to make clear distinctions on the type of KOS we are examining. This study is reviewing alignment methods on the KOS level of "taxonomies", we will be excluding discussions of alignment methods on other levels of the KOS (e.g., relational *schema* matching, *ontology* matching, etc.), and "taxonomy" alignment methods that view taxonomies synonymous as ontologies.

In later sections of this paper, we will review different definitions of taxonomies based on the systematic literature review results. Our own definition of a taxonomy in this paper are as follows: a taxonomy T is defined as a pair: $T = (\mathbf{C}, \mathbf{E})$, where $\mathbf{C}$ is a set of concepts, and $\mathbf{E}$ is a set of edges. In this study, we will follow our definition of taxonomy as inclusion and exclusion criteria of the review corpus. From hereafter, unless when describing the steps for keyword searching, we will only be using the term "taxonomy alignment" to refer to all acts of mapping, and merging work with taxonomies.

Methods

Given that the usage of taxonomies spans across different application areas and disciplines (e.g., computer science, e-commerce, biology), we leverage the techniques used in systematic literature review (Erfani & Abedin, 2018; Kelly & Sugimoto, 2013; Siqueira &

Martins, 2022) to retrieve all relevant articles relating to taxonomy alignment methods. According to Kelly and Sugimoto (2013), systematic literature reviews help researchers to be “exhaustive with their coverage of the literature” by identifying sources of articles, developing search strategies, and devising inclusion and exclusion criteria of the literature. By retrieving all necessary and relevant articles, researchers can attempt to “create generalizations” for analyzing the articles (Kelly & Sugimoto, 2013).

One of the advantages to leverage systematic literature review is that it allows transparency and reproducibility of the search process. We follow a similar five-step systematic literature review process from Kelly and Sugimoto (2013) and aims to document our steps as clearly as possible. These steps include:

- Step 1. Identifying sources for finding key papers
- Step 2. Develop inclusion and exclusion criteria for identifying articles
- Step 3. Search strategies and search terms
- Step 4. Studies selection
- Step 5. Final assessment and analysis of the articles

Step 1: Identifying Key Papers

Common sources for finding articles include Web of Science, Scopus, and domain specific databases, such as how Erfani and Abedin (2018), and Diqueira and Martins (2022) implemented in their studies. We perform searches in four databases (as of March 22, 2022): Scopus and Web of Science for all domains; EBSCO’s Library and Information Science Source, and Library, Information Science Technology Abstracts (LISTA) for domain-specific databases in the field of Library and Information Science.

Step 2: Developing Inclusion and Exclusion Criteria

We develop inclusion and exclusion criteria for identifying articles that are evidently dedicated to taxonomy alignment research.

The inclusion criteria are: (1) peer-reviewed journal papers or conference proceedings; (2) written in English; (3) papers that includes taxonomy alignment as a research process.

The exclusion criteria are: (1) studies that are surveying or reviewing many alignment methods; (2) studies that are not describing a specific alignment method or mapping problem; (3) studies that define taxonomy in a different KOS-level (e.g., ontologies).

Step 3: Search Strategies and Search Terms

Our search terms are "taxonomy alignment", "taxonomy mapping", "taxonomy matching", or "taxonomy merging" anywhere any of these terms appear in an article's title, keywords, abstract, or full-text. We enumerate the search terms in each database and the number of results below:

In Scopus, the following search terms return 71 results:

```
TITLE-ABS-KEY ( "taxonomy alignment"  OR  "taxonomy mapping"
OR  "taxonomy matching"  OR  "taxonomy merging" )
AND  ( LIMIT-TO ( LANGUAGE ,  "English" ) )
AND  ( LIMIT-TO ( DOCTYPE ,  "ar" )
OR  LIMIT-TO ( DOCTYPE ,  "cp" ) )
```

In Web of Science, the following search terms return 44 results:

```
ALL=("taxonomy alignment" OR "taxonomy mapping"
OR "taxonomy matching" OR "taxonomy merging" )
```

In EBSCO- Library and Information Science Source; and in Library, Information Science & Technology Abstracts (LISTA), the following search terms return 4 results, and 3 results, respectively:

```
"taxonomy alignment" OR "taxonomy mapping"
OR "taxonomy matching" OR "taxonomy merging"
```

Step 4: Studies Selection

To further select relevant studies from the 122 results we retrieved from the databases, we first deduplicate the 122 results based on articles' DOI, title, and author, and that left us with 75 articles.

An initial round of screening through these 75 articles' titles and abstracts are conducted to further eliminate any other survey or review papers (all about either ontology or schema surveys), as per our exclusion criteria – we eliminate 20 papers that are not pertinent to taxonomy alignment processes.

A second round of screening articles from reviewing the full text of the remaining 55 articles is conducted to exclude articles that:

(1) the main scope of the research is not the alignment process (e.g., discussing natural group structure of protein; discussing forensic tools for blockchain technologies);

(2) the level of KOS discussed: if the researchers' definition of taxonomies is actually referring to other KOSs, it will be excluded. If the term taxonomies and ontologies are used interchangeably, we examine the researchers' definition to determine if the KOS they are working with have only one node type and one edge type (is-a). For example, we include the study from Raunich and Rahm (2014) because their taxonomies only consist of one node and edge type. (Their definition of taxonomies is "we will consider only ontologies $O = (C; C_i; I; R)$ where C_i contains only leaf nodes and R contains only is-a relationships between concepts. For this reason, in the following, we will use the terms ontology and taxonomy with the same meaning.")

(3) the paper contains an evaluation of their alignment process by comparing to other "ontology" matchers. For instance, we exclude Aanen et al. (2012)'s paper: though

the authors used the term “taxonomy mapping” throughout their paper, they compared their matcher SCHEMA with other *ontology* matchers by Noy and Musen (2003) and Park and Kim (2007).

(4) the paper conducts an evaluation of their alignment process by using benchmark “ontology” alignment datasets. For example, we exclude the study by Lin et al. (2019), because they defined taxonomy as “multi-relational data with numerous triple facts” and used the benchmark dataset from the 2017 Ontology Alignment Evaluation Initiative (OAEI)[2] for evaluating their method.

This process results in 28 papers for final assessment and analysis (Step 5). Step 5 of the systematic literature review involves final assessment and analysis, which we will discuss in detail in the subsequent Results section. The metadata of the 28 studies selected is provided in Appendix 1 of the Supplementary Materials. The indices of the studies in Appendix 1 mirror the numbers for the articles in Table I, II, and III. Duplicated numbers indicate that the paper is a follow-up study, or group research from the same group of authors.

Results

Based on our inclusion and exclusion criteria, 28 studies are identified for analysis. In this section, we first describe the various definitions of a taxonomy, then we provide a systematic organization of the characteristics of taxonomy alignment methods.

Definitions of a Taxonomy

In these 28 papers, we found at least four different variations of the definition of a taxonomy, quoted below.

Angermann et al., (2017) stated that:

[2] OAEI:`http://oaei.ontologymatching.org/2017/`

> “Formally, a Taxonomy T is an out-tree. $T = (C, E)$. which is using a set of concepts C for describing terms with the help of a label, and a set of edges E connecting less general with more general concepts.”

Chen et al. (2020) defined taxonomy as:

> “A taxonomy T_k represents a tree-based schema that consists of a set of concepts having certain relations and sharing a set of resources. It is denoted by a triple $T_k = (C, R, D)$ where: C... represents a set of concepts of the taxonomy; R...represents a set of semantic relations between concepts... In this article, only the relation ‘is-a’ is considered; D... represents a corpus of document.”

In Giabelli et al. (2022), the authors mentioned that:

> “A taxonomy T is a 4-tuple $T = (C, W, H, F)$. C is a set of concepts (aka, nodes)...; W is a set of words (or entities, or leaf concepts) belonging to the domain of interest...; H is a directed taxonomic binary relation between concepts...; and F is a directed binary relation mapping words into concepts...”

In Reynaud and Safar (2006), the authors denoted:

> “A taxonomy is a pair (C, H_C) consisting of a set of concepts C arranged in a subsumption hierarchy H_C. A concept is only defined by two elements: a label and subclass relations. The label is a string which can be an expression composed of several words. Subclass relations establish links with other concepts...It is the single semantic association used in the hierarchy.”

Despite the many variations to define a taxonomy, there are many commonalities across all papers: taxonomy as a tree, taxonomy as a hierarchy, a taxonomy contains a set of concepts, and a taxonomy contains a subset relation. Here we reiterate our definition of a taxonomy in this study: A taxonomy T is defined as a pair: $T = (\mathbf{C}, \mathbf{E})$, where $\mathbf{C}$ is a set of concepts, and $\mathbf{E}$ is a set of edges.

Possible interpretations of concepts can be taxa (e.g., genera), spatial regions (e.g., the United States), classes (e.g., persons), or other types (Lehmann, 1992; Sowa, 1992). Possible interpretations of edges can be subset relation, proper-part-of relation, or other semantic relations noted in the literature (Brachman, 1983; Chaffin et al., 1988; Gerstl & Pribbenow, 1995; Woods, 1975). While concepts and edges can have different interpretations, it is important that the interpretation of concepts and edges in a given taxonomy should be homogeneous to avoid semantic anomalies.

Characteristics of Taxonomy Alignment Methods

In this section, we categorize the characteristics associated with taxonomy alignment methods based on the 28 studies we examined. To date, there is no existing survey or review papers solely on "taxonomy alignment" approaches, so a more comprehensive investigation into these approaches has yet to be provided. Here we adopt the term "matching dimensions" from foundational schema mapping surveys ((Rahm & Bernstein, 2001; Shvaiko & Euzenat, 2005) to address the common pipeline (input, approach, output) of a mapping research. We also develop new characteristics pertaining to taxonomy alignment work, based on the commonalities we observed across the 28 studies.

Tables I, II, and III present an overview of the characteristics and the annotation results on the studies we examined. Within the 28 studies, some studies are follow-up studies, or group research from the same group of authors. We identified those studies and group them together. This eventually left us with 20 different methods in the three results tables. We use Cohen's Kappa to measure the inter-agreement (McHugh, 2012) between the two authors' annotation task on the categorical characteristics of the 28 studies (20 methods). The two authors reach a moderate agreement on the first round of annotating four categorical characteristics (symmetry, methodology, unit of matching, comparison type) with Cohen's Kappa κ=0.49. Upon iterative discussions, the authors reach a substantial agreement on the second round of annotating the same four categorical characteristics, with Cohen's Kappa κ=0.79. Three new categorical characteristics (tree shapes, number of merged solutions, information preserved) are added during our second round of discussion. With all seven categorical characteristics incorporated, the two authors still reach substantial agreement with Cohen's Kappa κ=0.68. The codebook for the annotation task is attached in the Appendix 2 of the Supplementary Materials.

Table I

Characteristics of Taxonomy Alignment Methods -1

		ATOM [1]	**Chen** [2]	**CROEA** [3]	**Daraio** [4]	**Euler/X** [5]	**HAMSTER** [6]	**Jung** [7]
Application Areas		automobile, anatomy, e-commerce	e-commerce	education	e-commerce	geography, biodiversity	e-commerce	digital library, virtual organizations
Matching Dimensions								
Input	**Tree shapes**	vertical tree, horizontal tree	texts, vertical tree	texts, vertical tree	vertical tree	vertical tree, horizontal tree	general hierarchy	vertical tree
	Variable names	Os, Ot	Tt, Ts1, Ts2,. . . , Tsm	lom = (w1m, w2m. . .)	D= {d1,. . . dn}	T1, T2..Tn	Ts, Tt	Ti, Tj
	Symmetry	asymmetric	asymmetric	asymmetric	asymmetric	symmetric	asymmetric	symmetric
Approach	**Methodology**	logic-based	logic-based, heuristic-based, NLP-unsupervised	NLP-supervised	NLP-supervised	logic-based	NLP-unsupervised	NLP-unsupervised
	Unit of Matching	element-level, structure-level	element-level, structure-level	element-level	element-level	element-level, structure-level	element-level, structure-level	element-level
	Comparison Type	Pairwise	Multiple	Multiple	pairwise	multiple	pairwise	pairwise
	Relation Type	equivalence, is-a, inverse is-a	equivalence, is-a, inverse is-a	text similarity	text similarity	equivalent, disjoint, overlap, proper-part, proper-part inverse	text similarity	text similarity
Output	**Number of Merged Solution(s)**	N=1	N=1	N=1	N=1	N=1, N>1	N=1	N=1
	Information Preserved	Somewhat	Somewhat	No	No	Yes	No	No

[1] ATOM in Raunich and Rahm, 2014; [2]Chen et al., 2020; [3] CROEA in Mouriño-Garcıa et al., 2018; [4] Daraio et al., 2020; [5] Euler/X in Cheng et al., 2017; Cheng et al., 2020; Cheng and Ludäscher, 2020; Franz, Chen, et al., 2016; Franz et al., 2015; Johnston, 2016; Thau and Ludäscher, 2007; [6] HAMSTER in Nandi and Bernstein, 2009; [7] Jung, 2006, 2008

Application Areas

Application areas are the fields, domains, or areas of interests that the taxonomy alignment studies applied to. We infer the application areas of a study by either if the authors directly mentioned the areas they were inspecting in the paper, or by the example taxonomies they provided in the study.

Table II

Characteristics of Taxonomy Alignment Methods -2

		Kanehisa [8]	**Lee [9]**	**Li'11 [10]**	**Li'08 [11]**	**Maldonado [12]**	**Merge-Into [13]**	**Musgrove [14]**
Application Areas		biology	cross-lingual	finance	business	business	computer science, general ontologies	e-commerce
Matching Dimensions								
Input	**Tree shapes**	vertical tree	texts	horizontal tree	vertical tree	texts	vertical tree	texts
	Variable names	T1, T2	Chinese WordNet, WordNet Domains	Tszse, Tcas	Tsic, Ttarget	Tvendor, TGPC	Tsrc, Tdest	TDMOZ, TYahoo, TAbout, Tt
	Symmetry	symmetric	asymmetric	asymmetric	asymmetric	asymmetric	asymmetric	asymmetric
Approach	**Methodology**	heuristic-based	NLP-supervised	heuristic-based	NLP-unsupervised	NLP-supervised	heuristic-based	NLP-unsupervised
	Unit of Matching	element-level	element-level	structure-level	element-level	element-level	element-level, structure-level	element-level
	Comparison Type	pairwise	pairwise	pairwise	pairwise	pairwise	multiple	pairwise
	Relation Type	author-defined relations	text similarity	author-defined relations	text similarity	text similarity	author-defined relations	text similarity
Output	**Number of Merged Solution(s)**	N=1	N=1	N=1	N=1	N/A	N=1	N=1
	Information Preserved	Yes	No	No	No	No	Somewhat	No

[8]Kanehisa et al., 2022; [9]Lee et al., 2009; [10]J.-m. Li et al., 2011; [11]Z. Li et al., 2008; [12]Maldonado et al., 2021; [13] Merge-Into in Subramaniam et al., 2009; [14] Musgrove, 2006

In these 28 papers (20 methods) , at least nine of them were from the e-commerce or business domains – meaning they were mainly working with web directories or corporate taxonomies such as Amazon's product taxonomy (Angermann et al., 2017; Nandi & Bernstein, 2009); Walmart's taxonomy (Angermann et al., 2017; Chen et al., 2020); Italian company (Daraio et al., 2020); or other virtual organizations (Jung, 2006, 2008). Two papers were concerned with the domain area of cross-lingual mapping (Lee et al., 2009; Xu & Sun, 2007), mainly the mapping from leveraging English taxonomies to Chinese taxonomies. Other papers are from many different domains, from education (in CROEA) (Mouriño-Garcıa et al., 2018), to biology (Kanehisa et al., 2022), or to other general areas where the alignment process can be applicable to all domains (Cheng et al., 2017; Raunich & Rahm, 2014; Subramaniam et al., 2009).

Table III

Characteristics of Taxonomy Alignment Methods -3

		Pfeifer [15]	**Ponzetto** [16]	**Reynaud** [17]	**Taxo-Semantics** [18]	**WETA** [19]	**Xu** [20]
Application Areas		business	Wikipedia	general	e-commerce	business	cross-lingual
Matching Dimensions							
Input	**Tree shapes**	horizontal tree	vertical tree	vertical tree	texts, vertical tree	vertical tree	texts, vertical tree
	Variable names	Tsj, Ttk	T, T'	Ts, Tt	Ts,Tt, Background Knowledge	To, Td	TCE, TCODP
	Symmetry	asymmetric	asymmetric	asymmetric	asymmetric	asymmetric	asymmetric
Approach	**Methodology**	NLP-supervised	NLP-supervised	NLP-unsupervised	NLP-unsupervised	NLP-supervised	NLP-supervised
	Unit of Matching	element-level	element-level	element-level, structure-level	element-level, structure-level	element-level	element-level
	Comparison Type	pairwise	pairwise	pairwise	pairwise	pairwise	pairwise
	Relation Type	text similarity; equivalent, subtype, inverse subsumption, associative	text similarity	text similarity	text similarity	text similarity	text similarity
Output	**Number of Merged Solution(s)**	N=1	N=1	N=1	N=1	N=1	N=1
	Information Preserved	No	No	No	No	No	No

[15]Pfeifer and Peukert, 2013a, 2013b; [16]Ponzetto and Navigli, 2009; [17]Reynaud and Safar, 2006; [18] Taxo-Semantics in Angermann et al., 2017; [19] WETA in Giabelli et al., 2022; [20] Xu and Sun, 2007

Matching Dimensions

Shvaiko and Euzenat, 2005 first used the term “matching dimensions” to describe the overall pipeline all alignment methods would encounter. We adopt this terminology for taxonomy alignment, seeing that all alignment methods have a common pipeline that involves *Input*, *Approach*, and *Output*.

Input

In their book, Euzenat et al. (2007) elaborated that the ***Input*** is the schema, or in our case, taxonomies, that goes into comparison. In the following sections, we synthesize the characteristics tree shapes, variable names, and symmetry of the taxonomy alignment inputs.

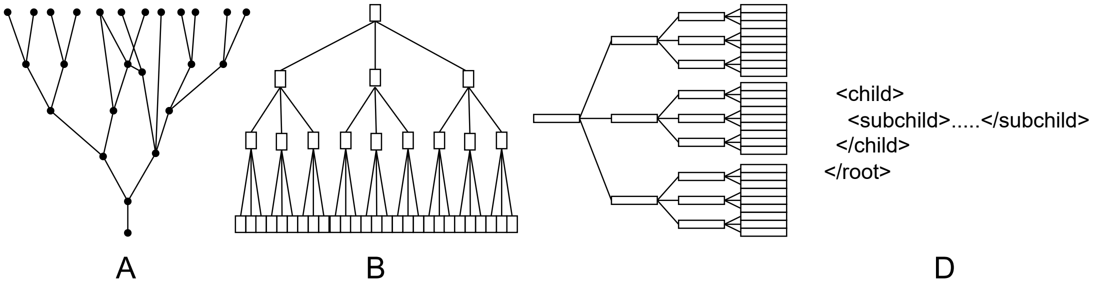


Figure 1

Tree Shapes of a taxonomy input: (A) figurative tree; (B) vertical tree; (C) horizontal tree; (D) general hierarchy. Terminologies for tree shapes (A), (B), (C) adopted from Zeng (2021).

Tree Shapes. In the 28 papers we examined, the inputs are all taxonomies, but the representations of these taxonomies can vary. Some could be in the form of texts, mentioned in the research narratives as concept or candidate pairs; some others can be a general hierarchical syntax in technical standards such as XML; yet some others are more illustrative, such as depicting an actual tree structure. We follow some of Zeng's organization of tree shapes, namely figurative trees, vertical trees, and horizontal trees (Zeng, 2021), along with our own observations on tree-structures (texts, general hierarchy) across the 20 approaches, and use the following terms for annotating the representations of trees in taxonomy alignment studies: texts, figurative trees, vertical trees, horizontal trees, general hierarchy (see Figure 1 and Appendix 2).

Out of the 20 methods, the most used tree shape was the vertical tree (14), followed by texts (7), horizontal tree (4), general hierarchy (1). Some studies used more than one tree shapes to depict their taxonomies, with combinations of either vertical trees with horizontal trees, or texts with vertical trees. None of the studies chose figurative trees to depict their taxonomies.

Variable Names. Variable names are the mathematical forms, axioms, or values used to describe the taxonomies compared. For example, when referring to a first taxonomy and a second taxonomy, one could use T_1 and T_2 as variable names for the two taxonomies. In these studies, subscript variables tend to vary, depending on the example taxonomy used (e.g. T_{szse}, T_{cas}). In general, many studies prefer to use numeric subscripts (e.g. T_1, $T_2 ... T_n$) or subscript variables that depicts the source and target taxonomies (e.g. T_s, T_t).

Symmetry. Raunich and Rahm (2014) defined the taxonomy alignment methods by symmetry. Asymmetric matching means there is a source taxonomy and a target taxonomy, and the alignment process are to fulfill the desired goals of the target taxonomy. In other words, in asymmetric mapping, structural inconsistencies from the source taxonomy may be disregarded to favor the consistency of the target taxonomy during the alignment process. Symmetric mapping means there are no apparent source or target taxonomies, two (or multiple) taxonomies are viewed as equals, and the alignment goal is usually to make all taxonomies logically and structurally consistent throughout the alignment process. We adopt this lingo of symmetric vs asymmetric because it is a helpful indicator to describe taxonomy alignment methods.

Most of the studies we examined are asymmetric – this means there are fewer constraints to satisfy all the taxonomies involved, and the alignment process may be able to scale up computationally, or more efficiently. On the other hand, it also means that there will be certain amount of information loss from the source taxonomy that are irreproducible. The works that are symmetric (Cheng et al., 2017; Jung, 2006; Kanehisa et al., 2022) means that they could be computationally more taxing to satisfy all taxonomies' constraints, but they can potentially provide more diverse viewpoints to help preserve more information from the original taxonomies.

Approach

The ***approach***, or process, is the methodology classification of these papers. The approach is often the core of a taxonomy alignment study, as it speaks to the matching process and the actual techniques in matching. It is hard to find two single papers that operates with the same approach, except for group papers (e.g., Euler/X team papers), or follow-up paper by the same authors (e.g., Pfeifer and Peukert, 2013a, 2013b). We synthesize the characteristics of the taxonomy alignment approach into: methodology, unit of matching, comparison type, and relation type.

Methodology.

Natural Language Processing Approaches

Natural Language Processing (NLP) approaches entail that concepts in two (or more) taxonomies are mapped based on the degree of similarity among the lexical features in the concepts. For instance, $T_1.cherry$ and $T_2.cherries$ may be mapped as equivalent because the concepts share the same word stem. In Rahm and Bernstein (2001), Shvaiko and Euzenat (2005), and Shvaiko and Euzenat (2011), the authors provided many categories for NLP related approaches: string-based, language-based, linguistic resources, name-based, text-oriented, etc. Here, we aggregated these as either NLP unsupervised or supervised methods.

Usually, unsupervised methods such as finding synonyms in synonym rings or existing word dictionaries (e.g., WordNet) are used while mapping two taxonomies (Angermann et al., 2017; Musgrove, 2006; Reynaud & Safar, 2006). Some studies in the unsupervied NLP methods category utilized continuous measures in specifying degree of similarity between, or among taxonomies to assign a confidence score (Z. Li et al., 2008; Pfeifer & Peukert, 2013a, 2013b); some others tried to determine word similarity on exact match or fuzzy match measures (Subramaniam et al., 2009; Xu & Sun, 2007).

Supervised NLP methods usually involve training datasets to determine word similarities in a vector space (Daraio et al., 2020; Giabelli et al., 2022; Maldonado et al., 2021;

Mouriño-Garcıa et al., 2018). It is also interesting to note that with the rise of supervised NLP neural network approaches, studies that employed supervised methods were more recent.

Usually, the goal for employing an NLP approach is to enrich the source taxonomy or an existing dictionary by automating the taxonomy alignment processes. The taxonomies examined usually comprise a large number of concepts, and by automatically aligning two (or more) taxonomies, the studies could come up more quickly with suggested candidate pairs that are ranked higher in similarity to be included in the updated taxonomy (Giabelli et al., 2022; Jung, 2006, 2008; Reynaud & Safar, 2006).

Logic-based Approaches

Logic-based similarity approaches align concepts in two (or more) taxonomies according to a set of formal, qualitative relations that could be used for logical reasoning. These relations are usually discrete measures rather than continuous measures; and they consist of relations other than equivalence between concepts.

In Raunich and Rahm (2014), equivalence, is-a, and inverse-is-a relations were introduced as the mapping mechanisms. Chen et al. (2020) followed Raunich and Rahm (2014)'s approach very closely, and also utilized the same three mapping mechanisms. We classified these two studies as logic-based approaches because these relations can be formally written into logic axioms. In studies relating to the use of Euler/X approach (Cheng et al., 2017; Cheng et al., 2020; Cheng & Ludäscher, 2020; Franz, Chen, et al., 2016; Franz et al., 2015; Johnston, 2016; Thau & Ludäscher, 2007), qualitative reasoning technique such as Region Connection Calculus was used to denote five base relations (RCC-5): equivalence, proper part, inverse proper part, disjointness, and overlapping. The number of relation types used in the mapping process is the major difference between Raunich and Rahm (2014)'s ATOM and the Euler/X approach. There are also other qualitative relations beyond the ATOM's three relations, or the RCC-5 relations, such as RCC-8, used mainly for geo-spatial reasoning purposes (Renz, 2002) rather than comparing semantic relations. There are even

16 relations proposed by Inants and Euzenat (2015), which are beyond the scope of taxonomy alignment studies.

Usually, the goals for logic-based approaches are to align concepts semantically as well as structurally in ways where the merged taxonomies could be logically consistent. To achieve this, in Raunich and Rahm (2014) and Chen et al., (2020), the authors focused more on the target taxonomy, and they remove concepts or nodes from the source taxonomies when they are not logically consistent. In the Euler/X approach, if structural inconsistencies were detected among the taxonomies, an iterative process to revise the mapping relations would be needed so both taxonomies are preserved with logically aligned concepts. Another difference between the other logic-based approach and the Euler/X approach is that the Euler/X approach allows for multiple merged taxonomies as output, if there are additional structurally possible solutions after merging. Further, Raunich and Rahm (2014) and Chen et al. (2020) aimed at a semi to fully automated logic-based approach towards taxonomy alignment, while the Euler/X studies were semi-automated, requiring human-in-the-loop process.

Heuristic-based Approaches

Only four studies were heuristic-based among the 28 papers we examined. We define heuristic-based approach as when the authors define their own discrete measures or rules that do not fit with the logic-based or the NLP approaches. For example, in Kanehisa et al. (2022), the authors worked with protein and genomics databases KEGG, and used their self-defined JOIN operation to conduct a match on key IDs in the taxonomies within the databases. In Li et al. (2011) and Subramaniam et al. (2009), the authors defined sets of rules that were written into the algorithm to transform the taxonomies and perform taxonomy alignment. In Chen et al., (2020), the authors implemented hybrid steps and combine logic-based, heuristic-based, and unsupervised NLP approaches. Based on our observation of these studies, usually, the goal of heuristic-based approaches is to provide a one-to-one match between concepts in both taxonomies, but the goals can also vary.

Unit of Matching. We adopted Rahm and Bernstein (2001)'s characteristics of Element vs. Structure matching, which element matching refers to the mapping between the concept nodes; and structure matching means that the levels where the nodes are situated at also need to be mapped.

All of the papers we examined were working on the level of element matching, that is, attempting to align the meanings of the concepts, classes, or nodes of the taxonomies; with the exception of Li et al. (2011), in which the authors explicitly stated that they are resolving "structural conflicts" between taxonomies. Some other studies, predominantly logic-based studies, were concerned with both element and structure matching. This means these studies considered both the structures of the taxonomies as well as the semantics of the concepts when aligning the taxonomies.

Comparison Type. Most alignment methods we examined were aligning taxonomies in a pair-wise manner. This means that during the alignment process, each time there were only two taxonomies compared. Chen et al. (2020), largely based on Raunich and Rahm (2014)'s ATOM, claimed their approach to be multi-ways, whereas ATOM could only work with pairwise alignment. Subramaniam et al. (2009) explicitly discussed the multi-way merge problem, and named four questions to be considered: "(1) Is there a preferred order in which they should be merged? (2) Does this preferred order always exist? (3) If it does, is it unique? (4) How can we determine the preferred order for multi-way merging?". They claimed their approach to be "order-free" but mentioned that all methods are bound to have information loss. Franz et al. (2016), with the Euler/X approach, attempted to align multiple taxonomies by aligning two at a time. The Euler/X approach could potentially align three, or more taxonomies at a time, but at present the studies were still mainly working with pairwise alignments (Cheng et al., 2017; Cheng & Ludäscher, 2020).

Relation Type. Relation type is used to explicate how two concepts in either taxonomies are inter-linked or compared. For instance, if $T_1.concept\ A$ is marked as "equivalent" to $T_2.concept\ B$, then the "equivalent" linkage here between the two concepts is the "relation type" of this alignment process.

In the studies we examined, the most sought-after relation type was text similarity, that is, using a continuous metric to compare two concepts in most NLP-approach studies. Discrete relation types such as equivalent, proper-part, and proper-part inverse were used across logic-based studies (Chen et al., 2020; Cheng et al., 2017; Raunich & Rahm, 2014)– though some referred to the proper-part relation as "is-a" and "inverse-isa". In heuristic-based approaches, the relation types were mostly author-defined, given that these studies were operating on their own algorithmic matching processes (Kanehisa et al., 2022; J.-m. Li et al., 2011). It is also worth noting that in Pfeifer and Peukert (2013a)'s NLP-based approach, on top of the text-mining service as a main mapping relation type, the authors included other mapping relation types: equivalent, subtype (similar to proper-part), and inverse subsumption (similar to proper-part inverse).

Output

The ***output*** of a taxonomy alignment method is the results of comparing, aligning, mapping two or more taxonomies. What and how the results look like can vary in different studies. Here we synthesize the output of taxonomy alignment in terms of the number of merged solutions produced, and whether the information of the original taxonomies is preserved.

Number of Merged Solution(s). Almost all studies, regardless of methodologies (logic-based, NLP-based, heuristic-based) or comparison type (pairwise vs. multiple), were capable of producing one merged solution (N=1). In one study (Maldonado et al., 2021), it was difficult to discern or infer from the narratives if the alignment successfully output any results.

In the logic-based Euler/X studies, the output may have more than one merged solutions (N>1) (Cheng et al., 2020). The Euler/X studies called the merged solutions as "possible worlds". The authors believed that given that there are multiple perspectives on how two taxonomies should be merged, there could be more than one merged solutions. This means that the ambiguity of the original taxonomies themselves, or the underspecified relations between the concepts in two taxonomies can create more than one outcomes. In general, for the studies that were based on NLP or heuristic approaches, the output was usually word pairs, a list of words, a list of categories, or numeric scores on how similar two concepts are that could help with enhancing or expanding the taxonomy. These word pairs or lists ultimately made up one merged solution.

Information Preserved. This output category speaks specifically towards the information of the original taxonomies. We are interested in knowing whether, after the taxonomy alignment process, one can still recognize the original taxonomies T_1 and T_2 (or T_n) from the merged solution(s). Knowing if the original taxonomies are preserved could potentially be valuable in situations when input taxonomies are lost or not given. When re-alignment is needed, being able to infer the original taxonomies from the merged solution can help reproduce the alignment outcome. Situations where we-aligning taxonomies can occur are when new version(s) of the same taxonomies appear, or when one tries to align the original T_1 and T_2 with other new taxonomies.

In the 28 studies, if the alignment input was “asymmetric”, the study was more concerned with the “target taxonomy”. Hence, information of the source taxonomy would usually not be preserved. There were still exceptions on asymmetric studies that preserve both the source and target taxonomies: this is manifested in Raunich and Rahm (2014), Chen et al. (2020), and Subramaniam et al. (2009), in which the authors stated that they preserved some or parts of the source taxonomies (hence marked as “Somewhat” in the result table). For studies that were “symmetric” mapping studies – there was no distinction between source or target taxonomies – the original taxonomies’ information was preserved in the merged solution(s) (Cheng et al., 2017; Kanehisa et al., 2022). That is, from the merged solution(s), we could see the original node or concept names, or even the structure of the nodes from either taxonomies. The exception of this is Jung (2008), in which only scores of text similarities were shown in their results.

Conclusion and Discussion

In this research, we review 28 studies that employ mapping, matching, alignment methods on the KOS level of "taxonomies".

Matching Dimensions and Characteristics

Our first research question asks: what are the characteristics of a taxonomy alignment process? To address this question, we adopt the term "matching dimensions" from foundational schema mapping works (Rahm & Bernstein, 2001; Shvaiko & Euzenat, 2005) to refer to a mapping research pipeline. This pipeline includes the "input", the "approach", and the "output" of a mapping research. Through our systematic review, we come up with characteristics correspond to each of the three dimensions (input, approach, output).

In the Input dimension, the characteristics we develop are tree shapes, variable names, and symmetry. While variable names and symmetry can be used as characteristics in other KOS-level mapping work, the characteristic tree shapes are unique to taxonomies only. Tree shapes include texts, figurative trees, vertical trees, horizontal trees, and general hierarchy. We are interested in tree shapes as a characteristic because we recognize how the taxonomies are visualized and represented are crucial towards the research's objectives or their application domains. For instance, for e-commerce taxonomy alignment research, horizontal trees are more likely to be used to simulate product taxonomy on the user interface (e.g. Amazon website taxonomy).

The characteristics in the Approach dimension are methodology, unit of matching, comparison type, and relation type. These four characteristics are derived from the aforementioned schema mapping work and our synthesis of the 28 taxonomy alignment research. While these four characteristics may not be exclusive to taxonomy alignment studies and can be applied to other KOS-level mapping work, the categorical values we suggested for these characteristics are reflecting the taxonomy alignment research we reviewed. For instance, for Methodology, we categorize it as NLP-based approach, logic-based approach, and heuristic-based approach because these approaches are predominantly used in the 28 studies. Other three characteristics (unit of matching, comparison type, relation type) may be universal to alignment work at other KOS-level, but they are important aspects that should be included in taxonomy alignment studies as well.

The Output dimension should show the characteristics of the outcome(s) of an alignment work – we thus include two characteristics: Number of Merged Solution(s) and Information Preserved. Cognizant that some studies might eventually merged two taxonomies into various possible ways, we are interested to know how many merged solution(s) each taxonomy alignment method can offer. As for the Information Preserved characteristic, it is to address concerns about whether an alignment approach is capable for reproducing the taxonomy alignment process via its output solution.

Commonalities and Differences

In our second research question, we ask how similar or different the characteristics in taxonomy alignment research are. In terms of application areas, taxonomy alignment research are applied in many different application areas, with the most common in e-commerce, followed by business domains.

In the three matching dimensions and the characteristics corresponding to each dimension, if categorical values are presented, it means we observe commonalities of that characteristic across the 28 papers, so groupings of categories are possible. For example, the characteristic *Tree Shapes* has categorical values because there are 14 studies that represent their taxonomies using the vertical tree, seven studies used texts, four studies used horizontal trees, and one used a general hierarchy in XML.

Characteristics that we suggest as free inputs, or provide examples but having free input value options, are the characteristics that have more variety and differences across the 28 studies. For instance, for variable names, each study may manifest the axioms or mathematical forms of its taxonomies quite differently. For relation type, each study may also have its own defined rules for matching, or it may contain similar mechanisms as other studies (e.g., text similarity). "is-a" and "inverse is-a" might be interchangeable with "proper-part" and "proper-part inverse", but we retain each research's own way of naming these relation types. A full list of the characteristics that shows which ones are categorical (hence more studies have commonalities in that characteristics), which ones have free input values, can be found in our codebook, attached in the Appendix 2 of the Supplementary Materials.

While the characteristics we propose are mutually exclusive from each other, they may have interactions between or among each other. For example, we observe interaction between methodology and relation type: if a study is using an NLP-based approach, it is more likely to use the relation type "text similarity". Similarly, if it is a heuristic-based approach, it is more likely to use author-defined relation and rules. Interaction between symmetry and information preserved is also manifested: if a study is asymmetric, meaning a source taxonomy and a target taxonomy are provided, it is less likely to preserve the information of its taxonomies in the output merged solution. Conversely, if a study's taxonomy inputs are symmetric, it is more likely to still contain its original taxonomies in the output merged solution.

Study Limitations

Evidently, the limitation of this study lies within the limitations of systematic literature review method itself. The systematic review of this study is conducted through a set of predetermined search terms, exclusion criteria, and inclusion criteria. It is possible that there are papers about taxonomy alignment methods that we are missing because of these predetermined selection steps.

A second limitation of this study could be the reliability of coding the characteristics of the studies. The two authors reached substantial agreement (Cohen's Kappa κ= 0.68) on the categorical characteristics. However, given the complexity of each taxonomy alignment research work, the true essence of a taxonomy alignment method may not be fully conveyed and implemented in a research article, leading to a more interpretive coding analysis on our end.

We are cognizant that the analysis of this study may not reflect the entire taxonomy alignment literature and that it is not feasible to draw statistical conclusions such as correlations between and among the characteristics. However, we believe the systematic review and the 28 research studies selected can serve as a first core corpus to help improve the understanding of taxonomy alignment research.

Contributions

In this study, we take on a Knowledge Organization Systems perspective and differentiate taxonomy alignment methods from the mapping research of other KOS(s). To our knowledge, no other mapping survey research has discussed what they are comparing from a KOS point of view. As a theoretical implication, using a KOS lens is critical in understanding the broader picture of what other similar knowledge structures there might be. This lens therefore enables us to define taxonomies, the inclusion criteria, and the exclusion criteria in a more precise way.

Furthermore, this study review and synthesize 28 research articles about taxonomy alignment methods. As mentioned, a practical contribution of this study is that these 28 studies can serve as a core corpus of taxonomy alignment literature for those who are interested in matching, mapping, aligning taxonomies.

Finally, the main methodological contributions of this study are three-fold: (1) to enhance the understanding of a common pipeline, or matching dimensions, of a taxonomy alignment work. We synthesize and develop characteristics that one can use to understand taxonomy alignment research; (2) to provide a novel categorization of taxonomy alignment approach into natural language processing approach, logic-based approach, and heuristic-based approach; (3) to provide a guideline for the matching dimensions and key characteristics that should be included in future taxonomy alignment research.

Acknowledgments

Cheng would like to thank her doctoral committee members, Dr. Bertram Ludäscher, Dr. Allen Renear, Dr. Karen Wickett, and Dr. Nico Franz, for their kind feedback on this work that is a vital component of her dissertation. We would also like to express our gratitude towards the reviewers and the editor for their constructive comments.

References

Aanen, S. S., Nederstigt, L. J., Vandić, D., & Frăsincar, F. (2012). Schema-an algorithm for automated product taxonomy mapping in e-commerce. *Extended Semantic Web Conference*, 300–314.

Angermann, H., Pervez, Z., & Ramzan, N. (2017). Taxo-Semantics: Assessing similarity between multi-word expressions for extending e-catalogs. *Decision Support Systems*, *98*, 10–25.

Baca, M. (2016). *Introduction to metadata*. Getty Publications.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34–43.

Brachman, R. J. (1983). What is-a is and isn't: An analysis of taxonomic links in semantic networks. *Computer*, (10), 30–36.

Chaffin, R., Herrmann, D. J., & Winston, M. (1988). An empirical taxonomy of part-whole relations: Effects of part-whole relation type on relation identification. *Language and Cognitive processes*, *3*(1), 17–48.

Chan, L., & Zeng, M. (2006). Metadata interoperability and standardization–a study of methodology part i. *D-Lib magazine*, *12*(6), 1082–9873.

Chen, M., Wu, C., Yang, Z., Liu, S., Chen, Z., & He, X. (2020). A multi-strategy approach for the merging of multiple taxonomies. *Journal of Information Science*.

Cheng, Y.-Y. (2022). *Agreeing to disagree: Applying a logic-based approach to reconciling and merging multiple taxonomies* (Doctoral dissertation). University of Illinois at Urbana-Champaign.

Cheng, Y.-Y., Franz, N., Schneider, J., Yu, S., Rodenhausen, T., & Ludäscher, B. (2017). Agreeing to disagree: Reconciling conflicting taxonomic views using a logic-based approach. *Proceedings of the Association for Information Science and Technology*, *54*(1), 46–56.

Cheng, Y.-Y., Hoang, L., & Ludäscher, B. (2020). Cacao, Cocao, or Cocoa? Reconciliation of taxonomic names in Biodiversity Heritage Library. *ISKO*.

Cheng, Y.-Y., & Ludäscher, B. (2020). Reconciling taxonomies of electoral constituencies and recognized tribes of indigenous Taiwan. *Proceedings of the Association for Information Science and Technology*, *57*(1), e248.

Daraio, E., Cagliero, L., Chiusano, S., Garza, P., & Ricupero, G. (2020). An explainable data-driven approach to web directory taxonomy mapping. *Procedia Computer Science*, *176*, 1101–1110.

de Andrade, J., & de Lara, M. L. G. (2016). Interoperability and mapping between knowledge organization systems: Metathesaurus—unified medical language system of the national library of medicine. *KO KNOWLEDGE ORGANIZATION*, *43*(2), 107–112.

Dou, D., McDermott, D., & Qi, P. (2005). Ontology translation on the semantic web. In *Journal on data semantics ii* (pp. 35–57). Springer.

Erfani, S. S., & Abedin, B. (2018). Impacts of the use of social network sites on users' psychological well-being: A systematic review. *Journal of the Association for Information Science and Technology*, *69*(7), 900–912.

Euzenat, J., Shvaiko, P., et al. (2007). *Ontology matching* (Vol. 18). Springer.

Franz, N. M., Chen, M., et al. (2016). Names are not good enough: Reasoning over taxonomic change in the Andropogon complex. *Semantic Web*, *7*(6), 645–667.

Franz, N. M., Chen, M., Yu, S., Kianmajd, P., Bowers, S., & Ludäscher, B. (2015). Reasoning over taxonomic change: Exploring alignments for the Perelleschus use case. *PloS one*, *10*(2), e0118247.

Gerstl, P., & Pribbenow, S. (1995). Midwinters, end games, and body parts: A classification of part-whole relations. *International journal of human-computer studies*, *43*(5-6), 865–889.

Giabelli, A., Malandri, L., Mercorio, F., & Mezzanzanica, M. (2022). WETA: Automatic taxonomy alignment via word embeddings. *Computers in Industry*, *138*, 103626.

Gilchrist, A. (2003). Thesauri, taxonomies and ontologies–an etymological note. *Journal of documentation*.

Hedden, H. (2016). *The accidental taxonomist*. Information Today, Inc.

Hlava, M. M. (2014). The taxobook: Principles and practices of building taxonomies, part 2 of a 3-part series. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *6*(4), 1–164.

Hodge, G. (2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files.* ERIC.

Inants, A., & Euzenat, J. (2015). An algebra of qualitative taxonomical relations for ontology alignments. *International Semantic Web Conference*, 253–268.

Johnston, M. A. (2016). Redefinition of the Eleodes Eschscholtz subgenera Tricheleodes Blaisdell and Pseudeleodes Blaisdell, with the description of a new species (Coleoptera: Tenebrionidae). *Annales Zoologici*, *66*(4), 665–679.

Jung, J. J. (2006). Taxonomy alignment for interoperability between heterogeneous digital libraries. *International conference on asian digital libraries*, 274–282.

Jung, J. J. (2008). Taxonomy alignment for interoperability between heterogeneous virtual organizations. *Expert Systems with Applications*, *34*(4), 2721–2731.

Kanehisa, M., Sato, Y., & Kawashima, M. (2022). KEGG mapping tools for uncovering hidden features in biological data. *Protein Science*, *31*(1), 47–53.

Kelly, D., & Sugimoto, C. R. (2013). A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, *64*(4), 745–770.

Lee, L.-H., Yu, Y.-T., & Huang, C.-R. (2009). Chinese wordnet domains: Bootstrapping chinese wordnet with semantic domain labels. *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 1*, 288–296.

Lehmann, F. (1992). Semantic networks. *Computers & Mathematics with Applications*, *23*(2-5), 1–50.

Li, J.-m., Zhao, H.-z., & Du, M.-j. (2011). Detection and resolution of structural conflictions in heterogeneous xbrl taxonomies. *The 5th International Conference on New Trends in Information Science and Service Science*, *2*, 312–317.

Li, Z., Xiao, X., Wang, M., Wang, C., Wang, X., & Xie, X. (2008). Towards the taxonomy-oriented categorization of yellow pages queries. *ACM Transactions on Internet Technology (TOIT)*, *11*(4), 1–27.

Library of Congress. (2008). Marc to dublin core crosswalk.

Lin, H., Liu, Y., Zhang, P., & Wang, J. (2019). Representation learning of taxonomies for taxonomy matching. *International Conference on Computational Science*, 383–397.

Maldonado, A., Sharpe, S., & ter Horst, P. (2021). Bootstrapping supervised product taxonomy mapping with hierarchical path translations for the regulatory intelligence domain. *ceur-ws.org.*

Mazzocchi, F. (2018). Knowledge organization system (kos): An introductory critical account. *Knowledge Organization*, *45*(1), 54–78.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Mouriño-Garcıa, M., Pérez-Rodrıguez, R., Anido-Rifón, L., Fernández-Iglesias, M. J., & Darriba-Bilbao, V. M. (2018). Cross-repository aggregation of educational resources. *Computers & Education*, *117*, 31–49.

Musgrove, T. (2006). Recognizing emergent nodes in aligning multiple document taxonomies. *Ontology Matching*, 226.

Nandi, A., & Bernstein, P. A. (2009). HAMSTER: Using search clicklogs for schema and taxonomy matching. *Proceedings of the VLDB Endowment*, *2*(1), 181–192.

Noy, N. F., & Musen, M. A. (2003). The PROMPT suite: Interactive tools for ontology merging and mapping. *International journal of human-computer studies*, *59*(6), 983–1024.

Otero-Cerdeira, L., Rodrıguez-Martınez, F. J., & Gómez-Rodrıguez, A. (2015). Ontology matching: A literature review. *Expert Systems with Applications*, *42*(2), 949–971.

Park, S., & Kim, W. (2007). Ontology mapping between heterogeneous product taxonomies in an electronic commerce environment. *International Journal of Electronic Commerce*, *12*(2), 69–87.

Pfeifer, K., & Peukert, E. (2013a). Integration of text mining taxonomies. *International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management*, 39–55.

Pfeifer, K., & Peukert, E. (2013b). Mapping text mining taxonomies. *KDIR/KMIS*, 5–16.

Pieterse, V., & Kourie, D. G. (2014). Lists, taxonomies, lattices, thesauri and ontologies: Paving a pathway through a terminological jungle. *KO Knowledge Organization*, *41*(3), 217–229.

Ponzetto, S. P., & Navigli, R. (2009). Large-scale taxonomy mapping for restructuring and integrating wikipedia. *IJCAI*, *9*, 2083–2088.

Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *the VLDB Journal*, *10*(4), 334–350.

Raunich, S., & Rahm, E. (2014). Target-driven merging of taxonomies with ATOM. *Information Systems*, *42*, 1–14.

Renz, J. (2002). *Qualitative spatial reasoning with topological information*. Springer.

Reynaud, C., & Safar, B. (2006). When usual structural alignment techniques don't apply. *Ontology Matching*, 191.

Shvaiko, P., & Euzenat, J. (2005). A survey of schema-based matching approaches. In *Journal on data semantics iv* (pp. 146–171). Springer.

Shvaiko, P., & Euzenat, J. (2011). Ontology matching: State of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, *25*(1), 158–176.

Siqueira, J., & Martins, D. L. (2022). Workflow models for aggregating cultural heritage data on the web: A systematic literature review. *Journal of the Association for Information Science and Technology*, *73*(2), 204–224.

Smiraglia, R. P. (2014). Taxonomy. *The Elements of Knowledge Organization*, 51–55.

Soergel, D. (1999). The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*, *50*(12), 1119–1120.

Soergel, D. (2009). Knowledge organization systems: Overview.

Sowa, J. F. (1992). Conceptual graphs as a universal knowledge representation. *Computers & Mathematics with Applications*, *23*(2-5), 75–93.

Subramaniam, L. V., Nanavati, A. A., & Mukherjea, S. (2009). Enriching one taxonomy using another. *IEEE transactions on knowledge and data engineering*, *22*(10), 1415–1427.

Thau, D., & Ludäscher, B. (2007). Reasoning about taxonomies in first-order logic. *Ecological Informatics*, *2*(3), 195–209.

Van Hooland, S., & Verborgh, R. (2014). *Linked data for libraries, archives and museums: How to clean, link and publish your metadata*. Facet publishing.

Wang, Z., Lv, Q., Lan, X., & Zhang, Y. (2018). Cross-lingual knowledge graph alignment via graph convolutional networks. *Proceedings of the 2018 conference on empirical methods in natural language processing*, 349–357.

Woods, W. A. (1975). What's in a link: Foundations for semantic networks. In *Representation and understanding* (pp. 35–82). Elsevier.

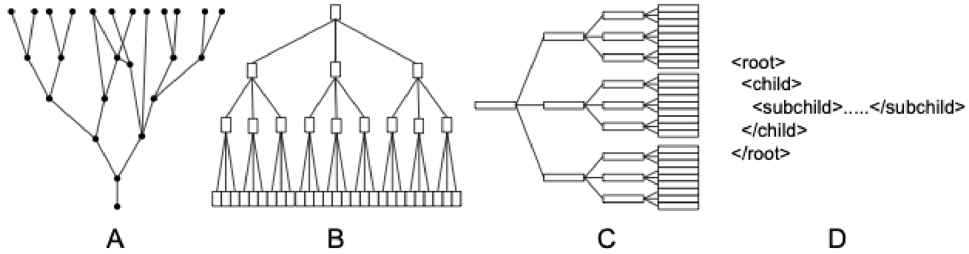
Xu, S., & Sun, M. (2007). Leveraging world knowledge in chinese text classification. *Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007)*, 33–38.

Zeng, M. (2008). Knowledge organization systems (KOS). *Knowledge Organization*, *35*(2-3), 160–182.

Zeng, M. (2019). Interoperability. *KO Knowledge Organization*, *46*(2), 122–146.

Zeng, M. (2021). Shape of trees.

Zeng, M., & Chan, L. (2006). Metadata interoperability and standardization-a study of methodology, part II. *D-Lib Magazine*, *12*(6), 1082–9873.



Tree Shapes of a taxonomy input: (A) figurative tree; (B) vertical tree; (C) horizontal tree; (D) general hierarchy. Terminologies for tree shapes (A), (B), (C) adopted from Zeng (2021).

228x76mm (72 x 72 DPI)

Table I

Characteristics of Taxonomy Alignment Methods -1

		ATOM [1]	**Chen** [2]	**CROEA** [3]	**Daraio** [4]	**Euler/X** [5]	**HAMSTER** [6]	**Jung** [7]
Application Areas		automobile, anatomy, e-commerce	e-commerce	education	e-commerce	geography, biodiversity	e-commerce	digital library, virtual organizations
Matching Dimensions								
Input	**Tree shapes**	vertical tree, horizontal tree	texts, vertical tree	texts, vertical tree	vertical tree	vertical tree, horizontal tree	general hierarchy	vertical tree
	Variable names	Os, Ot	Tt, Ts1, Ts2,. . . , Tsm	lom = (w1m, w2m. . .)	D= {d1,. . . dn}	T1, T2..Tn	Ts, Tt	Ti, Tj
	Symmetry	asymmetric	asymmetric	asymmetric	asymmetric	symmetric	asymmetric	symmetric
Approach	**Methodology**	logic-based	logic-based, heuristic-based, NLP-unsupervised	NLP-supervised	NLP-supervised	logic-based	NLP-unsupervised	NLP-unsupervised
	Unit of Matching	element-level, structure-level	element-level, structure-level	element-level	element-level	element-level, structure-level	element-level, structure-level	element-level
	Comparison Type	Pairwise	Multiple	Multiple	pairwise	multiple	pairwise	pairwise
	Relation Type	equivalence, is-a, inverse is-a	equivalence, is-a, inverse is-a	text similarity	text similarity	equivalent, disjoint, overlap, proper-part, proper-part inverse	text similarity	text similarity
Output	**Number of Merged Solution(s)**	N=1	N=1	N=1	N=1	N=1, N>1	N=1	N=1
	Information Preserved	Somewhat	Somewhat	No	No	Yes	No	No

[1] ATOM in Raunich and Rahm, 2014; [2]Chen et al., 2020; [3] CROEA in Mouriño-Garcıa et al., 2018; [4] Daraio et al., 2020; [5] Euler/X in Cheng et al., 2017; Cheng et al., 2020; Cheng and Ludäscher, 2020; Franz, Chen, et al., 2016; Franz et al., 2015; Johnston, 2016; Thau and Ludäscher, 2007; [6] HAMSTER in Nandi and Bernstein, 2009; [7] Jung, 2006, 2008

Table II

Characteristics of Taxonomy Alignment Methods -2

		Kanehisa [8]	**Lee [9]**	**Li'11 [10]**	**Li'08 [11]**	**Maldonado [12]**	**Merge-Into [13]**	**Musgrove [14]**
Application Areas		biology	cross-lingual	finance	business	business	computer science, general ontologies	e-commerce
Matching Dimensions								
Input	**Tree shapes**	vertical tree	texts	horizontal tree	vertical tree	texts	vertical tree	texts
	Variable names	T1, T2	Chinese WordNet, WordNet Domains	Tszse, Tcas	Tsic, Ttarget	Tvendor, TGPC	Tsrc, Tdest	TDMOZ, TYahoo, TAbout, Tt
	Symmetry	symmetric	asymmetric	asymmetric	asymmetric	asymmetric	asymmetric	asymmetric
Approach	**Methodology**	heuristic-based	NLP-supervised	heuristic-based	NLP-unsupervised	NLP-supervised	heuristic-based	NLP-unsupervised
	Unit of Matching	element-level	element-level	structure-level	element-level	element-level	element-level, structure-level	element-level
	Comparison Type	pairwise	pairwise	pairwise	pairwise	pairwise	multiple	pairwise
	Relation Type	author-defined relations	text similarity	author-defined relations	text similarity	text similarity	author-defined relations	text similarity
Output	**Number of Merged Solution(s)**	N=1	N=1	N=1	N=1	N/A	N=1	N=1
	Information Preserved	Yes	No	No	No	No	Somewhat	No

[8]Kanehisa et al., 2022; [9]Lee et al., 2009; [10]J.-m. Li et al., 2011; [11]Z. Li et al., 2008; [12]Maldonado et al., 2021; [13] Merge-Into in Subramaniam et al., 2009; [14] Musgrove, 2006

Table III

Characteristics of Taxonomy Alignment Methods -3

		Pfeifer [15]	**Ponzetto [16]**	**Reynaud [17]**	**Taxo-Semantics [18]**	**WETA [19]**	**Xu [20]**
Application Areas		business	Wikipedia	general	e-commerce	business	cross-lingual
Matching Dimensions							
Input	**Tree shapes**	horizontal tree	vertical tree	vertical tree	texts, vertical tree	vertical tree	texts, vertical tree
	Variable names	Tsj, Ttk	T, T’	Ts, Tt	Ts,Tt, Background Knowledge	To, Td	TCE, TCODP
	Symmetry	asymmetric	asymmetric	asymmetric	asymmetric	asymmetric	asymmetric
Approach	**Methodology**	NLP-supervised	NLP-supervised	NLP-unsupervised	NLP-unsupervised	NLP-supervised	NLP-supervised
	Unit of Matching	element-level	element-level	element-level, structure-level	element-level, structure-level	element-level	element-level
	Comparison Type	pairwise	pairwise	pairwise	pairwise	pairwise	pairwise
	Relation Type	text similarity; equivalent, subtype, inverse subsumption, associative	text similarity	text similarity	text similarity	text similarity	text similarity
Output	**Number of Merged Solution(s)**	N=1	N=1	N=1	N=1	N=1	N=1
	Information Preserved	No	No	No	No	No	No

[15]Pfeifer and Peukert, 2013a, 2013b; [16]Ponzetto and Navigli, 2009; [17]Reynaud and Safar, 2006; [18] Taxo-Semantics in Angermann et al., 2017; [19] WETA in Giabelli et al., 2022; [20] Xu and Sun, 2007

Appendix 1
Partial Metadata of the Taxonomy Alignment Studies Examined
(see csv file for the full metadata)

*Index No. in this table reflects the numbering of the articles/projects in the manuscript's Table I, II, and III. Duplicated numbers indicate that the studies are follow-up studies, or group research from the same group of authors.

No.	Authors	Methods	Title	Year	Source title	Volume	Issue	DOI
1	Raunich S., Rahm E.	ATOM	Target-driven merging of taxonomies with Atom	2014	Information Systems	42		10.1016/j.is.2013.11.001
2	Chen M., Wu C., Yang Z., Liu S., Chen Z., He X.	Chen et al 2020	A multi-strategy approach for the merging of multiple taxonomies	2020	Journal of Information Science			10.1177/0165551520952340
3	Mouriño-García M., Pérez-Rodríguez R., Anido-Rifón L., Fernández-Iglesias M.J., Darriba-Bilbao V.M.	CROEA	Cross-repository aggregation of educational resources	2018	Computers and Education	117		10.1016/j.compedu.2017.09.014
4	Daraio E., Cagliero L., Chiusano S., Garza P., Ricupero G.	Daraio et al 2020	An explainable data-driven approach to web directory taxonomy mapping	2020	Procedia Computer Science	176		10.1016/j.procs.2020.09.106
5	Cheng Y.-Y., Franz N., Schneider J., Yu S., Rodenhausen T., Ludäscher B.	Euler/X	Agreeing to disagree: Reconciling conflicting taxonomic views using a logic-based approach	2017	Proceedings of the Association for Information Science and Technology	54	1	10.1002/pra2.2017.14505401006
5	Cheng, YY; Hoang, KL; Ludascher, B	Euler/X	Cacao, Cocao, or Cocoa? Reconciliation of Taxonomic Names in Biodiversity Heritage Library	2020	KNOWLEDGE ORGANIZATION AT THE INTERFACE: PROCEEDINGS OF THE SIXTEENTH INTERNATIONAL ISKO CONFERENCE, 2020	17		
5	Cheng, YY; Ludascher, B	Euler/X	Reconciling taxonomies of electoral constituencies and	2020	Proceedings of the Association for Information Science & Technology	57	1	10.1002/pra2.248

No.	Authors	Methods	Title	Year	Source title	Volume	Issue	DOI
			recognized tribes of indigenous Taiwan					
5	Franz N.M., Chen M., Kianmajd P., Yu S., Bowers S., Weakley A.S., Ludäscher B.	Euler/X	Names are not good enough: Reasoning over taxonomic change in the Andropogon complex	2016	Semantic Web	7	6	10.3233/SW-160220
5	Franz N.M., Chen M., Yu S., Kianmajd P., Bowers S., Ludäscher B.	Euler/X	Reasoning over taxonomic change: Exploring alignments for the Perelleschus Use Case	2015	PLoS ONE	10	2	10.1371/journal.pone.0118247
5	Johnston M.A.	Euler/X	Redefinition of the Eleodes Eschscholtz Subgenera Tricheleodes Blaisdell and Pseudeleodes Blaisdell, with the Description of a New Species (Coleoptera: Tenebrionidae)	2016	Annales Zoologici	66	4	10.3161/00034541ANZ2016.66.4.018
5	Thau D., Ludäscher B.	Euler/X	Reasoning about taxonomies in first-order logic	2007	Ecological Informatics	2	3 SPEC. ISS.	10.1016/j.ecoinf.2007.07.005
6	Nandi A., Bernstein P.A.	HAMSTER	HAMSTER: Using search clicklogs for schema and taxonomy matching	2009	Proceedings of the VLDB Endowment	2	1	10.14778/1687627.1687649
7	Jung J.J.	Jung 2006 2008	Taxonomy alignment for interoperability between heterogeneous digital libraries	2006	Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)	4312 LNCS		10.1007/11931584_30
7	Jung J.J.	Jung 2006 2008	Taxonomy alignment for interoperability between heterogeneous virtual organizations	2008	Expert Systems with Applications	34	4	10.1016/j.eswa.2007.05.015
8	Kanehisa M., Sato Y., Kawashima M.	Kanehisa et al 2022	KEGG mapping tools for uncovering hidden features in biological data	2022	Protein Science	31	1	10.1002/pro.4172

No.	Authors	Methods	Title	Year	Source title	Volume	Issue	DOI
9	Lee L.-H., Yu Y.-T., Huang C.-R.	Lee et al 2009	Chinese WordNet domains: Bootstrapping Chinese WordNet with semantic domain labels	2009	PACLIC 23 - Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation	1		
10	Ji-Mei L., Hui-Zhou Z., Mei-Jie D.	Li at al 2011	Detection and resolution of structural conflictions in heterogeneous XBRL taxonomies	2011	Proceedings - 5th International Conference on New Trends in Information Science and Service Science, NISS 2011	2		
11	Li Z., Xiao X., Wang M., Wang C., Wang X., Xie X.	Li et al 2012	Towards the taxonomy-oriented categorization of yellow pages queries	2012	ACM Transactions on Internet Technology	11	4	10.1145/2109211.2109213
12	Maldonado A., Sharpe S., Horst P.T.	Maldonado et al 2021	Bootstrapping Supervised Product Taxonomy Mapping with Hierarchical Path Translations for the Regulatory Intelligence Domain	2021	CEUR Workshop Proceedings	3063		
13	Subramaniam L.V., Nanavati A.A., Mukherjea S.	Merge-Into	Enriching one taxonomy using another	2010	IEEE Transactions on Knowledge and Data Engineering	22	10	10.1109/TKDE.2009.189
14	Musgrove T.	Musgrove 2006	Recognizing emergent nodes in aligning multiple document taxonomies	2006	CEUR Workshop Proceedings	225		
15	Pfeifer K., Peukert E.	Pfeifer 2013a, b	Integration of text mining taxonomies	2015	Communications in Computer and Information Science	454		10.1007/978-3-662-46549-3_3
15	Pfeifer K., Peukert E.	Pfeifer 2013a, b	Mapping text mining taxonomies	2013	IC3K 2013; KDIR 2013 - 5th International Conference on Knowledge Discovery and Information Retrieval and KMIS 2013 - 5th International Conference on Knowledge Management and Information Sharing, Proc.			10.5220/0004500400050016

No.	Authors	Methods	Title	Year	Source title	Volume	Issue	DOI
16	Ponzetto S.P., Navigli R.	Ponzetto et al 2009	Large-scale taxonomy mapping for restructuring and integrating Wikipedia	2009	IJCAI International Joint Conference on Artificial Intelligence			
17	Reynaud C., Safar B.	Reynaud & Safar 2006	When usual structural alignment techniques don't apply	2006	CEUR Workshop Proceedings	225		
18	Angermann H., Pervez Z., Ramzan N.	Taxo-Semantics	Taxo-Semantics: Assessing similarity between multi-word expressions for extending e-catalogs	2017	Decision Support Systems	98		10.1016/j.dss.2017.04.001
19	Giabelli A., Malandri L., Mercorio F., Mezzanzanica M.	WETA	WETA: Automatic taxonomy alignment via word embeddings	2022	Computers in Industry	138		10.1016/j.compind.2022.103626
20	Xu S., Sun M.	Xu & Sun 2007	Leveraging world knowledge in chinese text classification	2007	Proceedings - ALPIT 2007 6th International Conference on Advanced Language Processing and Web Information Technology			10.1109/ALPIT.2007.105

Appendix 2:
Annotation Procedures for Taxonomy Alignment Method Systematic Review Task

The purpose of this annotation task is to annotate the matching dimensions involved in a taxonomy alignment method study. In particular, a taxonomy alignment study comprises three dimensions: input, approach, output. In each dimension, there are three to four characteristics of that dimension. Your task is to identify the **values** for each of the characteristic. For characteristics that are [categorical], use the exact same values mentioned here and on the spreadsheet to annotate the studies. For characteristics that have [free input available] as a choice, follow the example values mentioned here first – if none of the example values are applicable, use the values mentioned from the study to input into the spreadsheet table. Pick multiple values if more than one category applies.

1. **Application Area:** we define application area as the main topic of taxonomies the study mentions or present a use case or demonstrate an evaluation on. These can be *biology*, *e-commerce, cross-lingual, business, geography*, or other areas mentioned from the study. [free input available]

2. **Tree shapes**: we define tree shapes as *texts*, *figurative trees, vertical trees, horizontal trees, general hierarchy*. [categorical]

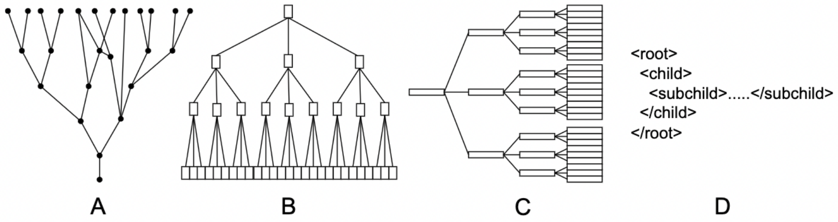


Figure 1

Tree Shapes of a taxonomy input: (A) figurative tree; (B) vertical tree; (C) horizontal tree; (D) general hierarchy. Terminologies for tree shapes (A), (B), (C) adopted from (Zeng, 2021).

3. **Variable names**: we define variable names as the input axioms, names, or labels the taxonomy alignment studies use to describe their inputs. For example, *T1, T2, Tn*. [free input available]

4. **Symmetry**: we define symmetry as *asymmetric* or *symmetric*. *Asymmetric* is when the taxonomy alignment study states that the input taxonomies have a ‘source’ and a ‘target’

taxonomy. *Symmetric* is when there is no clear distinction between source and target taxonomies. [categorical]

5. **Methodology**: we define methodology as *NLP-supervised, NLP-unsupervised, logic-based, heuristic-based*. *NLP-supervised* is when there is clearly a labeled training dataset. *NLP-unsupervised* is when there is no training dataset. *Logic-based* is when the approach uses RCC-5 like matching measures (equivalent, disjoint, overlap, proper-part, proper-part inverse). *Heuristic-based* is when the approach defines its own discrete measures or rules that do not fit with the logic-based or the NLP approaches. [categorical]

6. **Unit of matching:** we define unit of matching as *element-level* or *structure-level* matching. *Element-level* matching focuses more on the matching of concepts, names, nodes in the taxonomies, while *structure-level* focuses on the placement or position of the concepts in the taxonomies. [categorical]

7. **Comparison type:** we define comparison type as *pairwise* or *multiple* comparison. *Pairwise* comparison means the approach only compares two taxonomies. *Multiple* comparison means the approach compares more than two taxonomies. [categorical]

8. **Relation type**: For relation types, these could be RCC-5 like types – *equivalent, disjoint, overlap, proper-part, proper-part inverse*. Or these can be similarity-based relation types, such as *text similarity*. *Author-defined relations*. Other types might also exist, depending on how the study describes them. [free input available]

9. **Number of Merged Solutions:** If the alignment only produces one merged solution, then the number is *n=1*. If the alignment produces more than one solution, then the number of is *n>1*. If the alignment did not produce any solution (e.g. due to logical inconsistency or unforeseeable circumstances), then the number of possible worlds is *n=0.* There may also be cases when it is difficult to determine if a taxonomy alignment produces any one or multiple merged solution(s). If that is the case, use the label *N/A* for not applicable. [categorical]

10. **Information preserved:** This is a *Yes*, *Somewhat*, *No* field. If the taxonomy alignment solution still shows the original taxonomies from the input, then answer this field as ‘Yes’. Otherwise, answer ‘No’ for this field. If the authors stated in their paper about partially preserving the information, then answer ‘Somewhat’ for this field. [categorical]