Agreeing to Disagree: Reconciling Conflicting Taxonomic Views Using a Logic-based Approach

Yi-Yun Cheng
School of Information Sciences, University of Illinois at Urbana-Champaign, USA. yiyunyc2@illinois.edu

Nico Franz
School of Life Sciences, Arizona State University, USA. nico.franz@asu.edu

Jodi Schneider
School of Information Sciences, University of Illinois at Urbana-Champaign, USA. jodi@illinois.edu

Shizhuo Yu
Department of Computer Science, University of California at Davis, USA. szyu@ucdavis.edu

Thomas Rodenhausen
School of Information, University of Arizona, USA. thomas.rodenhausen@gmail.com

Bertram Ludäscher
School of Information Sciences, University of Illinois at Urbana-Champaign, USA. ludaesch@illinois.edu

ABSTRACT

Taxonomy alignment is a way to integrate two or more taxonomies. Semantic interoperability among datasets, information systems and knowledge bases is facilitated by combining the different input taxonomies into merged taxonomies that reconcile apparent differences or conflicts. We show how alignment problems can be solved with a logic-based region connection calculus (RCC-5) approach, using five base relations to compare concepts: *congruence*, *inclusion*, *inverse inclusion*, *overlap* and *disjointness*. To illustrate this method, we use different geo-taxonomies, which organize the United States into several, apparently conflicting, geospatial hierarchies. For example, we align T_{CEN}, a taxonomy derived from the Census Bureau's regions map, with T_{NDC}, from the National Diversity Council (NDC), and with T_{TZ}, a taxonomy capturing the U.S. time zones. Using these case studies, we show how this logic-based approach can reconcile conflicts between taxonomies. We have implemented these case studies with an open source tool called Euler/X which has been applied primarily for solving complex alignment problems in biological classification. In this paper, we demonstrate the feasibility and broad applicability of this approach to other domains and alignment problems in support of semantic interoperability.

KEYWORDS

taxonomy alignment, RCC-5, semantic interoperability

80th Annual Meeting of the Association for Information Science & Technology, Washington, DC | Oct. 27-Nov. 1, 2017

INTRODUCTION

[Amy and Tina meet at the water cooler]

***Tina**: Hey Amy, can you recommend a signature dish from where you live?*

***Amy**: Oh, definitely the half-smokes from the Northeast! They are these tasty half-pork and half-beef sausages.*

***Tina**: What a coincidence! We have half-smokes in the South, too! Where do you live in the Northeast? New York? Boston?*

***Amy**: Wrong guesses! Where do you live in the South?*

Tina and Amy together: Washington, D.C.

[The two of them look at each other, confused.]

Vocabulary misunderstandings are common in our everyday lives. According to Bowker and Star (2000), "in the face of incompatible information or data structures among users or among those specifying the system, attempts to create unitary knowledge categories are futile. Rather, parallel or multiple representational forms are required" (p.159). In this fictional dialogue, neither Tina nor Amy is wrong. Perhaps Tina viewed Washington D.C. from the National Diversity Council's perspective of the United States, whereas Amy saw the United States according to the Census Regions map (Figure 1). We could attempt to ask the two to agree on one of the two taxonomies, but other human communicators will still have other ways to classify the United States with yet other categories.

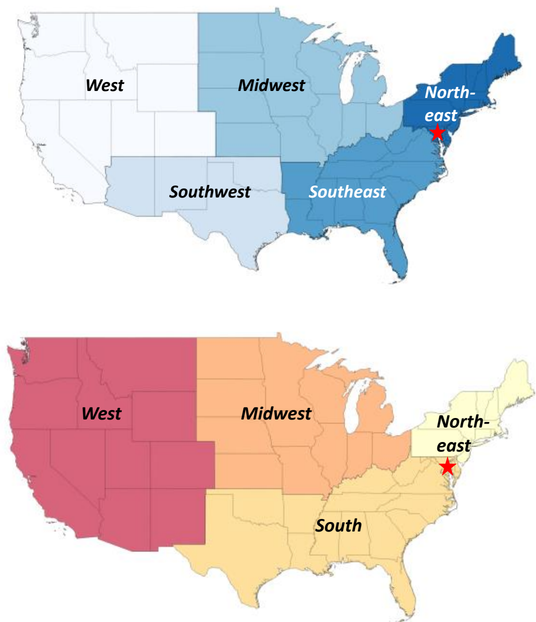

Figure 1. *Different geo-taxonomies of the contiguous United States, i.e., the 48 adjoining states and Washington DC (red star): National Diversity Council map NDC (top) and Census Bureau map CEN (bottom)*

The information sciences community has studied semantic interoperability extensively. For example, equivalence, hierarchical and associative relationships are often used to align multiple knowledge organization systems (KOS). However, these relationships can be defined ambiguously. For instance, equivalency among KOS can mean synonyms or near-synonyms; hierarchy may refer to exceedingly generic is-a relationships; and associative relationships may indicate everything that seems relevant. Such mapping relationships can be especially troublesome in crosswalks because they attempt to create a 'unitary' view of knowledge among different structures that are each internally coherent when used in specific contexts, but which appear to be in conflict when combined indiscriminately. In the end, this may result in data quality problems in the original KOS, such as granularity or meaning loss (de Andrade & Lopes Ginez de Lara, 2016; Chan & Zeng, 2006; Zeng & Chan, 2009).

We draw attention particularly to one type of KOS –taxonomies. Taxonomies, in a broad sense, are hierarchical structures that group similar objects together (Hodge, 2000). *Taxonomy alignment* is the term we use to address the issue of bridging, mapping or aligning two or more taxonomies. Taxonomies can be aligned in reference to a variety of similarity indicators, such as nomenclatural relationships, member composition or diagnostic features of child or parent nodes (Franz et al, 2015, 2016a, 2016b).

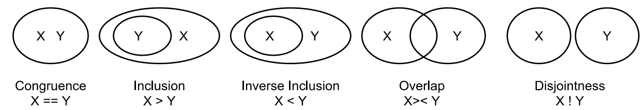

Figure 2*. The region connection relations used in RCC-5*

Many fields have contributed to taxonomy alignment research, but the key new idea in our line of work (Chen et al., 2014; Franz et al., 2015, 2016a, 2016b; Thau & Ludäscher, 2007) is to compare concepts *X* and *Y* via five exhaustive and mutually exclusive relationships – *congruence*, *inclusion*, *inverse inclusion*, *overlap* and *disjointness* (Figure 2) from the region connection calculus RCC-5 (Randell et al., 1992; Cohn & Renz, 2008).

Our objectives for this paper are two-fold: first, we show the feasibility of using a logic-based approach for taxonomy alignment with RCC-5. Second, we illustrate the value of multi-taxonomy alignments in a way that both preserves the internal coherence and context of the input taxonomies T_1 and T_2, while at the same time providing a reconciled combined view T_3. T_3 exposes all relations between concepts in T_1 and T_2, implied by the input articulations *A*. We do this by providing two case studies that use a logic-based tool named Euler/X to align several geo-taxonomies derived from U.S. maps. First, we will align the National Diversity Council (NDC) map with the Census Bureau's Census Regions map (CEN), on the state-level, taking a bottom-up approach. Second, we will align the United States Time Zone (TZ) and CEN, on a regional level, taking a top-down approach.

RELATED WORK

Semantic Interoperability

Semantic interoperability is the capability for communication ensuring that the *meaning* or semantics of data, information or knowledge across sources is bridged. Semantic interoperability is particularly relevant between different knowledge organization systems (KOS), which are controlled vocabularies used in organizing information. Types of KOS include term lists, classifications (subject headings or taxonomies) and relationship models (Hodge, 2000). If existing KOS are different in structure, domain, language or granularity, it is necessary for the KOS to be transformed, mapped or merged to enable interoperability. Semantic interoperability in taxonomies means that the definitions and relations within different taxonomy systems are well-mapped (Zeng & Chan, 2009).

Three main types of mapping relationships between KOS (especially in the thesauri context) are equivalence, hierarchical and associative relationships. In crosswalks, in fact, equivalent mapping (or exact mapping) is the most prevalent relationship for linking two classification schemes together, and this is termed *absolute crosswalking*. It is usually used when two concepts in two or more KOS are synonyms or near-synonyms; and if there is no equivalent counterpart, absolute crosswalking mapping will not occur.

A less strict crosswalking approach is relative crosswalking, which means that all entities in one KOS will be mapped to at least one entity in the other KOS. However, this mapping process disregards whether or not the two entities are semantically equivalent. As long as there is a relevant entity to be matched with, relative crosswalking will occur.

Each crosswalking approach has its pros and cons. Absolute crosswalking ensures the equivalencies in concepts, but the data values will be empty whenever the mapping did not occur; relative crosswalking can overcome the data conversion problem that will happen with absolute crosswalking, but it will result in data quality problems if we want to trace back and resurrect the original KOS, such as losing granularity or missing meanings in the final integrated mapping work (de Andrade, J., & Lopes Ginez de Lara, 2016; Chan & Zeng, 2006; Zeng & Chan, 2009).

Equivalent mapping further creates problems when the original KOS includes a membership is-a condition. Different KOS differ strongly in granularity, naming systems and modeling styles, regardless of using either manual or automatic matching tools, researchers usually only map the equivalency on the superclass-level. Members in a superclass that were not mapped just automatically inherit the equivalency to the counterpart superclass in the other KOS (Kless, Milton, Kazmierczak & Lindenthal, 2015; Pfeifer & Peukert, 2015).

The Taxonomy Alignment Problem (TAP)

From the KOS perspective, we have said that taxonomies, in a broad sense, are hierarchical structures that group similar objects together (Hodge, 2000). Taxonomies permit the assembly of multiple alternative, internally coherent hierarchies where all concepts derived from one hierarchy can be connected via parent-child ("is-a") relationships (Thau & Ludäscher, 2007; Thau, Bowers, & Ludäscher, 2008).

In this paper, we focus on *taxonomy alignment problems* (TAP), where the given taxonomies T_1, T_2 are inter-linked (or "formally crosswalked") via a set of input *articulations A*, defined as RCC-5 relations (see below), to yield a "merged" taxonomy T_3. Any solution taxonomy T_3 of a TAP must satisfy all logic constraints implied by the inputs T_1, T_2, and A (so T_3 must be *sound*) and must reveal all pairwise relationships between concepts in T_1 and T_2 (so T_3 must be *complete*). A TAP with inputs T_1, T_2, and A may have zero, one or many solutions T_3, in which case we call the input TAP, *inconsistent*, *unique* or *ambiguous*, respectively. If a TAP is inconsistent, this means it contains a logical contradiction, and some articulation constraints in A have to be relaxed (e.g., adding a disjunction to allow for multiple options), repaired (choosing a different articulation relationship) or even dropped. Multiple solutions T_3 are generated by the Euler/X tool, e.g., in the form of multiple *answer sets* (Gebser, Kaminski, Kaufmann & Schaub, 2012) or *possible worlds* (Denecker, Lierler, Truszczynski & Vennekens, 2012).

Different fields have contributed to taxonomy alignment research. In discussing semantic heterogeneity problems in digital libraries, Jung (2006) computes the similarity in distance between the concepts in different digital libraries and tries to align them. In text mining, Pfeifer and Peukert (2015) proposed a schema-based and instance-based similarity alignment for taxonomies using statistical metrics. We propose to use RCC-5 with its five base relations when aligning taxonomies, and we propose this as a complementary, "high resolution" approach for cross walking and integrating taxonomies.

Though taxonomy alignment problems are applicable in many domains, such as corporate taxonomies, website taxonomies and scientific taxonomies (Souza, Tudhope, & Almeida, 2012), Linnaean taxonomies are probably one of the oldest and most prominent examples. These taxonomies organize (typically) perceived species of living organisms into higher-level classifications (Franz & Thau, 2010). Concepts can be aligned in reference to a variety of similarity indicators, such as nomenclatural relationships, member composition or diagnostic features of child or parent nodes (Franz et al., 2015). Other domains include phonetics and genealogies and taxonomies in comparative sub-disciplines as well as in conservation (Franz et al., 2016b). As described in Franz et al. (2016a), biological classifications evolve in light of new knowledge while re-using existing names under a complex set of rules. Over time, this leads to many-to-many relationships between properly formed taxonomic names and their (former versus current) biological meanings (Remsen, 2016). The temporal evolution of biological classifications can be complex, due to the interaction of new and revised insights with naming rules established in this community. Therefore, if there is a need to integrate biodiversity data across taxonomies, using just taxonomic names and nomenclatural relationships will not be enough to understand and reconcile the different taxonomic perspectives (Franz et al., 2016a).

Euler/X

Euler/X (https://github.com/EulerProject/) is an open source logic-based tool that uses RCC-5 to align and reconcile taxonomies. Different underlying reasoners can be used by Euler/X, solving TAPs via FO (first-order), ASP (answer set programming) or direct RCC reasoning, i.e., X∈{FO, ASP, RCC}. A detailed system overview of Euler/X can be found in Chen et al. (2014). We have used the current Euler2 prototype of the toolkit in our research. The following are among the terms related to the Euler/X tool:

Articulations

An articulation is a constraint or rule that defines a relationship (a set constraint) between two concepts from different taxonomies. Each articulation is of the form X o Y, where X and Y are concepts from T_1 and T_2, respectively, and where "o" is a RCC-5 relation. Articulations can come from domain experts (who may link them to the relevant underlying evidence), or may be generated from data directly. Given two taxonomies T_1, T_2, an example of an articulation is [T_1.Cherry_Blossom **equals** T_2.Sakura]. Taxonomy T_1 has the concept Cherry_Blossom, whereas in T_2 its equivalent counterpart concept is named Sakura.

Region Connection Calculus (RCC)

RCC is used for qualitative (often, but not necessarily spatial) representation and reasoning and can be seen as a decidable fragment of first-order predicate logic (Cohn & Renz, 2008). This fragment is also closely related to set constraints and monadic first-order logic (Bachmair, Ganzinger & Waldmann, 1993; Bodirsky & Hils, 2012). RCC-5 is the variant of the RCC family used in Euler/X, consisting of five pair-wise disjoint and mutually exclusive relations (Figure 2), i.e., *congruence* (equals or "==" in Euler/X-generated figures), *proper inclusion* (includes or ">" in Euler/X), *inverse proper inclusion* (is_included_in or "<"), *overlap* (overlaps or "><"), and *disjointness* (disjoint or "!") .

Possible Worlds

When encoding and solving TAPs via ASP, the different answer sets represent alternative taxonomy merge solutions or possible worlds (PWs). Each PW satisfies all given TAP constraints and assigns exactly one of the basic five RCC-5 relations to each pair of concepts, removing disjunctive ambiguity. Thus, each PW provides a distinct solution for satisfying all input conditions.

RESEARCH DESIGN

We use a case study design with two cases. Figure 3 shows how we implemented logic-based mapping using Euler/X. The three main steps to implement this logic-based approach are:

Step 1. Supply input taxonomies T_1 and T_2
Step 2. Formulate RCC-5 articulations between T_1 and T_2
Step 3. Iteratively edit articulations in Euler/X

Step 1: Supply input taxonomies T_1 and T_2

Our two cases use these three maps rendering the contiguous United States in three different ways:

1. The Census Regions Map1 (CEN), consists of four regions: West, Midwest, Northeast and South, i.e., the contiguous 48 states and Washington D.C.
2. The National Diversity Council Map2 (NDC), consists of five regions: West, Southwest, Midwest, Northeast, Southeast, the 48 states and Washington D.C.
3. The Time Zone map3 (TZ), consists of four regions: Pacific, Mountain, Central and Eastern.

To prepare these regions for entry into Euler/X, we transcribe each map into a hierarchical taxonomic structure in plain text format, (**PARENT** CHILD1 CHILD2 …):

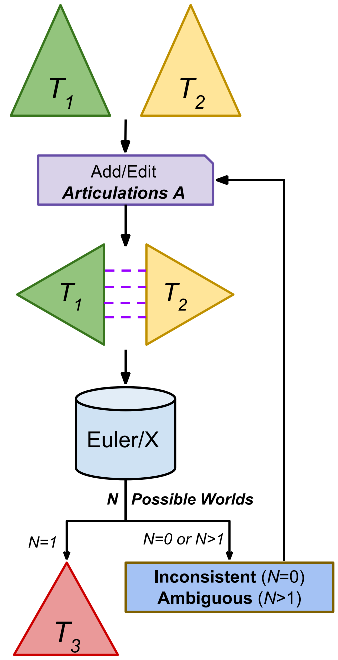

Figure 3. *The process of aligning taxonomies T1 and T2 with Euler/X*

1. CEN

For example, in the CEN case, the four regions Northeast, Midwest, South and West are all children of USA:

(USA Northeast Midwest South West**)**

Then, we indicate the states within each region (States abbreviated in 2 letters) in the same parent-child structure:

(Northeast CT MA ME NH NJ NY PA RI VT**)**

(Midwest IL IN IA KS MI MN MO NE ND OH SD WI**)**

(South AL AR DE DC FL GA KY LA MD MS NC OK SC TN TX VA WV**)**

(West AZ CA CO ID MT NV NM OR UT WA WY**)**

2. NDC

Analogously, we convert the NDC map into this taxonomic structure:

(USA Midwest Northeast Southeast Southwest West**)**

(Northeast CT DC DE MD MA ME NH NJ NY PA RI VT**)**

(Midwest IA IL IN KS MI MN MO ND NE OH SD WI**)**

(Southeast AL AR FL GA KY LA MS NC SC TN VA WV**)**

(Southwest AZ NM OK TX)

(West CA CO ID MT NV OR WA WY UT**)**

[1] CEN: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
[2] NDC: http://www.nationaldiversitycouncil.org/wp-content/uploads/2011/12/us_regions.jpg
[3] TZ:https://nationalmap.gov/small_scale/printable/images/pdf/reference/timezones4.pdf

3. TZ

In the TZ taxonomy, a single state can have counties with different time zones. For example, some parts of Indiana use Central Time while other parts use Eastern Time. This means that there are overlaps between the regional boundaries of times (e.g., Eastern, Central) and those of states (e.g., Indiana). Also, the TZ map itself (Figure 4) does not label the state-level concept names, therefore, we modelled the TZ taxonomy with just one line:

(USA Pacific Mountain Central Eastern**)**

Step 2: Formulate RCC-5 articulations

In our case studies, we are comparing (1) the CEN taxonomy with the NDC taxonomy and (2) the CEN taxonomy and the TZ taxonomy for the contiguous United States (48 states and D.C.). The comparisons are fairly straightforward from looking at the maps so that users can easily formulate their own articulations. In general, in order to compare taxonomies in a specific domain, we may need a domain expert's input on how to formulate the articulations between taxonomies. Using RCC-5 relationships equals (==), includes (>), is_included_in (<), overlaps (><) and disjoint (!), the following are the articulations we formulated:

(1) CEN vs. NDC (49 articulations, all with 'equal' relationships; only 5 shown here):

[CEN.AL equals NDC.AL]
[CEN.AR equals NDC.AR]
[CEN.AZ equals NDC.AZ]
[CEN.CA equals NDC.CA]
[CEN.CO equals NDC.CO]

(2) CEN vs. TZ (12 articulations):

[CEN.Midwest disjoint TZ.Pacific]
[CEN.Midwest overlaps TZ.Eastern]
[CEN.Midwest overlaps TZ.Mountain]
[CEN.Northeast is_included_in TZ.Eastern]
[CEN.South disjoint TZ.Pacific]
[CEN.South overlaps TZ.Central]
[CEN.South overlaps TZ.Eastern]
[CEN.South overlaps TZ.Mountain]
[CEN.USA equals TZ.USA]
[CEN.West disjoint TZ.Central]
[CEN.West disjoint TZ.Eastern]
[CEN.West overlaps TZ.Mountain]

To demonstrate how we formulated the (2) CEN-TZ articulations, we compare the Time Zone map (Figure 4) directly with the Census Regions map (Figure 5). We started by looking at one region from TZ first, and came up with all the articulations about that region, then move on to the next. For example, we can start with TZ.Eastern, and it is very obvious that TZ.Eastern cannot geographically relate to CEN.West, therefore we can determine that CEN.West is disjoint with TZ.Eastern. Next, we can also see clearly from the two maps that the area for TZ.Eastern is bigger than that of CEN.Northeast, so we can make the assumption that the whole area of CEN.Northeast is_included_in TZ.Eastern. However, not all TZ.Eastern is an inclusion of CEN.South, only the southern part of TZ.Eastern is included in CEN.South; likewise, only a few parts around the Michigan Lake Area for TZ.Eastern is included in the CEN.Midwest. Thus, we can formulate our articulations such that CEN.South and CEN.Midwest overlaps with TZ.Eastern. In the same manner, we can formulate articulations for the Central, Mountain, and Pacific Time Zone areas.

Step 3: Iteratively edit articulations within Euler/X

Given taxonomies T_1, T_2, and articulations A, Euler/X infers from these all *possible worlds* W_1, W_2, …, W_N. Each world W_i represent a solution taxonomy T_3, simultaneously satisfying T_1, T_2, and A. If there are no possible worlds (N=0) (Figure 3) then the input TAP is inconsistent (i.e., a contradiction can be derived) and the user must repair, relax, or drop some articulations in A. Conversely, if A is underspecified and ambiguous, a large (exponential) number of worlds (N>1) may be derived. Usually, the objective is to find a unique world (N=1) that includes the input taxonomies, or a small number of worlds that reflect ambiguities that cannot be resolved (Franz et al., 2015).

When formulating articulations, we may overlook some articulations that the reasoning process requires in order to find only one or few suitable alignments. This problem will become more severe as our taxonomy alignment problems become larger. For example, in the CEN versus NDC taxonomy case, it is understandable that when we are writing out articulations, we may lose track and forget one or two states in the articulations. Or, when specifying the input for the CEN versus TZ example, it might not be obvious how to write out all 12 articulations needed to obtain one possible world, and we might only have, e.g., 11 at first. For example, when comparing Figures 4 and 5, it may not be obvious that there is a small area in Texas of TZ.Mountain that happens to be overlapping with CEN.South area. These missing articulations will result in underspecified input TAPs and thus in multiple possible worlds.

Euler/X can identify which sets of articulations are ambiguous and derive articulations that are logically implied, but not part of the given user input. This automatic 'ambiguity check' is not available when we do other types of aligning work such as absolute crosswalking or relative crosswalking. If the user underspecified an element in a relative crosswalking method, there will be no simple means to detect it and make it explicit. By iteratively checking the results Euler/X generated, and fixing the equivocal articulations, we can eventually get to the single solution (possible world) we are seeking.

RESULTS

The first part of the results is on the CEN-NDC case, while the second part describes the CEN-TZ case. All the files and figures we showed in this study is available in this repository: https://github.com/EulerProject/ASIST17.

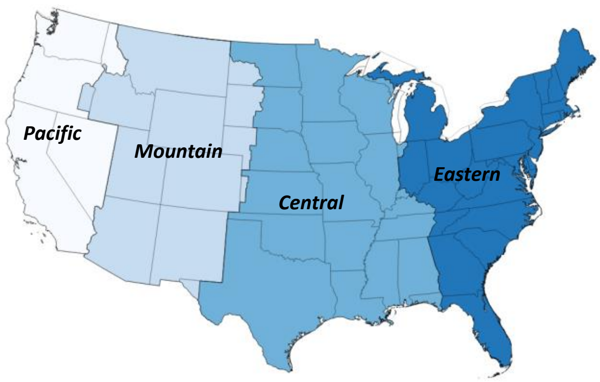

Figure 4. Time Zone map (TZ) for contiguous U.S

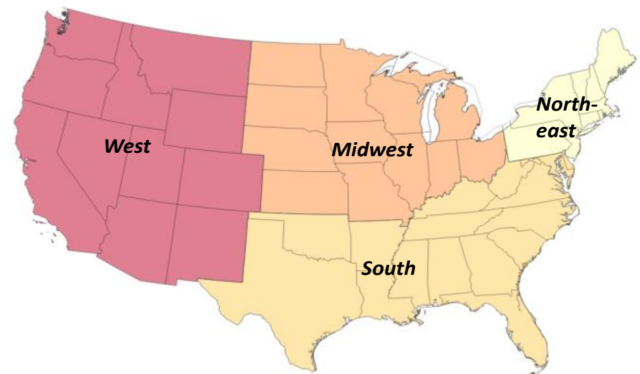

Figure 5. Census Regions (CEN) for contiguous UnitedStates

CASE 1: CEN VERSUS NDC

State-level alignments are all congruent

Case 1 aims to align the National Diversity Council taxonomy with the Census Regions taxonomy. This case serves as an example of the bottom-up approach of resolving taxonomy alignment problems. Corresponding to Figure 2, this use case entails the NDC taxonomy, CEN taxonomy and the 49 articulations that the users formulated as the input file into the Euler/X tool. Euler/X then can produce the input visualization shown in Figure 6 (next page). We can easily see that the 48 states and Washington D.C. from either side of the taxonomy are mapped with 'equal' signs. This is the input articulations that we put in, Euler/X has not yet inferred anything for us in this input visualization.

In Figure 7 (next page), the output is a single possible world, automatically generated and visualized by Euler/X. Euler/X depicts pairs of congruent concepts in the same grey rounded box. We can see that each of the states (and D.C.) from the two taxonomies are all in grey boxes, denoting that these lowest level (the children) alignments are congruent. In other words, both taxonomies agree at the lowest level (children) that the physical boundaries for each state are the same for either NDC or CEN. Interestingly, the two taxonomies also agree completely on the Midwest.

Inferred new articulations for regional-level alignments

On the upper-level classifications, we can see from the input visualization (Figure 6) that the big clusters of the CEN and NDC corresponding parents (CEN.West, South, Northeast, Midwest; NDC.West, Southeast, Southwest, Northeast, Midwest) are slightly off-center and do not align properly with each other like the states did. However, just from the input file we formulated on our own, it is not clear what the relationships between the upper-level classifications are, or what relationships the states have on upper-level concepts. Does NDC.Northeast equal CEN.Northeast? Does CEN.West include NDC.Southwest? The upper-level alignments were underspecified before Euler/X automatically inferred the relationships for us. The newly inferred articulations (Figure 7) made are depicted as red arrows, while the dotted lines tell us which concepts are overlapping.

The inferred articulations about the parents (CEN.West, South, Northeast, Midwest; NDC.West, Southeast, Southwest,Northeast, Midwest) are:

[CEN.USA equals NDC.USA]
[CEN.West includes NDC.West]
[CEN.South includes NDC.Southeast]
[CEN.South overlaps NDC.Southwest]
[CEN.South overlaps NDC.Northeast]
[CEN.Northeast is_included_in NDC.Northeast]
[CEN.Midwest equals NDC.Midwest]

The results explicitly constrain certain parent-level articulations that were not explicitly provided in the user's input articulations. In that sense, the reasoning has newly expressed information that we can now add to the original input to obtain a better-specified new input version.

In this case, we can also see our claim that using a RCC-5 logic-based approach in Euler/X allows us to preserve the original taxonomies in the merged view (Figure 7). This is a unique possible world, where the two taxonomies can both reside in and still maintain their own taxonomic names and hierarchical taxonomy structure.

Revisiting our introductory prompt about Tina and Amy and the half-smokes, the Euler/X-inferred graph (Figure 7) allows us to understand that although our speakers agreed on the lowest (state or city) level that they are congruent in both taxonomies, their disagreement is seen on a higher, regional-level. In Figure 7, marked in orange dotted boxes and bolded arrows, Washington D.C. is actually included in the South region in the CEN taxonomy, whereas in the NDC taxonomy, D.C. is included in the Northeast region.

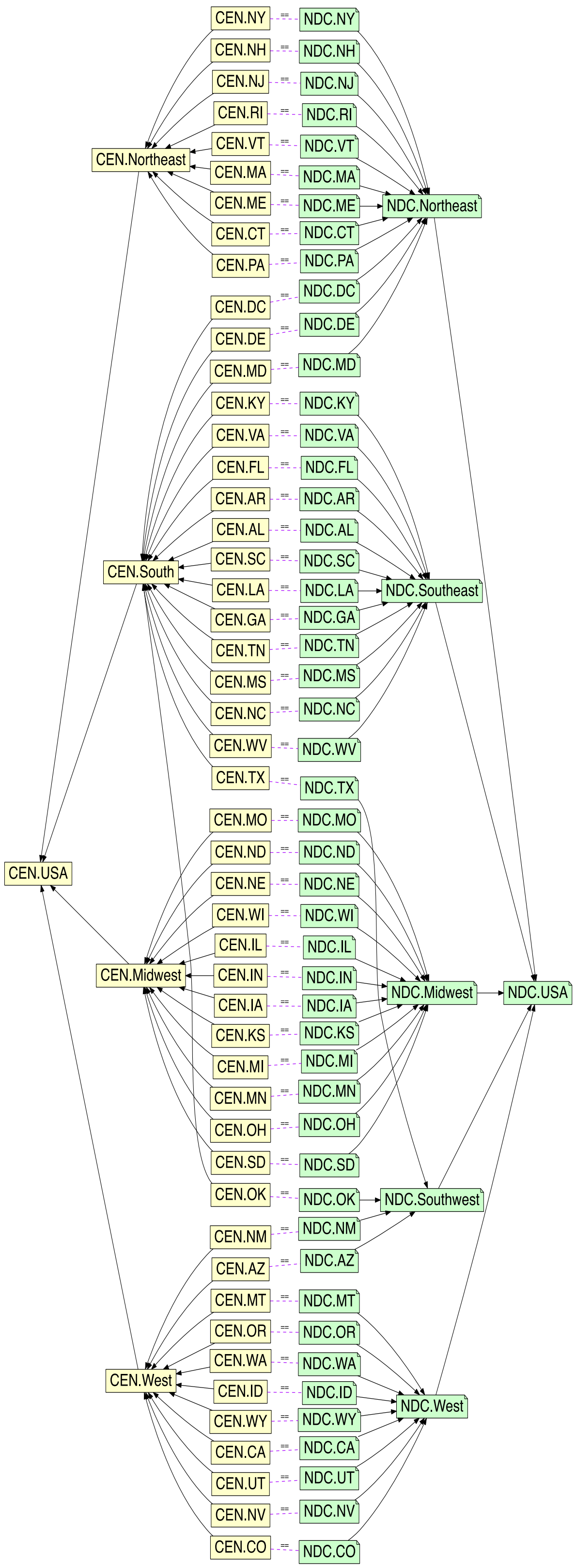

Figure 6. *CEN-NDC taxonomy alignment problem with 49 input articulations between T_{CEN} and T_{NDC}*

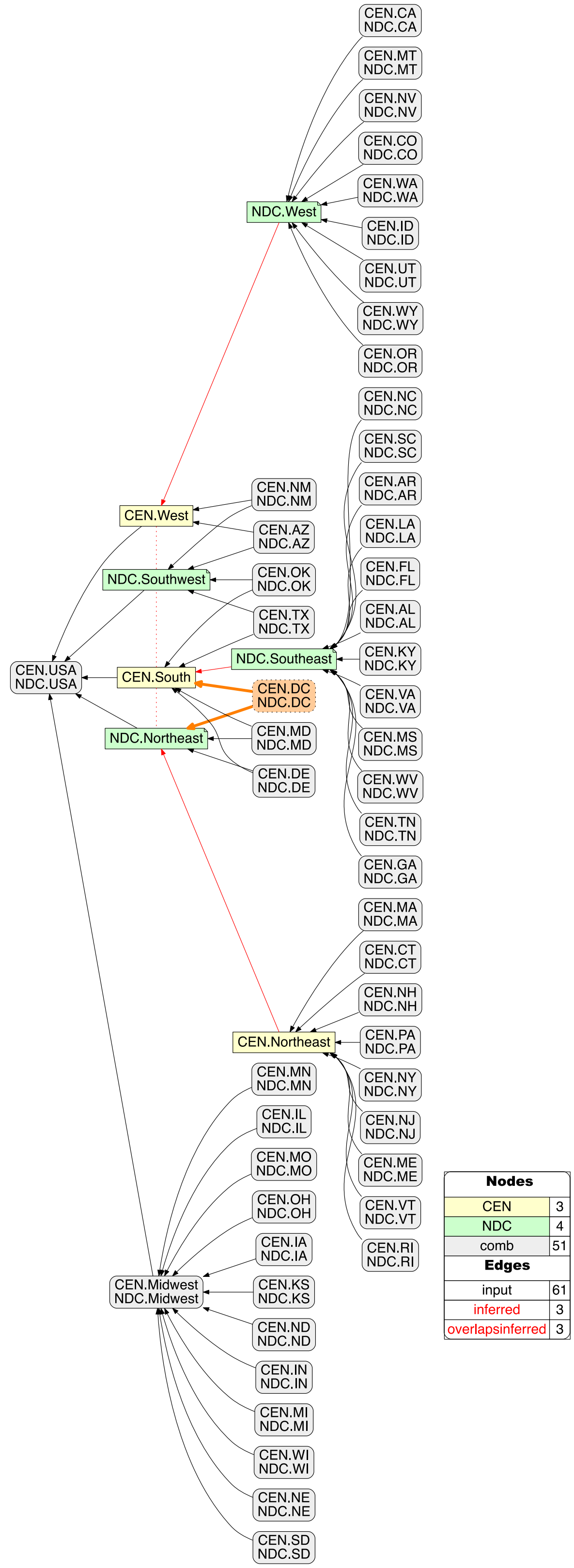

Figure 7. The unique possible world (PW) T_3 reconciling T_{CEN} and T_{NDC} via inferred relationships

Case 2: CEN versus TZ
Top-down regional alignment

The Census Regions taxonomy versus Time Zone case serves as a top-down example for taxonomy alignments. Since state boundaries are non-congruent with time zone boundaries, it is hard to model the articulations on the state-level alignments like we did with Case 1. For example, in Figure 4 where Mountain Time Zone and Central Time Zone overlap, we may be able to point out that the state on the very top is North Dakota, but we cannot really discern which counties of North Dakota are in the Mountain Time Zone and which other ones are in the Central Time Zone. If we want to model the situation as in Case 1 where all lowest-level alignments are equal, in this case, we may have to be more granular and model county-level articulations. This would work in theory, but it would be an onerous process since the contiguous (i.e., lower 48) US states consist of more than 3000 counties, requiring us to model as many county-level articulations.

An alternative way to represent the non-congruent parent regions is to provide the higher-level regional articulations directly to the Euler/X toolkit as input. The following input shows how we represent the corresponding two taxonomies with just 12 articulations:

taxonomy CEN Census_Regions
(**USA** Midwest South West Northeast)
taxonomy TZ Time_Zone
(**USA** Pacific Mountain Central Eastern)
articulations CEN TZ
[CEN.Midwest disjoint TZ.Pacific]
[CEN.Midwest overlaps TZ.Eastern]
[CEN.Midwest overlaps TZ.Mountain]
[CEN.Northeast is_included_in TZ.Eastern]
[CEN.South disjoint TZ.Pacific]
[CEN.South overlaps TZ.Central]
[CEN.South overlaps TZ.Eastern]
[CEN.South overlaps TZ.Mountain]
[CEN.USA equals TZ.USA]
[CEN.West disjoint TZ.Central]
[CEN.West disjoint TZ.Eastern]
[CEN.West overlaps TZ.Mountain]

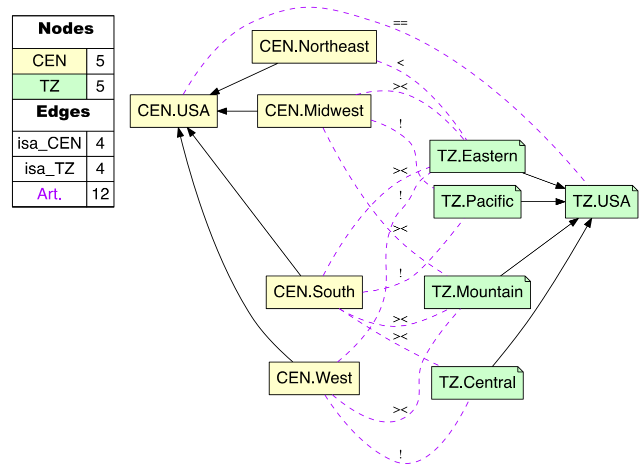

Figure 8*. Input visualization of the region-level input alignments between T_{CEN} and T_{TZ}*

Figure 8 shows the visualization for these input constraints, where none of the input regions show congruent (==) articulations. Every link is either overlapping, disjoint, proper inclusion.

The single possible world Euler/X generated for this alignment can be seen in Figure 9. This gives us a view in which Time Zone regions are aligned with the Census Regions but at the same time the original taxonomies are still preserved. We can see this by looking at the colored boxes in Figure 9: CEN concepts are in yellow boxes, while TZ concepts are in green clipped boxes. Euler/X also infers underspecified articulations in the input file; these are shown in the red arrows and dotted lines in Figure 9. For example, the Pacific Time Zone is properly included in the West Census Region, as shown by the red arrow pointing from TZ.Pacific to CEN.West.

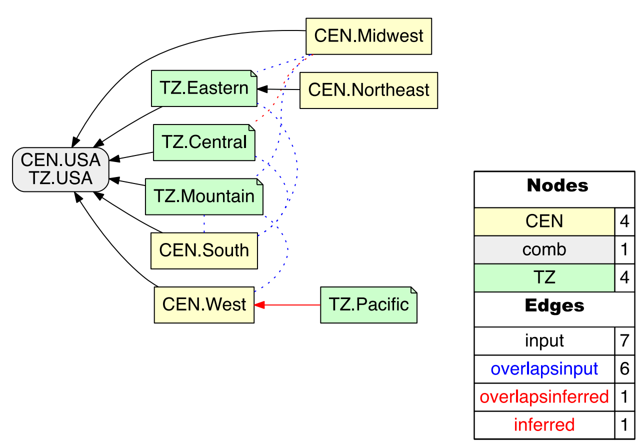

Figure 9. *The unique PW for the T_{CEN} with T_{TZ} alignment*

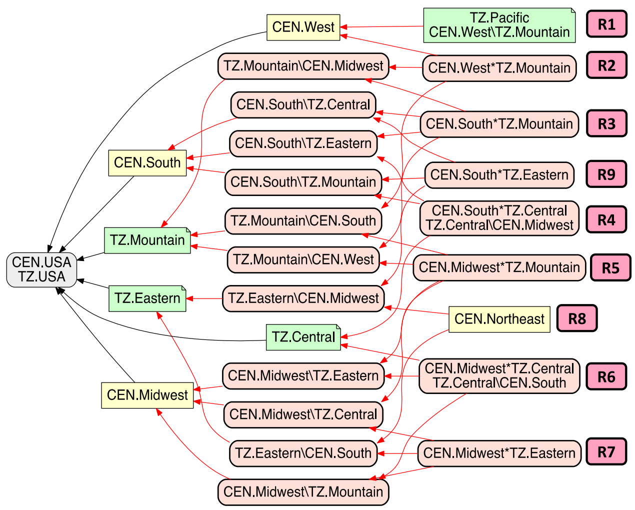

Figure 10. *Combined concepts solution for T_{CEN} and T_{TZ}*

Combined concepts solution for regional-level alignments

In the T_{CEN} and T_{TZ} case, since there are a lot of overlapping relationships (Figure 9), it seems even more necessary than in Case 1 (T_{CEN} and T_{NDC}) to represent which subregions overlap at a granular level. According to Euler/X's combined concepts solution (Figure 10), there are 18 new combined concepts regions (pink round-edges boxes) formed from our case. For instance, we can now tell that only the intersection of Census Regions West and Mountain Time Zone (denoted CEN.West*TZ.Mountain) can be counted as properly included in CEN.West. We can also infer from this graph that the intersection area of TZ.Central and CEN.Midwest (CEN.Midwest*TZ.Central) is congruent with the Central Time Zone area excluding the Census Region's South (TZ.Central \ CEN.South).

Comparing Euler solutions with a GIS solution

When overlaying the four regions in TZ and the five regions in CEN with a geographic information system (GIS), a new map with nine regions is created (Figure 11). These correspond to the nine leaf nodes produced by Euler/X (Figure 10). A total of 18 new concepts are generated by Euler/X to explain all containment and overlaps relationships in the result via concept *intersection A*B* and *difference A\B*.

By using only the merged GIS results, we can see what the new nine regions are, but there are no terms to describe each region except by creating new names. In Euler/X, when we resolve the output regions individually using the new labels, it tells us how regions are aligned – at the granular output region level – in order to form the 18 new regions. The Euler combined concept visualization also indicates that not all nine output regions are entirely new. The newly formed ones – those for which there is no adequate input region vocabulary – are the seven leaf-level regions in pink (Figure 10). We also labelled the corresponding nine regions of the GIS view from R1 to R9 in Figure 10.

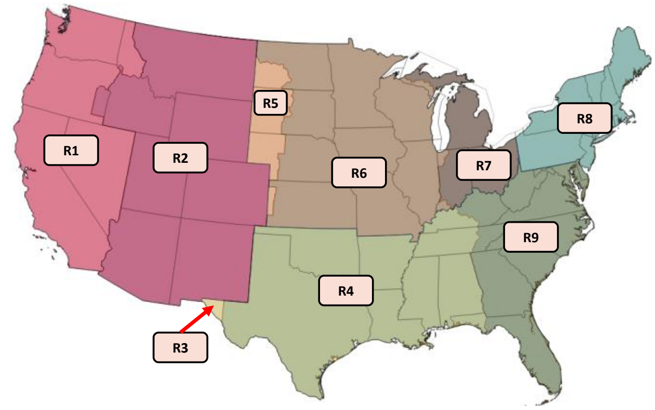

Figure 11. *A GIS overlay of CEN and TZ maps confirms the logically derived solution with 9 new regions*

Using the Euler/X results, we can name the new 18 regions based on our original taxonomies using *A*B* (again: "*" means *intersection* or AND), and *A\B* ("\"denotes *set difference* or NOT). For instance, we see that R1 is simply TZ.Pacific; alternatively we can say it is CEN.West without the TZ.Mountain area (CEN.West\TZ.Mountain). Similarly, we can describe R3 as the intersection between TZ.Mountain and CEN.South (TZ.Mountain*CEN.South).

In the special case of spatial taxonomies chosen for our study here, we are able to visualize how two maps relate to each other by overlaying them in a GIS. Here, the GIS merged results serve as a "ground truth" with which to independently assess the validity of the Euler/X results, and the former demonstrate clearly that the Euler/X alignments are an adequate method for taxonomy alignments.

However the more typical cases for taxonomy alignment require articulations between non-spatial taxonomies, i.e., for which no GIS route or direct visual cues about regional extensions are available. In general, the use of RCC-5 as an alignment vocabulary is a suitable approach to perform a wide range of multi-hierarchy reconciliations, whether these are spatial in an immediate sense or in a more abstract sense of aligning the extensions of different conceptual regions.

DISCUSSION

Based on the findings of this study, we hold that our logic-based taxonomy alignment approach can be used to solve a broad spectrum of crosswalking issues of concern to the information sciences community. If taxonomies are aligned not just based on their names or equivalent crosswalking, but based on the extent of congruence that each concept has with its counterparts, then we will be able to mitigate the membership condition problems that occur in equivalent crosswalking. By logically aligning each concept with its counterpart in another taxonomy, members of a superclass will have a comprehensive set of reasoner-validated relations to other concepts, instead of automatically inheriting the relationship from its superclass. In addition, our logic-based RCC-5 approach preserves the original taxonomies while providing an alignment view. Such alignments can solve data integration problems that happen in the more coarse-grained relative crosswalking, which otherwise is subjected to information loss. If future studies based off of our integrative alignments want to recover each unaltered input taxonomy, the semantics and structures in both taxonomies will be well-preserved and reproducible.

An implication is that our study also underscores the benefits of designing different alignment workflows, e.g., here *bottom-up* versus *top-down*, to match the needs of specific taxonomy alignment problems. The bottom-up approach seems to work well whenever we have non-overlapping relationships at the leaf-level (lowest-level) articulations, and we are not sure how the higher-level concepts should be aligned. The bottom-up alignment can help us to infer the underspecified articulations at the parent level(s). The top-down alignment approach, on the other hand, seems favorable when there is an expectation of certain higher-level articulations in conjunction with under-specified, complex, and often overlapping leaf-level relations. Expert input will frequently be needed to establish such expectations under the top-down approach.

CONCLUSIONS

In our study, we have explored the feasibility of using logic-based RCC-5 articulations to align different regional classifications of the United States. Our results demonstrate that this approach to taxonomy alignments is feasible and it suggests a promising way to enhance semantic interoperability.

In Case 1, we compared the National Diversity Council map with the Census Regions map in a bottom-up approach, in which 49 tip-level articulations (the 48 states and Washington D.C.) can be congruently asserted, not in the absolute cross-walking sense but using the logic-based RCC-5 relation. We call this case a bottom-up approach because Euler/X can leverage these tip-level articulations to unambiguously infer the higher, region-level articulations (e.g., that CEN.West includes NDC.West).

On the other hand, in Case 2, we aligned the U.S. Time Zone map with the Census Regions map in a top-down manner, realizing that it is laborious to use a bottom-up approach when different subsets of counties of one state participate in multiple time zones. We provided 12 articulations – none of them congruence – that show how each higher-level input region can be spatially reconciled with its respective counterparts. In this case, we were able to also visually represent alignments at the more granular level of the output regions generated by overlapping input regions. By contrasting Euler/X's alignments with the GIS results, we can see that Euler/X is able to identify nine new regions that result from partial input region overlap while at the same time still uses and newly combines the names the input regions to create unique output region labels (e.g. Region 9 is "TZ.Eastern*CEN.South"). This extended or hybrid vocabulary preserves the provenance of input region naming while generating new and more granular label sets needed to characterize each alignment region at the fine-grained level, whereas the GIS results provide no such provenance-aware region-identifying terminology for each output region.

To conclude, we have shown that the logic-based taxonomy alignments in RCC-5 can align distinct taxonomies to facilitate semantic interoperability, while preserving the internal coherence and context of each taxonomy. In future research, we envision investigating more sophisticated alignment cases in information science, such as aligning multiple metadata standards. We also envision that one day, in the face of conflicting views, we will be more willing to agree to disagree, and perhaps even redefine *synthesis* this way.

ACKNOWLEDGMENTS

Support of the authors' research through the National Science Foundation is kindly acknowledged (DEB-1155984, DBI-1342595, and DBI-1643002). The authors thank Professor Kathryn La Barre for her comments and suggestions. We would also like to thank Dr. Laetitia Navarro and Jeff Terstriep for help with creating map overlays in QGIS.

REFERENCES

Bachmair, L., Ganzinger, H., & Waldmann, U. (1993). Set constraints are the monadic class. *Logic in Computer Science*, 75–83. doi:10.1109/LICS.1993.287598.

Bodirsky, M., & Hils. (2012). Tractable set constraints. *J. Artif. Int. Res*. 45 (1): 731–759.

Bowker, G. C., & Star, S. L. (2000). Sorting things out: Classification and its consequences. MIT Press.

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

Chen, M., Yu, S., Kianmajd, P., Franz N., Bowers, S., and Ludäscher, B. (2014). *Provenance for explaining taxonomy alignments.* In *Provenance and annotation of data and processes*, 258–60. Springer, Cham. doi:10.1007/978-3-319-16462-5_27.

Cohn, A. G., & Renz, J. (2008). Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence*, 3, 551-596.

de Andrade, J., & Lopes Ginez de Lara, M. (2016). Interoperability and mapping between knowledge organization systems: Metathesaurus—unified medical language system of the National Library of Medicine. *Knowledge Organization, 43*(2).

Denecker, M., Lierler, Y., Truszczynski, M., & Vennekens, J. (2012). A Tarskian informal semantics for answer set programming. Intl. Conf. on Logic Programming (ICLP), 17:277–289. LIPIcs. Dagstuhl, Germany. doi:10.4230/LIPIcs.ICLP.2012.277.

Franz, N. M., Chen, M., Kianmajd, P., Yu, S., Bowers, S., Weakley, A. S., & Ludäscher, B. (2016a). Names are not good enough: Reasoning over taxonomic change in the Andropogon complex1. *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).doi:10.1371/journal.pone.0118247.

Franz, N. M., Pier, N. M., Reeder, D. M., Chen, M., Yu, S., Kianmajd, P. & Ludäscher, B. (2016b). Two influential primate classifications logically aligned. *Systematic Biology, 65*(4), 561-582.

Gebser, M., Kaminski, R., Kaufmann, B., & Schaub, T. (2012). Answer set solving in practice. *Synthesis Lectures on Artificial Intelligence and Machine Learning, 6*(3), 1-238.

Hodge, G. (2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files.* Washington, D.C.: Digital Library Federation, Council on Library and Information Resources.

Jung, J. J. (2006, November). Taxonomy alignment for interoperability between heterogeneous digital libraries. In International conference on asian digital libraries (p. 274-282). Berlin Heidelberg: Springer.

Kless, D., Milton, S., Kazmierczak, E., & Lindenthal, J. (2015). Thesaurus and ontology structure: Formal and pragmatic differences and similarities. *Journal of the Association for Information Science and Technology, 66*(7), 1348-1366.doi:10.1002/asi.23268.

Pfeifer, K., & Peukert, E. (2013, September). *Integration of text mining taxonomies.* In International joint conference on knowledge discovery, knowledge engineering, and knowledge management (p. 39-55). Berlin Heidelberg: Springer. doi:10.1007/978-3-662-46549-3_3.

Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. *Knowledge Representation and Reason*, *92*, 165-176.

Remsen, D. (2016). The use and limits of scientific names in biological informatics. *ZooKeys*, (550), 207.

Souza, R. R., Tudhope, D., & Almeida, M. B. (2012). Towards a taxonomy of KOS: Dimensions for classifying knowledge organization systems. *Knowledge Organization*, *39*(3), 179-192.

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

Thau, D., Bowers, S., & Ludäscher, B. (2008, October). Merging taxonomies under RCC-5 algebraic articulations. In Proceedings of the 2nd international workshop on ontologies and information systems for the semantic web (p. 47-54). ACM.

Zeng, M. L., & Chan, L. M. (2009). Semantic interoperability. Encyclopedia of Library and Information Sciences, 3rd Edition, 4645–62. CRC Press.