

© 2022 Yi-Yun Cheng

AGREEING TO DISAGREE:  
APPLYING A LOGIC-BASED APPROACH TO RECONCILING AND MERGING  
MULTIPLE TAXONOMIES

BY

YI-YUN CHENG

DISSERTATION

Submitted in partial fulfillment of the requirements  
for the degree of Doctor of Philosophy in Information Sciences  
in the Graduate College of the  
University of Illinois Urbana-Champaign, 2022

Urbana, Illinois

Doctoral Committee:

Professor Bertram Ludäscher, Chair  
Professor Allen Renear  
Assistant Professor Karen Wickett  
Professor Nico Franz, Arizona State University

# ABSTRACT

Taxonomies are used to classify concepts into hierarchies via parent-child (is-a) relationships. They are used in many contexts, ranging from everyday sorting decisions, enterprise website infrastructures, to controlled vocabularies in information systems. The proliferation of taxonomies leads to interoperability challenges when one attempts to incorporate existing taxonomies with a newly created taxonomy, or to integrate multiple data in a digital library that were annotated with different taxonomies.

To address these interoperability challenges, methods to systematically align two taxonomies and merge them into a single merged view have been developed in the extant literature. However, merging taxonomies into a unified representation may result in the loss of important information that was present in the original taxonomies. The goals of this dissertation are thus to align taxonomies that: (1) preserve the information in both taxonomies in the merged solution(s); (2) provide multiple possible solutions to merge the taxonomies.

Specifically, this dissertation explores the use of a logic-based approach to align taxonomies in which concepts in two taxonomies  $T_1$ ,  $T_2$  are inter-linked via a set of relations to yield merged solution(s). The merged solution(s) can be (1) a unique merged taxonomy  $T_3$  that preserves both  $T_1$  and  $T_2$ 's information; (2) an inconsistent result that suggests the relations linking the taxonomies are contradictory; or (3) multiple merged solutions that present different possible ways (or *possible worlds*) in which  $T_1$  and  $T_2$  can be aligned. In this dissertation, I apply this logic-based alignment approach to taxonomies in socio-geographic contexts, including the United States maps, country taxonomies, indigenous peoples' tribes, and historical sovereignties in biodiversity data. Through these use cases, it is demonstrated that the logic-based taxonomy alignment approach is feasible to reconcile conflicts in socio-geographic taxonomies.

*To my family Aili, Willy, Hubert, Chengmei, and Ly.*

# ACKNOWLEDGMENTS

Six years ago, I flew 6,000 miles away from Taipei to Illinois to begin my doctoral studies. I was accustomed to skyscrapers, crowds, and endless traffic back at home, so when endless prairies and cornfields unfolded in front of my eyes on I-57, I was perplexed by this unfamiliarity. Nervous yet excited, I stepped foot into the East Foyer of 501 E Daniel St.

When two worlds collide – was my first impression on interacting with my advisor, Dr.Bertram Ludäscher. We came from very different academic backgrounds. Our weekly meetings sometimes went on for more than two hours, when he tried to explain the logic axioms, work his magic with programming codes in Emacs, and squiggle euler diagrams with blue markers on the whiteboard in his NCSA office; I, at the same time, exchanged all that I know on metadata, knowledge organization systems, and all the wonders in LIS. I am grateful for the tireless hours of meetings we held, the inspiration for ideas and this dissertation, and the guidance throughout this journey.

Over the course of six years, I have learned many things from many of my academic mentors and collaborators: Dr.Allen Renear on his dean-ly advice on navigating the academia; Dr.Jodi Schneider on her support for professional growth; Dr.Nico Franz on the beauties of weevils and taxon concepts; Dr.Jim Whitfield on systematics and biodiversity informatics; Dr.Jane Greenberg on leadership in LIS; my committee members on guidance on this doctoral dissertation; and members of the conceptual foundations group on conceptual modeling. I am grateful for all the intellectual conversations that have shaped my scholarship.

The first person I met at 501 E Daniel turned out to be my best companion of this doctoral journey, the newly minted Dr.Ly Dinh. We shared the same excitement as first-years, the same self-doubts during each doctoral milestones, and the same ups-and-downs whenever we got a paper accepted or rejected. I am grateful for all our scholarly discussions,

the roaring laughter from our travels, and our persistence to have our sweats shed at group fitness classes everyday for the past years.

The School of Information Sciences at University of Illinois Urbana-Champaign has then become my second home from home. I have met wonderful humans, colleagues, and peers here, and I could not have completed this doctoral study without them: Linh Hoang, Nikolaus Parulian, Dr.Rhiannon Bettivia, Dr.Michael Gryk, Dr.Jacob Jett, Dr.Tim McPhillips, Dr.Stephen Downie, Jean Kanengoni, Margaret Buck, Janet Eke, MT Campbell, members of CIRSS, members of iDAKS, iSchool staffs, and of course, our academic parent, Dr.Linda Smith. I am grateful for the friendliness, hospitality, and support from our iSchool community.

It's definitely not easy to be so far away from home – but I know my family and friends always have my back. My parents, Aili and Willy, and my brother Hubert will probably never read this dissertation, but I am grateful for their unconditional love and support of my decision to leave home for this degree. My dear grandparents, Aju and Hsi-Chou, always bragged to their friends about their granddaughter doing a PhD. Though they did not have the chance to see the day I get hooded, I know they would be forever proud of me for reaching the finishing line. Friends who have been in the same boat with me – Ly, Phuc Huynh, Yvette Tien, Xiaomei Li, Esther Hsiao, Cassandra Huang, Wan-Chen Lee, I am grateful for the constant reminder to still lead a enjoyable life outside of work. I would also like to extend my gratitude to my mentor Dr. Sherry Chen from National Taiwan University, who inspired me to begin this academic journey in the first place.

6,000 miles, six years, and numerous restless hours are what it takes to finish this doctoral degree. I embarked on this journey with bewilderment and excitement. At the end of this degree, I still am bewildered by the boundless knowledge out there; and excited about the next phases of our lives. I am not sure if I will ever be accustomed to the quietness of the prairies and the cornfields, but the vastness of space and knowledge has taught me to always stay humble, and always be kind.

# TABLE OF CONTENTS

|                                                                          |      |
|--------------------------------------------------------------------------|------|
| LIST OF TABLES . . . . .                                                 | viii |
| LIST OF FIGURES . . . . .                                                | ix   |
| CHAPTER 1 INTRODUCTION . . . . .                                         | 1    |
| 1.1 Motivating Examples . . . . .                                        | 2    |
| 1.2 Dissertation Background and Problem Statement . . . . .              | 7    |
| 1.3 Goals of the Dissertation . . . . .                                  | 11   |
| 1.4 Research Questions . . . . .                                         | 12   |
| 1.5 Organization of Chapters . . . . .                                   | 13   |
| CHAPTER 2 PRELIMINARIES . . . . .                                        | 17   |
| 2.1 Definition: Taxonomy . . . . .                                       | 17   |
| 2.2 Definition: Region Connection Calculus (RCC) . . . . .               | 20   |
| 2.3 Definition: Taxonomy Alignment Problems (TAP) . . . . .              | 21   |
| 2.4 Chapter Summary . . . . .                                            | 22   |
| CHAPTER 3 TAXONOMY ALIGNMENT METHODS: A REVIEW . . . . .                 | 23   |
| 3.1 Knowledge Organization Systems, Taxonomies, and Ontologies . . . . . | 23   |
| 3.2 Myriads of Alignment Methods . . . . .                               | 25   |
| 3.3 Level of KOS examined . . . . .                                      | 26   |
| 3.4 Identifying Key Papers . . . . .                                     | 27   |
| 3.5 Definitions of a Taxonomy . . . . .                                  | 30   |
| 3.6 Characteristics of Taxonomy Alignment Approaches . . . . .           | 32   |
| 3.7 Chapter Summary . . . . .                                            | 39   |
| CHAPTER 4 TAXONOMY ALIGNMENT PROBLEM . . . . .                           | 41   |
| 4.1 Overview . . . . .                                                   | 41   |
| 4.2 Articulations . . . . .                                              | 42   |
| 4.3 Region Connection Calculus (RCC) . . . . .                           | 43   |
| 4.4 Possible Worlds . . . . .                                            | 46   |
| 4.5 Modeling and Interpretations . . . . .                               | 47   |
| 4.6 Euler/X and LeanEuler . . . . .                                      | 47   |
| 4.7 A Simple Example of TAP: “ Magnificent 7 ” . . . . .                 | 49   |
| 4.8 Chapter Summary . . . . .                                            | 50   |

|                                                                                                   |     |
|---------------------------------------------------------------------------------------------------|-----|
| CHAPTER 5 RECONCILING THE SOCIO-GEOGRAPHIC IN THE UNITED STATES MAPS . . . . .                    | 51  |
| 5.1 Introduction . . . . .                                                                        | 51  |
| 5.2 Related Work . . . . .                                                                        | 54  |
| 5.3 Taxonomy Alignment Problems . . . . .                                                         | 55  |
| 5.4 Results . . . . .                                                                             | 59  |
| 5.5 Chapter Summary . . . . .                                                                     | 69  |
| CHAPTER 6 RECONCILING THE SOCIO-GEOGRAPHIC IN COUNTRY TAXONOMIES . . . . .                        | 71  |
| 6.1 Introduction . . . . .                                                                        | 71  |
| 6.2 Related Work . . . . .                                                                        | 73  |
| 6.3 Taxonomy Alignment Problems . . . . .                                                         | 76  |
| 6.4 Results . . . . .                                                                             | 79  |
| 6.5 Chapter Summary . . . . .                                                                     | 89  |
| CHAPTER 7 RECONCILING THE SOCIO-GEOGRAPHIC IN INDIGENOUS PEOPLES' TRIBES . . . . .                | 92  |
| 7.1 Introduction . . . . .                                                                        | 92  |
| 7.2 Related Work . . . . .                                                                        | 94  |
| 7.3 Taxonomy Alignment Problems . . . . .                                                         | 99  |
| 7.4 Results . . . . .                                                                             | 107 |
| 7.5 Chapter Summary . . . . .                                                                     | 122 |
| CHAPTER 8 RECONCILING THE SOCIO-GEOGRAPHIC IN BIODIVERSITY DATA AND INFORMATION SYSTEMS . . . . . | 126 |
| 8.1 Introduction . . . . .                                                                        | 127 |
| 8.2 Related Work . . . . .                                                                        | 132 |
| 8.3 Proposed Conceptual Model . . . . .                                                           | 137 |
| 8.4 Use Case . . . . .                                                                            | 143 |
| 8.5 Modeling Challenges . . . . .                                                                 | 159 |
| 8.6 Chapter Summary . . . . .                                                                     | 163 |
| CHAPTER 9 CONCLUSION . . . . .                                                                    | 165 |
| 9.1 Dissertation Findings . . . . .                                                               | 165 |
| 9.2 Contributions . . . . .                                                                       | 169 |
| 9.3 Future Work . . . . .                                                                         | 172 |
| REFERENCES . . . . .                                                                              | 174 |
| APPENDIX A . . . . .                                                                              | 186 |

# LIST OF TABLES

|     |                                                                              |     |
|-----|------------------------------------------------------------------------------|-----|
| 3.1 | Characteristics of Taxonomy Alignment Approaches -1 . . . . .                | 33  |
| 3.2 | Characteristics of Taxonomy Alignment Approaches -2 . . . . .                | 34  |
| 3.3 | Characteristics of Taxonomy Alignment Approaches -3 . . . . .                | 34  |
| 4.1 | RCC-5 composition table . . . . .                                            | 43  |
| 4.2 | RCC-5 Relations to Euler/X notations and Meanings . . . . .                  | 46  |
| 4.3 | Maximally Informative Relations (MIR) table for the Magnificent 7 example    | 49  |
| 6.1 | Comparison of $T_{ISO}$ , $T_{DHS}$ , and $T_{CIA}$ . . . . .                | 78  |
| 7.1 | Taiwan's Recognized tribes over time . . . . .                               | 97  |
| 7.2 | Taiwan indigenous peoples population dataset . . . . .                       | 103 |
| 7.3 | Taiwan indigenous peoples re-scaled population . . . . .                     | 107 |
| 8.1 | Abridged species occurrence dataset of <i>Platalea minor</i> . . . . .       | 144 |
| 8.2 | Number of occurrences per geographic entity . . . . .                        | 145 |
| 8.3 | Modern day Japan . . . . .                                                   | 146 |
| 8.4 | Modern day Vietnam . . . . .                                                 | 146 |
| 8.5 | East Asia Sovereignities from 1880 to 1950 . . . . .                         | 148 |
| 8.6 | East Asia Sovereignities from 1960 to 2022 . . . . .                         | 149 |
| 8.7 | 13 Alignments and the corresponding counts of possible worlds in each type . | 156 |

# LIST OF FIGURES

|      |                                                                                  |    |
|------|----------------------------------------------------------------------------------|----|
| 1.1  | The U.S. Department of the Interior Regions. . . . .                             | 2  |
| 1.2  | The Fish and Wildlife Service Regions, as of April 1st, 2022. . . . .            | 3  |
| 1.3  | The National Park Service Regions from 2003. . . . .                             | 4  |
| 1.4  | COVID-19 tracking dashboard created by different organizations . . . . .         | 5  |
| 2.1  | Example when a taxonomy is a tree. . . . .                                       | 19 |
| 2.2  | Example when a taxonomy is a DAG. . . . .                                        | 20 |
| 4.1  | The region connection relations used in RCC-5 . . . . .                          | 43 |
| 4.2  | RCC-5 composition Euler diagrams for <b>EQ</b> , <b>DR</b> , <b>PO</b> . . . . . | 44 |
| 4.3  | RCC-5 composition Euler diagrams for <b>PP</b> , <b>PPi</b> . . . . .            | 45 |
| 4.4  | The alignment lifecycle of Euler/X . . . . .                                     | 48 |
| 4.5  | A simple taxonomy alignment problem with $T_1$ , $T_2$ and 7 possible worlds . . | 50 |
| 5.1  | Different geo-taxonomies of the contiguous United States . . . . .               | 53 |
| 5.2  | Time Zone map (TZ) for contiguous U.S. . . . .                                   | 56 |
| 5.3  | Taxonomy alignment problem $T_{CEN}$ and $T_{NDC}$ . . . . .                     | 60 |
| 5.4  | One possible world (PW) $T_3$ reconciling $T_{CEN}$ and $T_{NDC}$ . . . . .      | 61 |
| 5.5  | Region-level input alignments between $T_{CEN}$ and $T_{TZ}$ . . . . .           | 66 |
| 5.6  | One PW for the $T_{CEN}$ with $T_{TZ}$ alignment . . . . .                       | 66 |
| 5.7  | Combined concepts solution for $T_{CEN}$ and $T_{TZ}$ . . . . .                  | 67 |
| 5.8  | A GIS overlay of CEN and TZ maps . . . . .                                       | 67 |
| 6.1  | Major U.S. Airline web interface. . . . .                                        | 75 |
| 6.2  | ISO's presupposed hierarchical structure for the entities . . . . .              | 76 |
| 6.3  | Input alignment for $T_{ISO}-T_{DHS}$ . . . . .                                  | 80 |
| 6.4  | Abridged input alignment for $T_{ISO}-T_{DHS}$ . . . . .                         | 82 |
| 6.5  | One possible world for $T_{ISO}-T_{DHS}$ (full) . . . . .                        | 83 |
| 6.6  | One possible world for $T_{ISO}-T_{DHS}$ (abridged) . . . . .                    | 84 |
| 6.7  | Input alignment for $T_{ISO}-T_{CIA}$ . . . . .                                  | 85 |
| 6.8  | One possible world for $T_{ISO}-T_{CIA}$ . . . . .                               | 86 |
| 6.9  | Input alignment for $T_{CIA}-T_{DHS}$ . . . . .                                  | 86 |
| 6.10 | Seven possible worlds for $T_{CIA}-T_{DHS}$ . . . . .                            | 87 |
| 6.11 | Combined concept solution for $T_{CIA}-T_{DHS}$ . . . . .                        | 88 |
| 7.1  | Taiwan's Indigenous electoral constituencies map . . . . .                       | 98 |

|      |                                                                                       |     |
|------|---------------------------------------------------------------------------------------|-----|
| 7.2  | Taiwan’s current 16 recognized tribes and their distribution map . . . . .            | 99  |
| 7.3  | Input alignment for $T_1$ .LH and $T_2$ .9 tribes . . . . .                           | 108 |
| 7.4  | One PW for $T_1$ .LH and $T_2$ .9 tribes . . . . .                                    | 109 |
| 7.5  | Input alignment for $T_1$ .LH and $T_3$ .16 tribes . . . . .                          | 111 |
| 7.6  | One PW for $T_1$ .LH and $T_3$ .16 tribes. . . . .                                    | 113 |
| 7.7  | GIS map overlaying solution for $T_1$ .LH and $T_3$ .16 tribes . . . . .              | 114 |
| 7.8  | Input visualizations comparison based on two different thresholds . . . . .           | 115 |
| 7.9  | PW comparison based on two different thresholds . . . . .                             | 116 |
| 7.10 | Input alignment for $T_1$ .LH and $T_4$ .26 tribes . . . . .                          | 118 |
| 7.11 | One PW for $T_1$ .LH and $T_4$ .26 tribes . . . . .                                   | 119 |
| 7.12 | Input alignment for $T_1$ and $T_5$ . . . . .                                         | 121 |
| 7.13 | One PW for the re-scaled modeling TAP $T_1$ and $T_5$ . . . . .                       | 122 |
| 8.1  | GBIF’s interface of the occurrence records of <i>Platalea minor</i> . . . . .         | 127 |
| 8.2  | A table in historical text comparing different occurrences of <i>Platalea Minor</i> . | 128 |
| 8.3  | Possible sightings and occurrences of <i>Platalea Minor</i> in the world . . . . .    | 129 |
| 8.4  | Proposed conceptual model for aligning historical sovereignties taxonomies .          | 138 |
| 8.5  | The taxonomy of sovereignty and geographic entities in a given year . . . . .         | 141 |
| 8.6  | Input alignment for $T_{Current}$ and $T_{Past}$ . . . . .                            | 142 |
| 8.7  | Four types of possible worlds . . . . .                                               | 143 |
| 8.8  | The 14 taxonomies constructed in this study . . . . .                                 | 150 |
| 8.9  | East Asia in 2022 . . . . .                                                           | 150 |
| 8.10 | East Asia in 1940 . . . . .                                                           | 150 |
| 8.11 | 13 input alignments for $T_C$ and $T_P$ . . . . .                                     | 153 |
| 8.12 | One of the 13 alignments: $T_{2022}$ and $T_{1940}$ . . . . .                         | 154 |
| 8.13 | Illustration of a simplified example possible world . . . . .                         | 156 |
| 8.14 | Example of two Type I possible worlds for $T_{2022}$ and $T_{1940}$ . . . . .         | 158 |

# CHAPTER 1

## INTRODUCTION

*“This book first arose out of a passage in Borges...in which it is written that ‘animals are divided into: (a) belonging to the Emperor, (b) embalmed, (c) tame, (d) suckling pigs, (e) sirens, (f) fabulous, (g) stray dogs, (h) included in the present classification, (i) frenzied, (j) innumerable, (k) drawn with a very fine camelhair brush, (l) et cetera, (m) having just broken the water pitcher, (n) that from a long way off look like flies’. In the wonderment of this taxonomy, the thing we apprehend in one great leap, the thing that, by means of the fable, is demonstrated as the exotic charm of another system of thought, is the limitation of our own, the stark impossibility of thinking that.”*

– Michel Foucault, *The Order of Things*, Preface

## 1.1 Motivating Examples

### 1.1.1 The U.S. Department of the Interior Region Reorganization

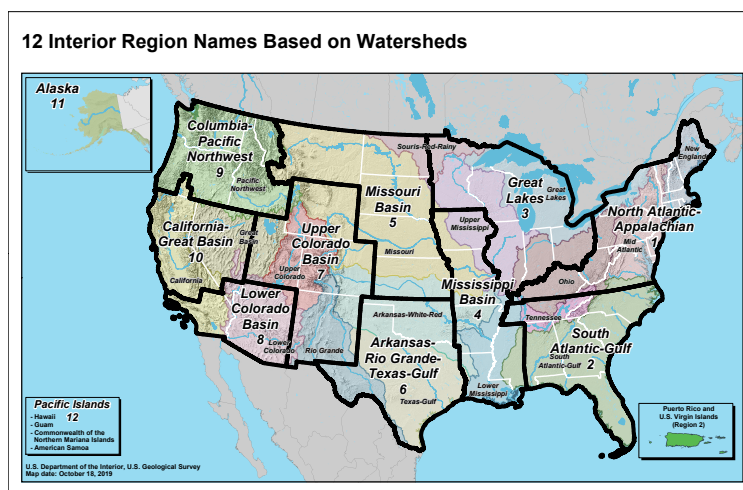


Figure 1.1: The U.S. Department of the Interior Regions.

The United States Department of the Interior (DOI) announced that its reorganization of regions of the United States is complete on August 22, 2018<sup>1</sup>. The now 12 unified regional boundaries, drawn by the U.S. Geological Survey, are roughly based on watersheds. The regions are assigned numeric values (e.g., Region 1), along with new regional names (e.g., North Atlantic-Appalachian), as shown in Figure 1.1. The DOI also provided a crosswalk<sup>2</sup> to map the new DOI regions to each of the bureau offices' original regions.

<sup>1</sup>DOI announcement on the unified interior regional boundaries:<https://www.doi.gov/employees/reorg/unified-regional-boundaries>

<sup>2</sup>DOI crosswalk: <https://www.doi.gov/sites/doi.gov/files/uploads/doi-regions-bureau-executive-assignments.pdf>

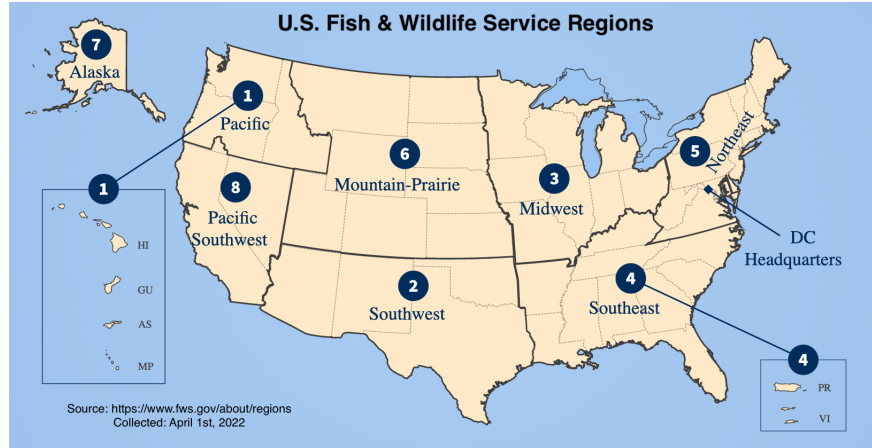


Figure 1.2: The Fish and Wildlife Service Regions, as of April 1st, 2022.

The Fish and Wildlife Service (FWS) is one of the bureaus within the Department of the Interior. As of the time this dissertation was written, the FWS still has its own division of eight regions<sup>3</sup>, with another set of numeric number assigned to each region (Figure 1.2). The crosswalk provided by the DOI attempted to align the eight FWS offices with the new regions, but there are still unresolved discrepancies between the two taxonomies. For example, the DOI cross-walked its Region 3 to the *FWS Midwest Bloomington, Minnesota office* with the state coverage of Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin. However, on the FWS regions, the Midwest comprises not only those six states, but also Iowa and Missouri. Iowa and Missouri are under the jurisdiction of DOI's Region 4 Mississippi Basin area, which the DOI roughly corresponds to the *FWS's Atlanta, Georgia (Southeast) office*.

<sup>3</sup>FWS regions: <https://www.fws.gov/about/regions>



Figure 1.3: The National Park Service Regions from 2003.

The National Park Service (NPS) is another bureau within the DOI. The NPS is actually making an effort to align its organizational structure<sup>4</sup> with the DOI's new numeric regions and crosswalk. However, given that the DOI splits California and Nevada each into two parts (Northern and Southern) and lists them in different regions, the NPS assign Region 8 in two different groupings, creating confusion of which region have jurisdiction over California and Nevada. Moreover, NPS's six regions still very much resemble its 2003 classification of the U.S. regions<sup>5</sup> (Figure 1.3).

The U.S. DOI's region reorganization example is not a singular case. People make arbitrary decisions on how to group items, places, or persons in relation to geographic locations all the time. Different organizations can have different ways to organize their regions; and it is often challenging to reconcile those differences. In this case, an overarching office struggles to align multiple taxonomies from their branch offices, raising more confusion than perhaps, before the reorganization of these regions.

<sup>4</sup>NPS organizational structure:<https://www.nps.gov/aboutus/organizational-structure.htm>

<sup>5</sup>Source:[https://www.nps.gov/gis/documents/nps\\_regions\\_11x8-5-new.pdf](https://www.nps.gov/gis/documents/nps_regions_11x8-5-new.pdf)

## 1.1.2 The (missing) COVID-19 regions

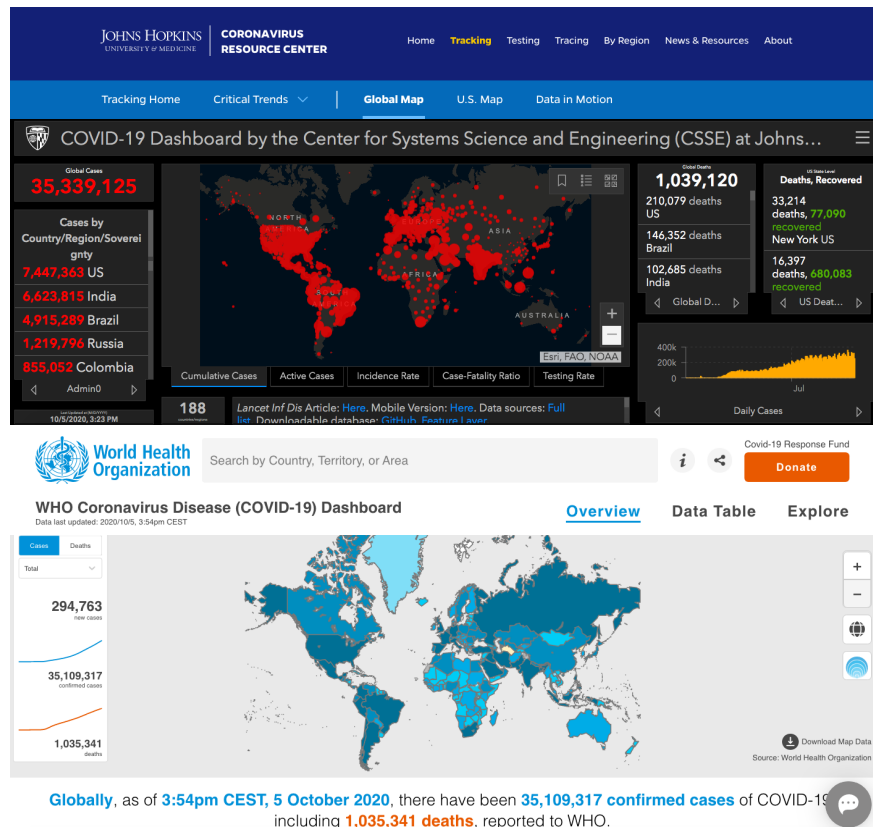


Figure 1.4: COVID-19 tracking dashboard created by different organizations  
Johns Hopkins Dashboard (top). World Health Organization Dashboard (bottom).

Sometimes, differences in taxonomies may be subtle to detect. The top figure in Figure 1.4 is the dashboard from Johns Hopkins University (JHU) to document COVID-19 cases <sup>67</sup>, and the bottom figure is from the World Health Organization (WHO) <sup>8</sup>. At first glance, the differences between the two dashboards are noticeable with the colors on the two map visualizations. However, it is difficult to identify right away the differences between the data sources used by the two institutions. When further examining the datasets in detail, there are actually four locations listed in the JHU dataset but are missing from the WHO dataset

<sup>6</sup>JHU dashboard: <https://coronavirus.jhu.edu/map.html>

<sup>7</sup>JHU dataset: <https://github.com/CSSEGISandData/COVID-19>

<sup>8</sup>WHO: <https://covid19.who.int/>

(as of May 5th, 2020). These countries/regions are *Taiwan*, *West Bank and Gaza*, *Kosovo*, and *Western Sahara*.

This example shows that the two organizations have used different country taxonomies to devise their datasets. JHU is one of the first teams to provide a real-time COVID-19 cases tracking dashboard, and their dataset is aggregated via sources from all over the world – possibly the reason that it is showing more diverse locations than that of the WHO’s dataset. On the other hand, the WHO, being part of the United Nations (UN), possibly only lists the countries that are recognized by the UN. Taxonomies such as the ones presented in this example can have their implicit assumptions (e.g., if they are recognized by certain organizations). If we do not recognize these assumptions, certain perspective can overwhelm other valid perspectives. In this case, we may overlook some areas in this world that have been resilient towards the COVID-19 outbreak.

Socio-geographic taxonomies are ubiquitous. From everyday decisions such as purchasing properties based on the fact whether the houses are in flood zones<sup>9</sup>, to redlining practices in urban planning that are susceptible to segregating neighborhoods into the rich and the poor (Plumer, Popovich, & Renault, 2020), to the invisible boundaries of Downtown and the South Side Chicago that deters different groups of Chicagoans to “cross the border” to the other side (Enos, 2017) – many situations such as these concern groupings of people and society in a hierarchical order in relation to space, geographic location, or maps.

The two examples in this section suggest that there is an apparent need to resolve the differences in these taxonomies to “agree to disagree”. Multiple taxonomies can be used to classify the same subject matters; and assumptions embedded within these taxonomies may obstruct a more comprehensive view of the subject matter, or even lead to missing information in datasets. The reconciliation process of these conflicts can amplify the confusion among the taxonomies (e.g., the DOI’s crosswalk), or it can potentially resolve the conflicts and make the implicit assumptions salient (e.g., some other alignment approaches we will

---

<sup>9</sup>Flood maps:<https://www.fema.gov/flood-maps>

mention and use in this dissertation).

## 1.2 Dissertation Background and Problem Statement

### 1.2.1 Taxonomy and Taxonomy Alignment

***Taxonomy*** is a scheme that classifies concepts into hierarchies via parent-child (is-a) relationships (Giabelli, Malandri, Mercurio, & Mezzanzanica, 2022; Reynaud & Safar, 2006). For instance, in organizing closets, we can sort the clothes into outerwear, tops, and bottoms. For outerwear, we can further sort it into jackets, coats, and ponchos. For tops, we can organize it into sweaters, shirts, and blouses. For bottoms, we can have jeans, shorts and pants. In organizing clothes into categories and subcategories, we are forming a hierarchical scheme – a *taxonomy*.

Taxonomies are used in many contexts, ranging from everyday sorting decisions (Hedden, 2016), enterprise website infrastructure (Angermann, Pervez, & Ramzan, 2017), phylogeny (Franz et al., 2015), to classification schemes or ontologies in computer science, information sciences, or knowledge organization communities (Olson, 2002). Specifically, taxonomies are one type of knowledge organization systems (KOS) for organizing information. Different types of KOS include term lists (e.g., dictionaries, glossaries), classification schemes (e.g., subject headings, classification systems, taxonomies), and relationship models (e.g., thesauri, ontologies) (Hodge, 2000; Zeng, 2008).

The need for reconciliation of taxonomies arises when multiple taxonomies about the same topic can disagree with each other; and even one single taxonomy can disagree with itself over time (Franz, Chen, et al., 2016; Franz, Pier, et al., 2016). To enhance interoperability between two vocabularies of one type of KOS, generic models such as derivation and expansion have been discussed in great detail (Chan & Zeng, 2006; Zeng & Chan, 2010). More specific mapping methods for a particular type of KOS have also been explored, such

as ontology matching methods (Otero-Cerdeira, Rodríguez-Martínez, & Gómez-Rodríguez, 2015). Metadata crosswalks are one of the major means to map one metadata standard to another, such as 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 (of Congress, 2008).

Alignment approaches specifically for taxonomies are also abundant: Jung (2006) computes word similarities between concepts in different digital library taxonomies to come up with candidate pairs of words that are similar to each other; Pfeifer and Peukert (2013a) employ a schema-based and instance-based similarity matcher that give results to the mappings between two taxonomies and the confidence scores between pairs of concepts. In systematically reviewing 28 taxonomy alignment relevant papers, I have categorized taxonomy alignment approaches into logic-based, constraint-based, or NLP-based (Natural Language Processing) methods, discussing each approaches’ matching dimensions in terms of inputs, processes, and outputs. Detailed review on taxonomy alignment research is provided in Chapter 3 of this dissertation. Prior work by scholars who have developed and implemented the logic-based taxonomy alignment approach employed in this dissertation uses region connection calculus (RCC-5) with its five base relations when aligning taxonomies (Thau, Bowers, & Ludäscher, 2008; Thau & Ludäscher, 2007). The key notions of this approach are discussed in detail in Chapter 4 of this dissertation.

### 1.2.2 Taxonomies with Socio-geographic Contexts

Social geography addresses “social identities and groupings” that are organized by spatial, or geographic relations (Del Casino Jr, 2009; Del Casino Jr, Thomas, Cloke, & Panelli, 2011). From a system’s perspective, socio-geographic boundaries are often created and maintained from the top-down (Trist, 1981). These socio-geographic boundaries can be manifested by taxonomies, which are hierarchical in structure, and often constructed from the top-down as

well. For instance, the Department of the Interior reorganization of the U.S. regions shows that taxonomies are created from an authority’s (from the top) perspective with a set of rules to divide the boundaries. In Bowker and Star (2000)’s book, the authors explain how government officials imposed a racial classification of people that constrained their mobility to move, work, or study in different areas during the Apartheid South Africa.

These system-imposed, socio-geographic taxonomies may encode latent assumptions that can be susceptible to issues such as ghettoization, bias, erasure, or omission (Adler & Tennis, 2013). Over the years, KO scholars have examined these issues in a wide range of research topics, such as in the classification of gender (Olson, 2002), racial groups (Higgins, 2016), and indigenous populations (Littletree & Metoyer, 2015). They find that misrepresented bibliographic classifications tend to provide a one-sided perspective that favors a dominant world view (Adler, 2016; Olson, 2002).

Furthermore, the emergence of heterogeneous knowledge organization systems and large-scale datasets makes it increasingly difficult to decode the underlying dominant world view within complex information systems, especially when socio-geographic contexts are involved (Del Casino Jr, 2009; Del Casino Jr et al., 2011; Enos, 2017). For instance, people’s beliefs on the geographic categorization of indigenous people’s tribes are highly informed by the government (Jarvis, 2017; McCulloch & Wilkins, 1995). This knowledge is deeply entrenched in peoples’ beliefs and may later be woven into political discourse about indigenous peoples’ territories, and even, voting rights (Templeman, 2018).

Efforts have been made to uncover latent assumptions about minorities in bibliographic classification by advocating for awareness, or to design classifications for future changes (Adler, 2016; Tennis, 2012). Other projects attempt to create taxonomies serving minorities (Kam, 2007; Lilley, 2015; Littletree & Metoyer, 2015), but these classifications are easily overshadowed by mainstream classifications that assimilate biased concepts about minorities Higgins (2016). Alignment of the lesser-known to the mainstream could be useful to bridge the connection of two taxonomies, but most alignment results offer a unitary merged solution

that may result in information loss to the original taxonomies or KOSs (Chan & Zeng, 2006; de Andrade & de Lara, 2016; Zeng & Chan, 2010).

### 1.2.3 Biodiversity Taxonomies with Socio-geographic Contexts

Though taxonomy alignment practices occur in many domains, Linnaean taxonomies are one of the oldest known taxonomies. These taxonomies organize perceived species of living organisms into higher-level ranks: Kingdom, Phylum, Class, Order, Family, Genus, Species. Concepts can be aligned in reference to a variety of defining characteristics, such as nomenclatural relationships, member composition, or diagnostic features of child or parent concepts (Franz et al., 2015). Other domains include phenetics, and genealogy, and taxonomies in comparative sub-disciplines as well as in conservation (Franz, Pier, et al., 2016). As described in Franz, Chen, et al. (2016), 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, whenever 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, Chen, et al., 2016).

What further exacerbates the situation is the difficulty to maintain a seamless and explicit navigation between biodiversity datasets and biodiversity literature, due to lack of consensus of the heterogeneous taxonomies used. Aggregated biodiversity databases such as the Global Biodiversity Information Facility (GBIF) have been the forefront in biodiversity data integration endeavors over the years. GBIF has become the de facto source that biodiversity researchers worldwide obtain species occurrences datasets (Robertson et al., 2014). GBIF aggregates numerous species occurrence datasets from different data sources

and collaborating institutions, but the quality of these datasets can be improved. *Within* the aggregators, numerous unresolved problems relating to integrating heterogeneous data already existed, such as consolidating different data types, reconciling multiple standards and taxonomies, and linking physical specimens, OCR-scanned legacy documents, to born-digital data and items (Bowker, 2000b; Cragin, 2003; Lapp, Morris, Catapano, Hobern, & Morrison, 2011). To connect items *across* not well-curated aggregated databases is more problematic when species taxonomies are different, geographic taxonomies are different, or even, the databases' web taxonomies are different. Biodiversity researchers who rely on these linked databases to interpret new species information, to classify species into the tree of life, or to recommend new conservation policy can potentially make flawed decisions (Bowker, 2000b; Cragin, 2003).

To reliably retrieve species information from a species occurrence dataset may already be difficult. It becomes more complex when socio-geographic aspects are involved. Biodiversity researchers rely heavily on additional information relating to the occurrence of species (Lapp et al., 2011). Integrated analyses are of great importance to biodiversity informatics to perform various tasks such as ecological niche-modeling, invasive species control, or habitat conservation, but geographic, time, or climatology data are created with different purposes, making it difficult to re-purpose them to fit with the needs of the biodiversity community (Bowker, 2000b; Cragin, 2003; Lapp et al., 2011). When geographic information is coupled with centuries-old historical data and publication, it can be even more challenging to find the correct place-names or discover new things from these biodiversity legacy documents (Bowker, 2000b; Lapp et al., 2011).

### 1.3 Goals of the Dissertation

This dissertation demonstrates the utility of the logic-based alignment approach in several socio-geographic contexts, including the United States maps, country taxonomies, indige-

nous peoples' tribes, and historical sovereignties in biodiversity data. The first three contexts are the basis for showing the feasibility of this approach in reconciling socio-geographic ambiguities; the last context stems from discussions with biodiversity scholars who pointed out a much-needed solution to reconcile not only species' names ambiguities, but also geographic ambiguities in biodiversity data (Nico M. Franz, James B. Whitfield).

This logic-based taxonomy alignment approach enables a symmetric mapping process to reduce information loss of all taxonomies involved. At the same time, the result of an alignment may produce multiple merged solutions. Specifically, the goals of this dissertation are to map concepts in taxonomies that:

- (1) preserve the original information in both taxonomies in the merged solution(s); and
- (2) allow for multiple possible solutions.

## 1.4 Research Questions

In line with the goals of this dissertation, I develop my research questions central to taxonomy alignment approaches. The questions are as follows:

**RQ1** Given a pair of taxonomies, how do we align them so that the information from both taxonomies is preserved?

Taxonomy alignment approaches usually operate on an asymmetric mapping process. This means that there are a source and a target taxonomy, and the alignment result is anchored on the target taxonomy. If the alignment process from the source to the target taxonomy is not logically consistent, asymmetric mapping eliminates the conflicting information (concepts and edges) from the source taxonomy. The logic-based alignment approach employed in this dissertation, on the other hand, is a symmetric mapping process in which the taxonomies have equal standings (no source nor target). The

information from both taxonomies is supposedly still all retained after the alignment occurred.

This research question is addressed recurrently in Chapters 5, 6, 7, and 8. From these chapters, I analyze and discuss how and what information is preserved in the logic-based taxonomy alignment process.

**RQ2** Given a pair of taxonomies, how, if at all, do we get multiple merged solutions when aligning the taxonomies?

The result of a taxonomy alignment problem is usually a single, unified merged solution. This research question aims to demonstrate how the logic-based approach employed in this dissertation to alignment taxonomies can yield more than one merged solutions.

This research question is addressed in Chapters 6, and 8, where there are multiple merged solutions in the use cases.

## 1.5 Organization of Chapters

**Chapter 2** formally defines key concepts and examples that are recurrent in this dissertation.

**Chapter 3** provides a systematic review of existing taxonomy alignment approaches, the characteristics, and the matching dimensions of these approaches.

**Chapter 4** explains in depth the key concepts relating to the logic-based taxonomy alignment approach used in this dissertation. Parts of this chapter are adapted from the following publications in which I am the lead author (Cheng et al., 2017a; Cheng & Ludäscher, 2018):

Cheng, Y. Y., Franz, N., Schneider, J., Yu, S., Rodenhausen, T., and 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., and Ludäscher, B. (2018, January). Full of beans: A study on the alignment of two flowering plants classification systems. *In 18th European Networked Knowledge Organization Systems Workshop (NKOS) Workshop Proceedings* (Vol. 2200, p. 53-63). CEUR-WS.

**Chapter 5** introduces the two use cases of conflicting taxonomies, each with two taxonomies, about regions of the United States. This use case is similar to the ones brought up in the beginning of this chapter, demonstrating the socio-geographic taxonomies that were created from the top-down. This chapter is adapted from the following publication (Cheng et al., 2017a), which I am the lead author:

Cheng, Y. Y., Franz, N., Schneider, J., Yu, S., Rodenhausen, T., and 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.

**Chapter 6** introduces the use case of conflicting taxonomies in geopolitics with three input country taxonomies. This chapter takes on Hiemstra and Mountz (2011)’s point of view that geopolitics is a socio-geographic phenomenon that ties power and relations to space. This chapter is adapted from the following publication (Cheng & Ludäscher, 2019b), of which I am the lead author:

Cheng, Y. Y., and Ludäscher, B. (2019). Exploring Geopolitical Realities through Taxonomies: The Case of Taiwan. *NASKO*, 7(1), 77-93.

**Chapter 7** introduces the use case of conflicting taxonomies of the indigenous peoples’ Tribes in Taiwan and their electoral voting groups, with five input taxonomies. This dis-

sertation agrees with Coombes, Gombay, Johnson, and Shaw (2011)’s stance on that it is contested to even consider indigenous peoples’ geographies or groupings as “conceptually discrete”. Cognizant of the debate, this chapter considers the taxonomies imposed by the Taiwanese government of the indigenous peoples’ tribes into discrete groupings a socio-geographic concern. This chapter chapter is adapted from the following publication, of which I am the lead author:

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).

**Chapter 8** highlights the complexity of biodiversity data, and how socio-geographic information plays a major role in connecting biodiversity data to legacy literature. The chapter introduces the use case of 14 conflicting taxonomies in historical geographic sovereignties in biodiversity data and provides a conceptual model for aligning these taxonomies.

**Chapter 9** answers the research questions and concludes with the contributions of this dissertation.

For chapters that have been adapted from the *Proceedings of the Association for Information Science and Technology*, the policy for thesis reuse is located at the Section on Permissions-AUTHORS<sup>10</sup>:“ If you wish to reuse your own article (or an amended version of it) in a new publication of which you are the author, editor or co-editor, prior permission is not required (with the usual acknowledgements)”.

For chapters that have been adapted from the *Proceedings of the North American Symposium of Knowledge Organization*, the policy for re-distributing and publishing the work can be found on the Copyright Notice section<sup>11</sup>:“Authors are able to enter into separate,

---

<sup>10</sup><https://asistdl.onlinelibrary.wiley.com/hub/journal/23739231/homepage/permissions>

<sup>11</sup><https://journals.lib.washington.edu/index.php/nasko/about/submissions#copyrightNotice>

additional contractual arrangements for the non-exclusive distribution of the journal’s published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal”.

For chapters that have adapted from the Proceedings of the 18th European Networked Knowledge Organization Systems (NKOS) Workshop, the policy for thesis reuse copyright is listed on the proceedings’ webpage<sup>12</sup>: “Copyright for the individual papers by the papers’ authors. Copying permitted for private and academic purposes”.

---

<sup>12</sup><http://ceur-ws.org/Vol-2200/>

# CHAPTER 2

## PRELIMINARIES

In this chapter, I define and formalize taxonomies, RCC-5, and taxonomy alignment problems (TAP). The formalizations follow closely those who reason about taxonomies in the same logic-based taxonomy alignment approach this dissertation employed (Chen, 2014; Thau, 2010; Yu, 2017).

### 2.1 Definition: Taxonomy

In this dissertation, a taxonomy  $T$  is defined as a pair:

$$T = (\mathbf{C}, \mathbf{E}) \tag{2.1}$$

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, Herrmann, & Winston, 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. For example, the concepts in the taxonomies in Chapter 5 are interpreted as spatial regions, and the edges denote (proper) spatial containment.

When formalizing taxonomies in first-order predicate logic (FOL), it is common to model concepts  $A$ ,  $B$  as unary predicates  $A(x)$  and  $B(x)$ , respectively. Then an edge such as  $(A, B)$  is formalized using a FOL sentence as follows:

$$\forall x(A(x) \rightarrow B(x)) \quad (2.2)$$

The taxonomies I consider in this dissertation is either tree(s) (especially for input taxonomies) or a directed acyclic graph(s) (DAG), i.e., partial order. Here I provide examples for both scenarios.

### 2.1.1 Example: When a taxonomy is a tree

Consider the following example of a simplified view of the animal kingdom, revised from Thau (2010). Let the taxonomy  $T = (\mathbf{C}, \mathbf{E})$  be consisted of:

- $\mathbf{C} = \{\text{Animals, Vertebrates, Invertebrates, Mammals, Reptiles, Insects, Mollusks, Turtles, Crocodiles, Beavers, Lions, Wasps, Ants, Snails, Squid}\}$ .
- $\mathbf{E} = \{(\text{Vertebrates, Animals}), (\text{Invertebrates, Animals}), (\text{Mammals, Vertebrates}), (\text{Reptiles, Vertebrates}), (\text{Insects, Invertebrates}), (\text{Mollusks, Invertebrates}), (\text{Turtles, Reptiles}), (\text{Crocodiles, Reptiles}), (\text{Beavers, Mammals}), (\text{Lions, Mammals}), (\text{Wasps, Insects}), (\text{Ants, Insects}), (\text{Snails, Mollusks}), (\text{Squid, Mollusks})\}$ .

An example when a taxonomy  $T$  is a tree is shown in Figure 2.1.

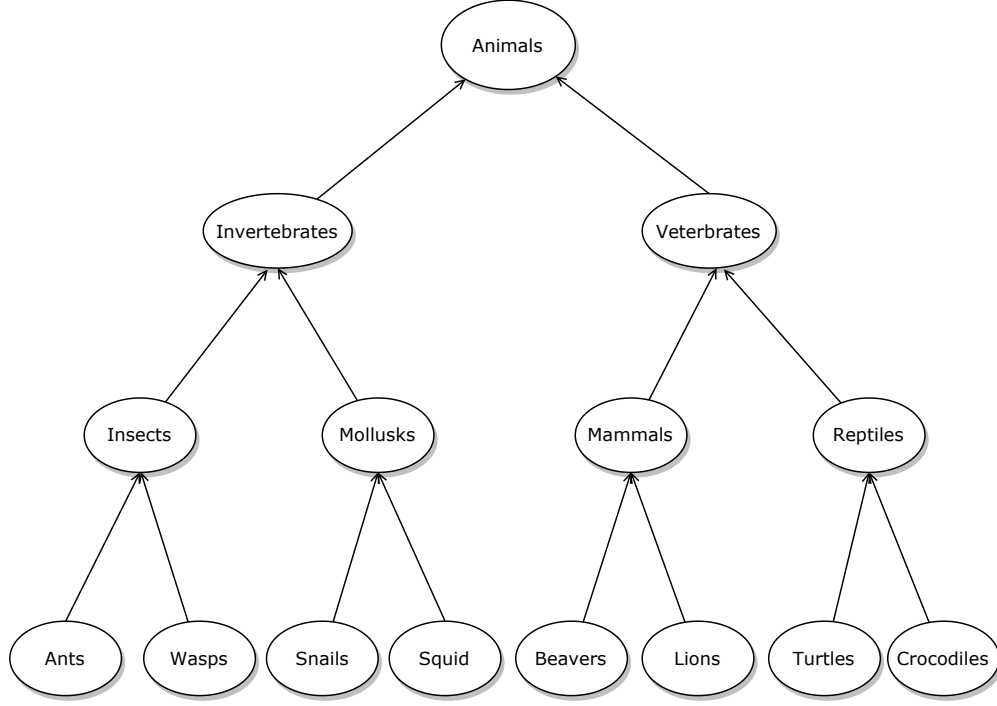


Figure 2.1: Example when a taxonomy is a tree.

### 2.1.2 Example: When a taxonomy is a DAG

Example 2 is an alternative classification to classify the above-mentioned animals. Here the taxonomy  $T = (\mathbf{C}, \mathbf{E})$  consists of:

- $\mathbf{C} = \{\text{Animal, Aquatic Animals, Terrestrial Animals, Turtles, Crocodiles, Beavers, Lions, Wasps, Ants, Snails, Squid}\}$ .
- $\mathbf{E} = \{(\text{Aquatic Animals, Animals}), (\text{Terrestrial Animals}), (\text{Squid, Aquatic Animals}), (\text{Turtles, Aquatic Animals}), (\text{Turtles, Terrestrial Animals}), (\text{Crocodiles, Aquatic Animals}), (\text{Crocodiles, Terrestrial Animals}), (\text{Beavers, Aquatic Animals}), (\text{Beavers, Terrestrial Animals}), (\text{Lions, Terrestrial Animals}), (\text{Wasps, Terrestrial Animals}), (\text{Snails, Terrestrial Animals}), (\text{Ants, Terrestrial Animals})\}$ .

An example when a taxonomy  $T$  is a DAG is shown in Figure 2.2.

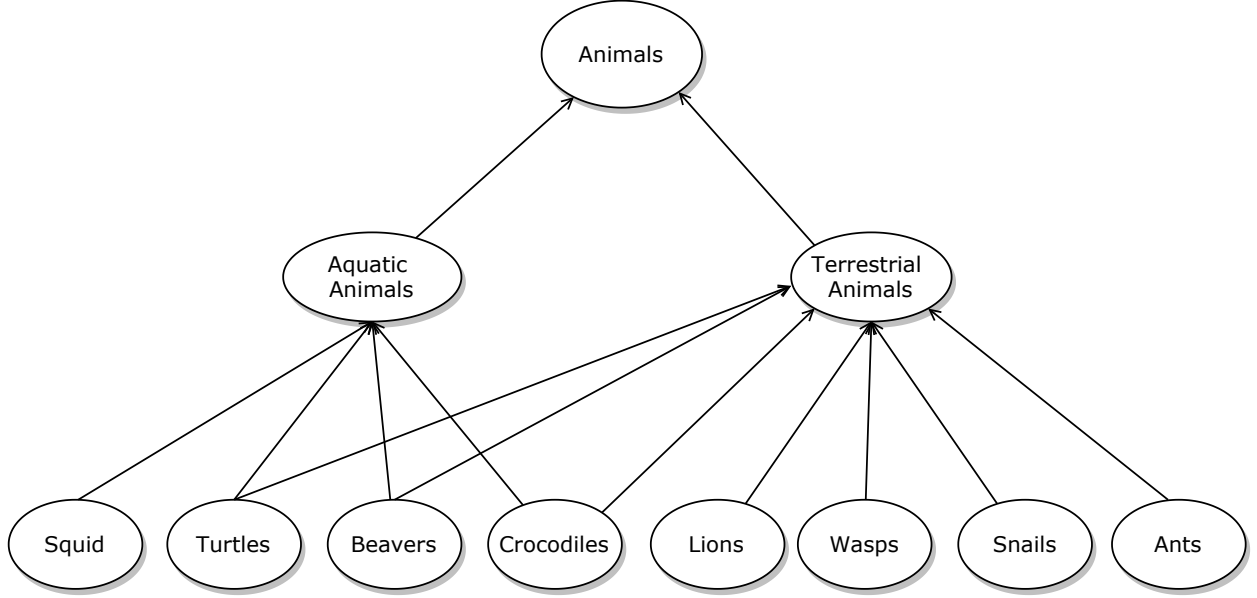


Figure 2.2: Example when a taxonomy is a DAG.

When the taxonomy is a tree, such as that in Example 1, all the sibling concepts are disjoint from each other. When a taxonomy is a DAG, sibling concepts are not disjoint from one another on the leaf-level, nor the upper-level classifications.

## 2.2 Definition: Region Connection Calculus (RCC)

As noted by Brachman (1983); Chaffin et al. (1988); Gerstl and Pribbenow (1995); Woods (1975), there are many interpretations of is-a edges. For the two examples above, the edges can possibly be interpreted as part-whole relations. However, it might be difficult to describe how aquatic animals relate to terrestrial animals by is-a terms. To specify more ways to relate concepts of taxonomies, we briefly introduce here the notion of Region Connection Calculus (RCC) here. Details on RCC are described in Chapter 4. In particular, we use RCC-5, a variant of the RCC family.

A pair of concepts  $A, B$  can be defined by one of the five base relations of RCC-5, mainly: equivalent (**EQ**), disjointness (**DR**), overlap (**PO**), proper part (**PP**), and proper

part inverse (**PPi**). RCC-5 uses different relationships  $R$  to describe how two concepts (or taxa) are related. This is denoted as  $A R B$ , where  $R \in \mathbf{EQ}, \mathbf{DR}, \mathbf{PO}, \mathbf{PP}, \mathbf{PPi}$ , and relating  $A$  to  $B$  will generate  $2^5 = 32$  pairwise relations. To define each of the RCC-5 relations in FOL terms:

When  $A$  is *equivalent* to  $B$ , i.e.,  $A \mathbf{EQ} B$ , we can express this as:

$$\forall x(A(x) \leftrightarrow B(x)) \quad (2.3)$$

When  $A$  is a *disjoint region* from  $B$ , i.e.,  $A \mathbf{DR} B$ , we can express this as:

$$\neg \exists x(A(x) \wedge B(x)) \quad (2.4)$$

When  $A$  *partially overlaps* with  $B$ , i.e.,  $A \mathbf{PO} B$ , in FOL terms it is denoted as:

$$\exists x(A(x) \wedge B(x)) \wedge \exists x(A(x) \wedge \neg B(x)) \wedge \exists x(B(x) \wedge \neg A(x)) \quad (2.5)$$

When  $A$  is a *proper part* of  $B$ , i.e.,  $A \mathbf{PP} B$ , we can express this as:

$$\forall x(A(x) \rightarrow B(x)) \wedge \exists x(B(x) \wedge \neg A(x)) \quad (2.6)$$

When  $A$  is *proper part inverse*  $B$ , i.e.,  $A \mathbf{PPi} B$ , in FOL terms it is denoted as:

$$\forall x(B(x) \rightarrow A(x)) \wedge \exists x(A(x) \wedge \neg B(x)) \quad (2.7)$$

## 2.3 Definition: Taxonomy Alignment Problems (TAP)

Consider the above two examples of the different ways to classify animals. Suppose we want to align these two taxonomies to see how they relate to each other, we can link them via RCC-5 relations, and the whole process constitutes a “taxonomy alignment problem”. Specifically,

here we define a taxonomy alignment problem (TAP) as:

Given two (or more) taxonomies,  $T_1$ ,  $T_2$  are inter-linked (or “formally crosswalked”) via a set of input relations (or articulations  $A$ ), defined as RCC-5 relations 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 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.

## 2.4 Chapter Summary

In this chapter, I define what a taxonomy is and provide examples on when taxonomy is a tree as opposed to when a taxonomy is a DAG. I then define each of the RCC-5 relations in first order logic terms. The RCC-5 relations are used extensively throughout this dissertation to relate concepts in different taxonomies in a taxonomy alignment problem, which I defined at the last section of this chapter. The definitions and formalizations provided in this chapter will be used throughout this dissertation.

## CHAPTER 3

# TAXONOMY ALIGNMENT METHODS: A REVIEW

This chapter reviews literature relating to taxonomy mapping, matching, or alignment methods. I take on an angle from the knowledge organization systems (KOS) framework (Hodge, 2000; Zeng, 2008), and specifically examining the level of KOS on “taxonomies”.

### 3.1 Knowledge Organization Systems, Taxonomies, and Ontologies

Knowledge organization systems are all kinds of knowledge structures used to organize information and concepts (Hodge, 2000; Mazzocchi, 2018; Zeng, 2008). Hodge (2000) identified types of knowledge organization systems into term lists, classifications and categories, and relationship lists; Zeng (2008) added “metadata-like models” and further denote the different dimensions of each KOS. “Taxonomies”, in Zeng (2008)’s model, are 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 arise 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. 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).

According to Hodge (2000), ontologies:

“...represent *complex relationships* among objects, and include rules and axioms...”

And taxonomies:

“...provide a *hierarchical arrangement* of numeric or alphabetic notation to represent broad topics.”

Note 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 usually 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 due to how both KOSs are used to classify natural languages; and the techniques to classify concepts in a taxonomy are similar to the automatic indexing techniques used in ontologies (Gilchrist, 2003). 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; Soergel, 2009; Zeng, 2008).

## 3.2 Myriads of Alignment Methods

The usage of taxonomies is ubiquitous. Many have attempted to define what a taxonomy is, but the definition of the term often varies. Gilchrist (2003), as cited by Mazzocchi (2018), gives at least five different meanings of a taxonomy:

1. web directories;
2. taxonomies to support automatic indexing;
3. taxonomies created by automatic categorization;
4. front end filters;
5. corporate taxonomies.

Even though there are many definitions 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; Van Hooland & Verborgh, 2014). We provide our definition of taxonomies in the previous chapter: 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 chapter, we will follow our definition of taxonomy as inclusion and exclusion criteria of the review corpus.

The desire to make taxonomies, or knowledge organization systems, interoperable with each other, is also a common practice (Chan & Zeng, 2006; 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), where schemas can

also be considered as a type of KOS. 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, Lv, Lan, & Zhang, 2018) have been emerging since Berners-Lee, Hendler, and Lassila’s disruptive concept of the semantic web (Berners-Lee, Hendler, & Lassila, 2001). In Information Sciences, various interoperability approaches have been discussed extensively since the beginning of classification systems and metadata standards (Chan & Zeng, 2006; Zeng & Chan, 2006). 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<sup>1</sup> (Baca, 2016), or one of the many crosswalks from MARC to other standards by the Library of Congress (of Congress, 2008). Spanning across different fields, *taxonomy alignment* (mapping, matching, merging) work flourishes in biodiversity (Franz, Chen, et al., 2016), biology (Kanehisa, Sato, & Kawashima, 2022), e-commerce (Aanen, Nederstigt, Vandić, & Fräsincar, 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 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 invented.

### 3.3 Level of KOS examined

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

---

<sup>1</sup>Getty crosswalk: [http://www.getty.edu/research/publications/electronic\\_publications/intrometadata/crosswalks.html](http://www.getty.edu/research/publications/electronic_publications/intrometadata/crosswalks.html)

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, McDermott, & Qi, 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.

This chapter 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.

### 3.4 Identifying Key Papers

Given that the usage of taxonomies spans across different application areas and disciplines, 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” by developing a coding scheme for analyzing the articles (Kelly & Sugimoto, 2013). Common sources for finding articles include Web of Science, Scopus, and domain specific databases, such as how Erfani and Abedin (2018), and Siqueira and Martins (2022) implemented in their studies. One of the advantages to leverage systematic literature review is that it allows transparency and reproducibility of the search process.

### 3.4.1 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: studies that are surveying or reviewing many alignment methods.

### 3.4.2 Source Identification and Search Strategies

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 LIS. Our search terms are "taxonomy alignment", "taxonomy mapping", "taxonomy matching", or "taxonomy merging" anywhere any of these terms appear in an article title, keyword, abstract, or full-text. We enumerate the search terms in each databases 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"
```

### 3.4.3 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, 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 “ontologies” or other KOSs, it will be excluded. If the term taxonomies and ontologies are used interchangeably, we based on the researchers' definition to determine if the KOS they are examine have only one node type and one edge type (is-a) (For example, we include the study from Raunich and Rahm (2014), their definition 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 conducted an evaluation of their alignment process by comparing to other

“ontology” matchers (e.g., Aanen et al. (2012), though using the term “taxonomy mapping” throughout their paper, compare their matcher SCHEMA with other ontology matchers by Noy and Musen (2003) and Park and Kim (2007))

- (4) the paper conducted an evaluation of their alignment process by using benchmark “ontology” alignment datasets (e.g., we exclude the study by Lin, Liu, Zhang, and Wang (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)<sup>2</sup> for evaluating their method.

This process results in 28 papers for final assessment.

### 3.5 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:

“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.”

---

<sup>2</sup>OAEI:<http://oei.ontologymatching.org/2017/>

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 dissertation:

A taxonomy  $T$  is defined as a pair:

$$T = (\mathbf{C}, \mathbf{E}) \tag{3.1}$$

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. For example, the concepts in the taxonomies in Chapter 5 are interpreted as spatial regions, and the edges denote (proper) spatial containment.

## 3.6 Characteristics of Taxonomy Alignment Approaches

In this section, we attempt to provide a systematic way to categorize all the characteristics associated with taxonomy alignment, mapping, matching, merging approaches based on the 38 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. It could be because taxonomies are often viewed as “very simple ontologies reduced to classification structures” (Reynaud & Safar, 2006) that makes it difficult to distinguish ontology alignment methods with taxonomy alignment methods. It could also be because taxonomies are prevalent used in many fields, as shown that the 38 papers we gathered came from many application areas, and that makes it more difficult to provide a systematic examination on these approaches. Here we adopt some of the characteristics in Rahm and Bernstein (2001), Shvaiko and Euzenat (2005), and Shvaiko and Euzenat (2011)’s work from schema matching and ontology matching to make sense of taxonomy alignment approaches. Tables 3.1, 3.2, and 3.3 present an overview of the characteristics on the studies we examined.

### 3.6.1 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, T1.cherry and T2.cherries may be mapped as equivalent given the concepts share the same word stem. In Rahm and Bernstein (2001), Shvaiko and Euzenat

Table 3.1: Characteristics of Taxonomy Alignment Approaches -1

|                             | ATOM [1]                        | Chen [2]                        | CROEA [3]          | Daraio [4]           | Euler/X [5]                   | HAMSTER [6]                           | Jung [7]                                     |
|-----------------------------|---------------------------------|---------------------------------|--------------------|----------------------|-------------------------------|---------------------------------------|----------------------------------------------|
| <b>methodology</b>          | logic-based                     | logic-based                     | NLP                | NLP                  | logic-based                   | NLP                                   | NLP                                          |
|                             |                                 |                                 | supervised         | supervised           |                               | unsupervised                          | unsupervised                                 |
| <b>application areas</b>    | general                         | e-commerce                      | education          | e-commerce           | geography, biodiversity       | e-commerce                            | digital library                              |
| <b>element vs structure</b> | element and structure           | element and structure           | element            | element              | element and structure         | element                               | element                                      |
| <b>symmetry</b>             | asymmetric                      | asymmetric                      | asymmetric         | asymmetric           | symmetric                     | asymmetric                            | symmetric                                    |
| <b>pairwise vs multiple</b> | pairwise                        | multiple                        | multiple           | pairwise             | pairwise and multiple         | pairwise                              | pairwise                                     |
| <b>Matching Dimensions</b>  |                                 |                                 |                    |                      |                               |                                       |                                              |
| <b>input</b>                | $T_s, T_t$                      | $T_s, T_t$                      | document           | $T_s, T_t$           | $T_1, T_2 \dots T_n$          | $T_s, T_t$ , schemas                  | $T_1, T_2$                                   |
| <b>approach</b>             | equivalence, is-a, inverse is-a | equivalence, is-a, inverse-is-a | training documents | TF-IDF, Vector space | RCC-5                         | unsupervised matching with query logs | distance similarity                          |
| <b>output</b>               | one merged taxonomy             | candidate pairs                 | labelled documents | word pairs           | one or more merged taxonomies | aggregate class of query distribution | similarity metrics between pairs of concepts |

[1] ATOM in Raunich and Rahm (2014); [2] Chen et al. (2020); [3] CROEA in Mouriño-García, Pérez-Rodríguez, Anido-Rifón, Fernández-Iglesias, and Darriba-Bilbao (2018); [4] Daraio, Cagliero, Chiusano, Garza, and Ricupero (2020); [5] Euler/X in Cheng et al. (2017a); Cheng, Hoang, and Ludäscher (2020); Cheng and Ludäscher (2020b); 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)

(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. We aggregated these classes and simply group them as 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; L.-H. Lee et al., 2009; Musgrove, 2006; Ponzetto & Navigli, 2009; Reynaud & Safar, 2006). Some studies in the unsupervised NLP methods category tend to utilize 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 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-García et al., 2018). It is also interesting to note that with the rise of supervised NLP neural network approaches, studies that employed supervised methods tend to

Table 3.2: Characteristics of Taxonomy Alignment Approaches -2

|                             | Kanehisa [8]     | Lee [9]                                      | Li'11 [10]                        | Li'08 [11]                               | Maldonado [12]                              | Merge-Into [13]                                 | Musgrove [14]            |
|-----------------------------|------------------|----------------------------------------------|-----------------------------------|------------------------------------------|---------------------------------------------|-------------------------------------------------|--------------------------|
| <b>methodology</b>          | constraint-based | NLP                                          | constraint-based                  | NLP                                      | NLP                                         | NLP                                             | NLP                      |
|                             | ID-match         | unsupervised                                 | author defined rules              | unsupervised                             | supervised                                  | unsupervised                                    | unsupervised             |
| <b>application areas</b>    | biology          | cross-lingual                                | business                          | business                                 | business                                    | general                                         | e-commerce               |
| <b>element vs structure</b> | element          | element                                      | structure                         | element                                  | element                                     | element                                         | element                  |
| <b>symmetry</b>             | symmetric        | asymmetric                                   | asymmetric                        | asymmetric                               | asymmetric                                  | asymmetric                                      | asymmetric               |
| <b>pairwise vs multiple</b> | pairwise         | pairwise                                     | pairwise                          | pairwise                                 | pairwise                                    | multiple                                        | pairwise                 |
| <b>Matching Dimensions</b>  |                  |                                              |                                   |                                          |                                             |                                                 |                          |
| <b>input</b>                | $T_1, T_2$       | Chinese WordNet                              | $T_s, T_t$                        | $T_s, T_t$                               | $T_s, T_t$                                  | $T_s, T_t$                                      | $T_s, T_t$               |
| <b>approach</b>             | Join operator    | synonyms with DDC and domain lexico taxonomy | transforming tuples to dimensions | TF-IDF                                   | word embeddings; equivalency between labels | synonyms, node structure was taken into account | WordNet, word similarity |
| <b>output</b>               | updated taxonomy | predictions of Chinese WordNet domains       | updated taxonomy                  | target categories with confidence scores | updated taxonomy                            | a merged solution                               | ontology expansion       |

[8]Kanehisa et al. (2022); [9]L.-H. Lee, Yu, and Huang (2009); [10]J.-m. Li, Zhao, and Du (2011); [11]Z. Li et al. (2008); [12]Maldonado, Sharpe, and ter Horst (2021); [13] Merge-Into in Subramaniam, Nanavati, and Mukherjea (2009); [14] Musgrove (2006)

Table 3.3: Characteristics of Taxonomy Alignment Approaches -3

|                             | Pfeifer [15]                                                | Ponzetto [16]                                   | Reynaud [17]                    | Taxo-Semantics [18]                 | WETA [19]                       | Xu [20]                           |
|-----------------------------|-------------------------------------------------------------|-------------------------------------------------|---------------------------------|-------------------------------------|---------------------------------|-----------------------------------|
| <b>methodology</b>          | NLP                                                         | NLP                                             | NLP                             | NLP                                 | NLP                             | NLP                               |
|                             | unsupervised                                                | unsupervised                                    | unsupervised                    | unsupervised                        | supervised                      | supervised                        |
| <b>application areas</b>    | e-commerce                                                  | Wikipedia                                       | general                         | e-commerce                          | human resources                 | cross-lingual                     |
| <b>element vs structure</b> | element                                                     | element                                         | element and structure           | element                             | element                         | element                           |
| <b>symmetry</b>             | asymmetric                                                  | symmetric                                       | asymmetric                      | asymmetric                          | symmetric                       | asymmetric                        |
| <b>pairwise vs multiple</b> | pairwise                                                    | pairwise                                        | pairwise                        | pairwise                            | pairwise                        | pairwise                          |
| <b>Matching Dimensions</b>  |                                                             |                                                 |                                 |                                     |                                 |                                   |
| <b>input</b>                | $T_s, T_t$                                                  | Wiki-taxonomy                                   | $T_s, T_t$                      | $T_s, T_t$                          | $T_1, T_2$                      | $T_s, T_t$                        |
| <b>approach</b>             | hybrid matchers to compute word similarities                | compare trees from wikipedia to WordNet synsets | Lin similarity measure, WordNet | WordNet, word similarity            | Word embeddings; neural network | exact word match                  |
| <b>output</b>               | Mappings; relations between the pair, and confidence scores | weights on the similarities                     | suggested mapping candidates    | assessed semantically similar pairs | candidate pairs                 | category label performance scores |

[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)

be 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 are introduced as the mapping mechanisms. Chen et al. (2020) followed Raunich and Rahm (2014)’s approach very closely, and also utilize the 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 (Cheng et al., 2017a; Cheng, Hoang, & Ludäscher, 2020; Cheng & Ludäscher, 2020b; Franz, Chen, et al., 2016; Franz et al., 2015; Johnston, 2016; Thau & Ludäscher, 2007), qualitative reasoning such as Region Connection Calculus is used to denote five base relations (RCC-5): equivalence, proper part, inverse proper part, disjointness, and overlapping. A more thorough discussion on this approach will be provided in the next chapter. The number of formal relations used to work with the mapping 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. The 16 relations proposed by Inants and Euzenat (2015) are not within the scope of taxonomy alignment – the authors are trying to align ontologies, and 16 relations are needed for matching both the class-level and instance-level relations.

Usually, the goal for logic-based approaches is not only to align concepts semantically, but also 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 removed cycles from the source taxonomies when the

concepts are not logically consistent. In the Euler/X approaches, if there are inconsistencies among the taxonomies, an iterative process to revise the mapping relations will be needed so both taxonomies are preserved with logically aligned concepts. It is also worth noting that Euler/X approach allows for multiple merged taxonomies as output, if there are additional structurally possible solutions after merging. Raunich and Rahm (2014) and Chen et al. (2020) aims at a semi to fully automated logic-based approach towards taxonomy alignment, while the Euler/X approaches are semi-automated, requiring human-in-the-loop process.

### **Constraint-based Approaches**

There were only two studied we viewed as constraint-based among the 28 papers we examined. Constraint-based approaches are usually “based on keys and relationships” Rahm and Bernstein (2001), so it may be used more in schema matching contexts when relational databases are involved. In fact, in Kanehisa et al. (2022), the authors work with protein and genomics databases KEGG, and used a JOIN operation to conduct a match on key IDs in the taxonomies within the databases. J.-m. Li et al. (2011) is a more peculiar case, where the authors defined sets of rules that were written into the algorithm to transform the taxonomies and perform taxonomy alignment. Usually, the goal of constraint-based approaches is to provide a 1:1 match between concepts in both taxonomies.

#### **3.6.2 Application Areas**

In these 28 papers (20 matchers), nine of them were from the e-commerce or business domains – meaning they were mainly working with corporate, web directories, or product 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 are concerned with the domain area of cross-lingual mapping (L.-H. Lee et al., 2009; Xu & Sun, 2007), mainly

the mapping from leveraging English taxonomies to Chinese taxonomies. Other papers span across different domains, from education (Mouriño-García et al., 2018), to biology (Kanehisa et al., 2022), to human resources (Giabelli et al., 2022), or to the general areas where the alignment process can be applicable to all domains (Cheng et al., 2017a; Raunich & Rahm, 2014; Subramaniam et al., 2009).

### 3.6.3 Element vs. Structure

We adopt 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 are working on the level of element matching, that is, aligning concepts, classes, or other types of metadata elements; with the exception of logic-based studies, which work with both element and structure matching, concerning both the structures of the taxonomies as well as the semantics. Another exception is Reynaud and Safar (2006), where the authors employed two structural alignment techniques to work with more complex taxonomies.

### 3.6.4 Symmetry

Raunich and Rahm (2014) defined the taxonomy alignment approaches by symmetry. Asymmetric matching means there is a source taxonomy and a target taxonomy, and the alignment process are to fulfill the goal of the target taxonomy. Symmetric matching means there are no apparent source or target taxonomies, two (or multiple) taxonomies are treated equally, and the alignment goal is usually to preserve all the information in all taxonomies. We adopt this lingo of symmetric vs asymmetric because we find it 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 information loss from the source taxonomy which may be irreproducible. The works that are symmetric means that it could be computationally more taxing to satisfy all taxonomies’ constraints, but it provides a more diverse viewpoints which preserve more information from the original taxonomies.

### 3.6.5 Pairwise vs. Multiple

Most alignment methods we examined operate on a pairwise-level. Each time during the process, there are two taxonomies being aligned. 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 discuss 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, Chen, et al. (2016) attempts to align multiple taxonomies by aligning two at a time, whereas the Euler/X approach could also align three, or more taxonomies at a time.

### 3.6.6 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 approaches, seeing that all alignment methods have a common pipeline that involves *Input*, *Approach*, and *Output*.

In their book, Euzenat et al. (2007) elaborate that the ***Input*** is the kind of things that goes into comparison. In the 28 papers we examined, input are taxonomies in hierarchical

forms, but the representations can vary. Some could be in the form of XML files, WordNet dictionaries, or a document.

The ***approach***, or process, is mainly the basis to our methodology classification of these papers in the prior sections of this chapter. This step is the core of the taxonomy alignment approach – the matching process. 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)).

The ***output*** of a taxonomy alignment approach also varies from papers. In the logic-based Euler/X works, the output is one or more merged taxonomies (possible worlds) in a DAG visualization. For the NLP approaches works, the output is usually word pairs that could helped with enhancing or expanding the taxonomy. The output of some of these approaches may not necessary be a taxonomy, it could be a list of words, a list of categories, or numeric scores on how similar two concepts are in the taxonomies.

## 3.7 Chapter Summary

In this chapter, I review 28 studies that employ mapping, matching, alignment methods on the KOS level of “taxonomies”. I review various definitions of taxonomies in these studies, the characteristics of the alignment methods, and I categorize these alignment methods to Natural Language Processing approaches, logic-based approaches, and constraint-based approaches. While all matching methods have similar alignment pipeline (i.e., matching dimensions), involving two or more input taxonomies, the actual matching processes, and the output results, these methods differ in terms of application areas, structures, and other characteristics. Notably, most methods work with pairwise alignment, meaning the approach is only aligning two taxonomies at a time.

In the three logic-based approaches examined, two approaches focus on asymmetric mapping, meaning the outcomes of the alignments favor toward the target taxonomy. This pri-

oritizes efficiency over consistency of a taxonomy alignment processes, and can lead to the loss of information in both the source and the target taxonomies. The Euler/X approach I implemented in this dissertation focuses on symmetric mapping, meaning that both taxonomies are valued at the same level. If logical inconsistencies arise during the alignment process, the Euler/X approach requires us to resolve the inconsistencies. The results of the Euler/X logic-based alignment aim to preserve information from both taxonomies. Details on the Euler/X logic-based approach will be further discussed in the next chapter.

# CHAPTER 4

## TAXONOMY ALIGNMENT PROBLEM

*“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...”* – Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out*, p.159.

This chapter provides reviews of the aspects on the logic-based taxonomy approach used in this dissertation, including the mapping relations (articulations), qualitative reasoning mechanism for mapping (Region Connection Calculus), merged solution (possible world(s)), and the toolkit (Euler/X).

### 4.1 Overview

The definition of taxonomy within the scope of our work states that a taxonomy  $T$  is either a tree or a directed, acyclic graph (DAG) (Thau et al., 2008; Thau & Ludäscher, 2007) (see Chapter 2 for definitions). In addition, we apply additional constraints on the taxonomies in which each node is *disjoint* from all sibling nodes, meaning that any two nodes on the same level in the tree are mutually exclusive of each other (*sibling disjointness rule*). The children in each branch of the tree are also considered to be the only children of the parent with no prodigal sons or long, lost siblings (*parent coverage rule*).<sup>1</sup>

In this dissertation, I 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

---

<sup>1</sup>Parts of this chapter are adapted from Cheng et al. (2017a); Cheng and Ludäscher (2018)

A, defined as RCC-5 relations (see Chapter 2 and details below in this chapter), 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 the input TAP is either 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, e.g., in the form of multiple answer sets (Gebser, Kaminski, Kaufmann, & Schaub, 2012), or possible worlds (Denecker, Lierler, Truszczyński, & Vennekens, 2012).

## 4.2 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  $A \Phi B$ , where A and B are concepts from  $T_1$  and  $T_2$ , respectively, and where “ $\Phi$ ” is an 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.

Table 4.1: RCC-5 composition table

|                                    | Equivalent ( <b>EQ</b> ) | Disjoint ( <b>DR</b> ) | Overlap ( <b>PO</b> ) | Proper part ( <b>PP</b> ) | Proper part inverse ( <b>PPi</b> ) |
|------------------------------------|--------------------------|------------------------|-----------------------|---------------------------|------------------------------------|
| Equivalent ( <b>EQ</b> )           | EQ                       | DR                     | PO                    | PP                        | PPi                                |
| Disjoint ( <b>DR</b> )             | DR                       | EQ, DR, PO, PP, PPi    | PO, DR, PP            | DR, PO, PP                | DR                                 |
| Overlaps ( <b>PO</b> )             | PO                       | DR, PO, PPi            | EQ, DR, PO, PP, PPi   | PO, PP                    | DR, PO, PPi                        |
| Proper part ( <b>PP</b> )          | PP                       | DR                     | DR, PO, PP            | PP                        | EQ, DR, PO, PP, PPi                |
| Proper part inverse ( <b>PPi</b> ) | PPi                      | DR, PO, PPi            | PO, PPi               | EQ, PO, PP, PPi           | PPi                                |

Source:(Ragni & Wöfl, 2005)

### 4.3 Region Connection Calculus (RCC)

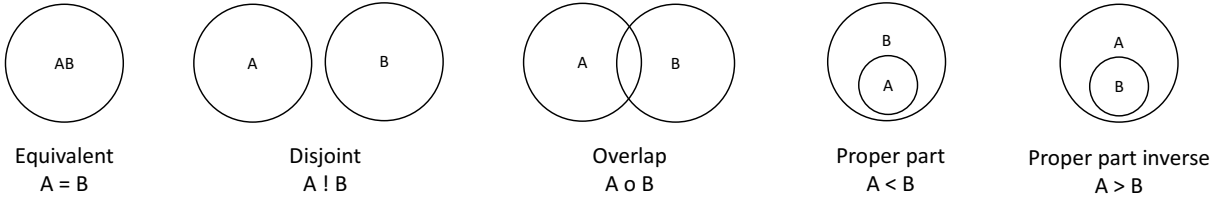


Figure 4.1: The region connection relations used in RCC-5

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, consisting of five pairwise disjoint and mutually exclusive relations, i.e., congruence (equals (**EQ**) or “=”), disjointness (disjoint (**DR**) or “!”), overlap (overlaps (**PO**) or “><”), proper part (is\_included\_in (**PP**) or “<”), and proper part inverse (includes (**PPi**) or “>”). See Figure 4.1 as well as the RCC-5 composition table in Table 4.1 (revised based on (Ragni & Wöfl, 2005)).

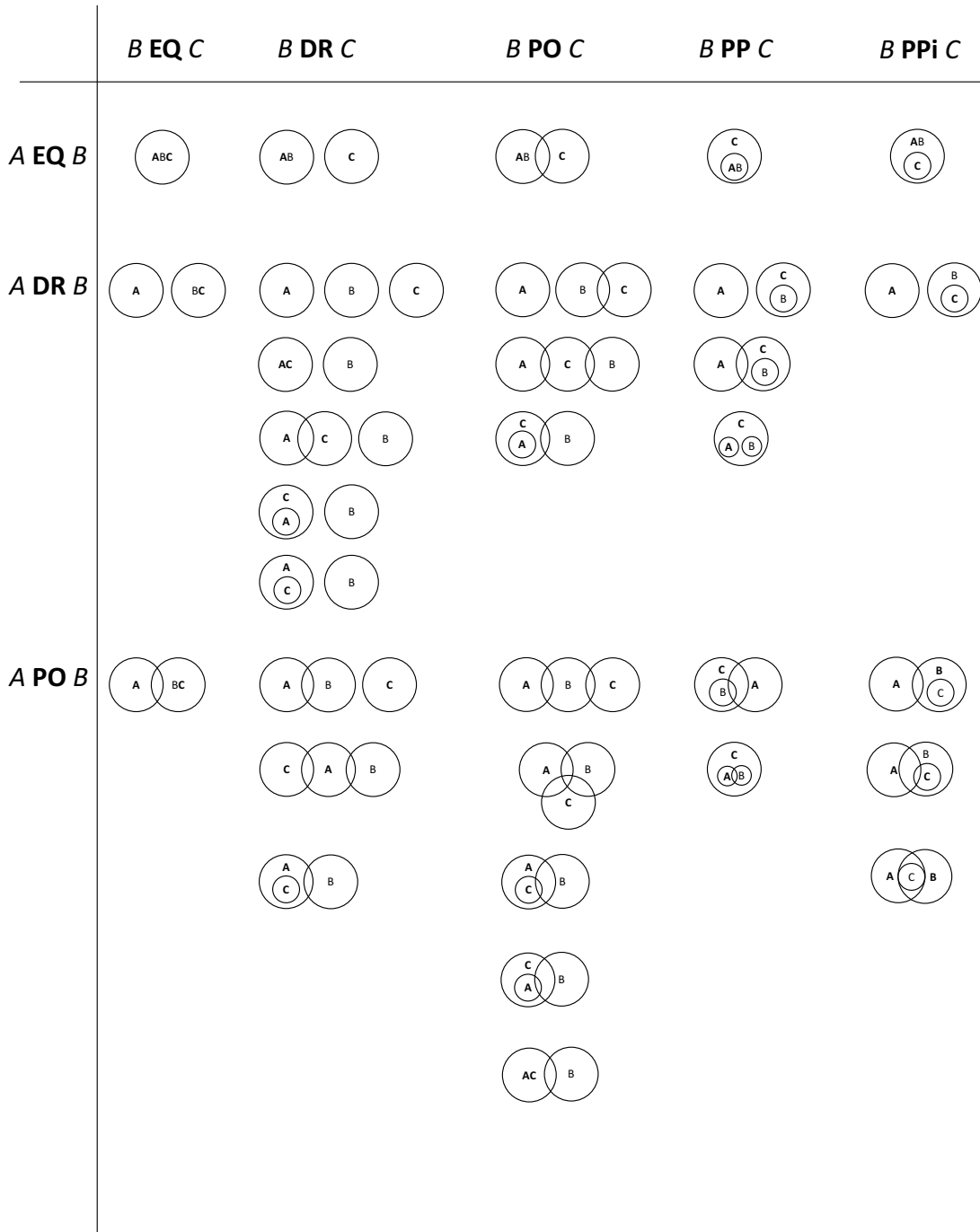


Figure 4.2: RCC-5 composition Euler diagrams for **EQ**, **DR**, **PO**

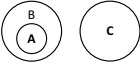
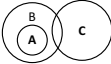
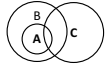
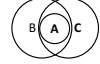
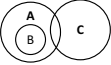
|                    | $B \text{ EQ } C$                                                                 | $B \text{ DR } C$                                                                   | $B \text{ PO } C$                                                                   | $B \text{ PP } C$                                                                     | $B \text{ PPi } C$                                                                  |
|--------------------|-----------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| $A \text{ PP } B$  |  |    |    |    |  |
|                    |                                                                                   |                                                                                     |    |                                                                                       |  |
|                    |                                                                                   |                                                                                     |    |                                                                                       |  |
|                    |                                                                                   |                                                                                     |                                                                                     |                                                                                       |  |
|                    |                                                                                   |                                                                                     |                                                                                     |                                                                                       |  |
| $A \text{ PPi } B$ |  |    |    |    |  |
|                    |                                                                                   |  |  |  |                                                                                     |
|                    |                                                                                   |  |                                                                                     |  |                                                                                     |
|                    |                                                                                   |                                                                                     |                                                                                     |  |                                                                                     |

Figure 4.3: RCC-5 composition Euler diagrams for **PP**, **PPi**

An example of the Euler diagrams created based on the RCC-5 composition table describes how the relations of A, B, and C are in Figure 4.2 and Figure 4.3. For instance, if

| <b>RCC-5 Relation</b> | <b>Euler/X notation</b> | <b>Meaning</b>                        |
|-----------------------|-------------------------|---------------------------------------|
| $A \mathbf{EQ} B$     | $A = B$                 | $A$ is equal to (or congruent to) $B$ |
| $A \mathbf{DR} B$     | $A \neq B$              | $A$ is disjoint from $B$              |
| $A \mathbf{PO} B$     | $A \circ B$             | $A$ overlaps with $B$                 |
| $A \mathbf{PP} B$     | $A < B$                 | $A$ is included in $B$                |
| $A \mathbf{PPi} B$    | $A > B$                 | $A$ includes $B$                      |

Table 4.2: RCC-5 Relations to Euler/X notations and Meanings

$A \mathbf{EQ} B$ ,  $B \mathbf{EQ} C$ , then the relation between  $A$  and  $C$  will also be **EQ**; if  $A \mathbf{PO} B$ ,  $B \mathbf{PO} C$ , then the relation between  $A$  and  $C$  can be all five relations (**EQ**, **DR**, **PO**, **PP**, **PPi**); if  $A \mathbf{PP} B$ , and  $B \mathbf{PO} C$ , then  $A$  and  $C$  can be **DR**, **PO**, or **PP**.

Table 4.2 shows the translation from RCC-5 relations to the tool’s (Euler/X, described later in this chapter) notation, and the actual meaning of the relations. Note in later chapters I will mainly be using Euler/X notations or verbatim meaning to describe the applications.

## 4.4 Possible Worlds

When encoding and solving TAPs via ASP, the different answer sets represent alternative merged taxonomy 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. A TAP may conclude with either 0, 1, or numerous PWs, meaning that the input taxonomies  $T_1$ ,  $T_2$ , and  $A$  are, respectively, an inconsistent result, a unique world, or ambiguous solutions. When the result is inconsistent (PW = 0), it means that the input  $A$  has logical contradictions or is overly specified, and some  $As$  need to be removed in order to make the TAP consistent. On the contrary, underspecified  $As$  cause ambiguous solutions (PW  $\geq 2$ ), and further constraints need to be added to refine the alignment results.

## 4.5 Modeling and Interpretations

Taxonomies sometimes may not be readily available in a tree-like structure. Thereby, conscious human decisions may be involved to model the taxonomies into a tree. Use cases in the following chapters of this dissertation will demonstrate how I transpose the taxonomies from maps into a tree.

Further, recognizing that existing taxonomies or TAP may not have a *ground truth* to verify against to form articulations, human interpretations of the articulations for TAPs may be needed. For use cases that are of highly specialized knowledge, domain experts’ insights are needed to specify how  $T_1$  relates to  $T_2$ . Such cases can be found in works by (Franz et al., 2015; Franz, Pier, et al., 2016), in which the authors have explored the use of logic-based RCC-5 alignment in the field of biodiversity systematics with detailed expert’s articulations.

In Chapter 2, I mention that there are many possible interpretations of concepts and edges in a taxonomy. In the scope of this dissertation to model cases in socio-geographic contexts, the concepts in the taxonomies can be interpreted as spatial regions. The edges can be interpreted as “containment” relations (contained in).

## 4.6 Euler/X and LeanEuler

Euler/X<sup>2</sup> 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. Euler/X has been developed over time from work within the scope of reasoning about taxonomies (Chen, 2014; Thau, 2010; Yu, 2017). Its predecessor CleanTax was developed by Thau, Bowers, and Ludäscher (2009), while Euler/X was mainly developed by Chen, Yu, Franz, Bowers, et al. (2014). A detailed system overview of Euler/X can be found in Chen, Yu, Kianmajd, et

---

<sup>2</sup>Euler/X: <https://github.com/EulerProject/>, the “X” in Euler/X stands for the ability to choose different reasoners or encoding within reasoners in the tool.

al. (2014). An interactive notebook-style version of Euler/X is also available (LeanEuler<sup>3</sup>) (Gupta & Ludäscher, 2019).

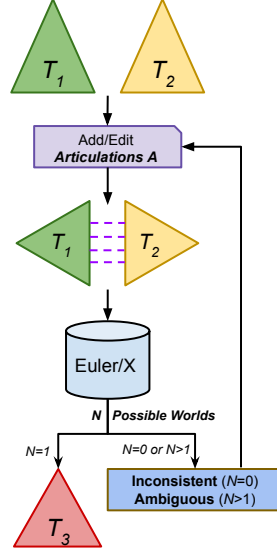


Figure 4.4: The alignment lifecycle of Euler/X

Figure adapted from (Cheng et al., 2017a).

Figure 4.4 shows the steps of implementing 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

Given taxonomies  $T_1$ ,  $T_2$ , and articulations  $A$ , Euler/X infers from these all possible worlds  $W_1, W_2, \dots, W_N$ . Each world  $W_i$  represents a solution taxonomy  $T_3$ , simultaneously satisfying  $T_1$ ,  $T_2$ , and  $A$ . If there are no possible worlds ( $N=0$ ) (Figure 4.4) then the input TAP is inconsistent (i.e., a contradiction can be derived) and the articulations of the alignments should be repaired, relaxed, or dropped. Conversely, if  $A$  is underspecified and

<sup>3</sup>LeanEuler: <https://github.com/idaks/LeanEuler>

| <b>T1/T1+T2</b> | <b>A</b> | <b>B</b>          | <b>C</b>          | <b>X</b> | <b>Y</b>          | <b>Z</b>          |
|-----------------|----------|-------------------|-------------------|----------|-------------------|-------------------|
| <b>A</b>        | =        | >                 | >                 | =        | >                 | >                 |
| <b>B</b>        | <        | =                 | !                 | <        | ! , = , o , < , > | ! , = , o , < , > |
| <b>C</b>        | <        | !                 | =                 | <        | ! , = , o , < , > | ! , = , o , < , > |
| <b>X</b>        | =        | >                 | >                 | =        | >                 | >                 |
| <b>Y</b>        | <        | ! , = , o , < , > | ! , = , o , < , > | <        | =                 | !                 |
| <b>Z</b>        | <        | ! , = , o , < , > | ! , = , o , < , > | <        | !                 | =                 |

Table 4.3: Maximally Informative Relations (MIR) table for the Magnificent 7 example

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).

## 4.7 A Simple Example of TAP: “ Magnificent 7 ”

To show a simple TAP example and the multiple possible worlds that exist, consider two taxonomies  $T_1$  and  $T_2$ , each with two children (Figure 4.5). Assume that we only know about these that the highest-level nodes are congruent to each other ( $T_1.a = T_2.x$ ); how  $T_1.b$ ,  $T_1.c$  relate to  $T_2.y$ ,  $T_2.z$  is unknown. With these *underspecified* articulations in our TAP, ambiguities arise. Similar to the RCC-5 composition table (4.1), we express this example in a table that demonstrates the *Maximally Informative Relations* (MIR) among the nodes in Table 4.3.

We end up with seven possible worlds where PW1 and PW2 depict how the concepts can be aligned congruently, PW3 to PW6 depict similar but subtle differences on how  $T_1.b$ ,  $T_1.c$  is included in or includes  $T_2.y$ ,  $T_2.z$ , while PW7 depicts  $T_1.b$ ,  $T_1.c$ ,  $T_2.y$ ,  $T_2.z$  all overlapping each other (red dashed lines). In reality, if we have more information about the taxonomy, we can specify more articulation to reduce the ambiguities of the alignment problem, which will then result in fewer, or one possible world.

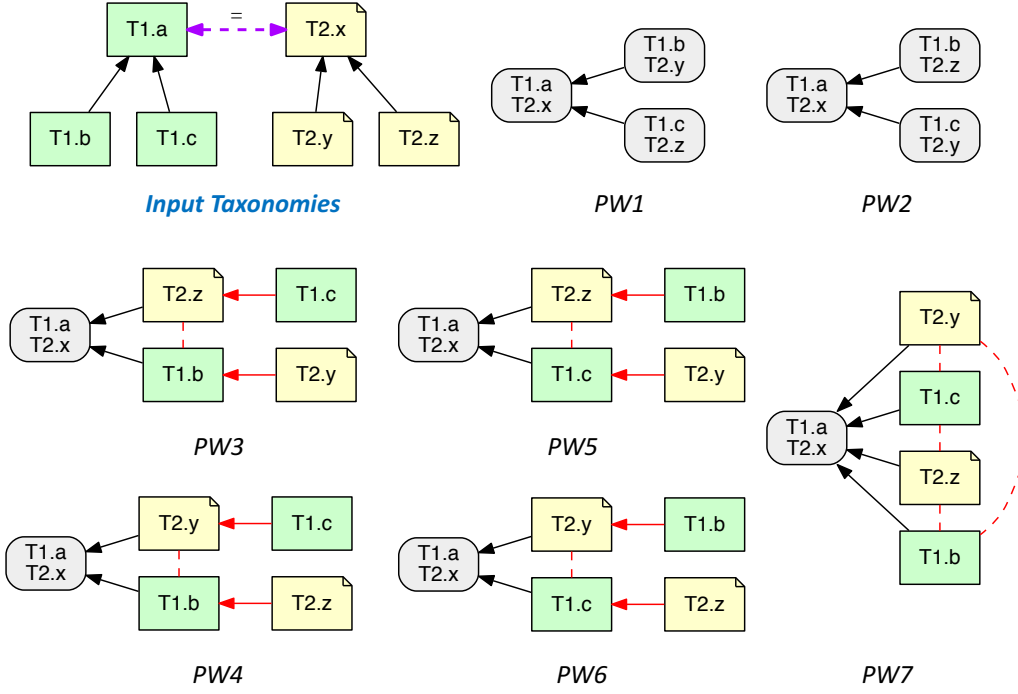


Figure 4.5: A simple taxonomy alignment problem with  $T_1$ ,  $T_2$  and 7 possible worlds

Figure adapted from (Cheng & Ludäscher, 2018).

## 4.8 Chapter Summary

In this chapter, I provide further details on the logic-based taxonomy alignment approach used in this dissertation. I explain the key notions used in this approach, namely the articulations, RCC-5, the possible worlds, the modeling and interpretations of an alignment, and the Euler/X toolkit. I also provide a simple illustrative taxonomy alignment problem example that aligns two taxonomies with one articulation. These key notions will be used recurrently in this dissertation, especially in the following chapters when discussing specific use cases of a taxonomy alignment problem.

## CHAPTER 5

# RECONCILING THE SOCIO-GEOGRAPHIC IN THE UNITED STATES MAPS

In this chapter, we discuss two use cases of reconciling maps of the United States where the United States is divided into conflicting geographic hierarchies. In one case, we align  $T_{CEN}$ , a taxonomy derived from the Census Bureau’s regions map, with  $T_{NDC}$ , from the National Diversity Council (NDC). In the second case, we align  $T_{CEN}$  with  $T_{TZ}$ , a taxonomy capturing the U.S. time zones. Using these case studies, we show the utility of this logic-based approach in socio-geographic contexts, and the approach’s ability to preserve the information of both input taxonomies in the merged solution.<sup>1</sup>

### 5.1 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.)*

---

<sup>1</sup>This chapter is adapted from 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 et al., 2017a).

Terminology confusion are common in our everyday lives. 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 5.1). 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.

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 (Chan & Zeng, 2006; de Andrade & de Lara, 2016; Zeng & Chan, 2010).

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, Chen, et al., 2016; Franz et al., 2015; Franz, Pier, et al., 2016) .

Many fields have contributed to taxonomy alignment research, but the key new idea in our line of work (Chen, Yu, Kianmajd, et al., 2014; Franz, Chen, et al., 2016; Franz et al., 2015; Franz, Pier, et al., 2016; Thau & Ludäscher, 2007) is to compare concepts A and

B via five exhaustive and mutually exclusive relationships: *congruence*, *inclusion*, *inverse inclusion*, *overlap*, and *disjointness* 4.1 from the Region Connection Calculus RCC-5 (Cohn & Renz, 2008; Randell, Cui, & Cohn, 1992).

Our objectives for this chapter are two-fold: first, we show the feasibility of using a logic-based approach for taxonomy alignment with RCC-5 through two use cases. 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 (details about articulations are discussed in Section 5.3.2).

The two use cases presented here 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 ( $T_{NDC}$ ) map with the Census Bureau’s Census Regions map ( $T_{CEN}$ ), on the state-level, taking a bottom-up approach. We call this case a “bottom-up” approach because Euler/X can leverage these leaf-level articulations to unambiguously infer the higher, region-level articulations.

Second, we will align the United States Time Zone ( $T_{TZ}$ ) and  $T_{CEN}$ , on a regional level, taking a top-down approach. WE call this a “top-down” approach because the articulations will not be on the leaf-level nodes, but rather the higher-level nodes in the taxonomies.

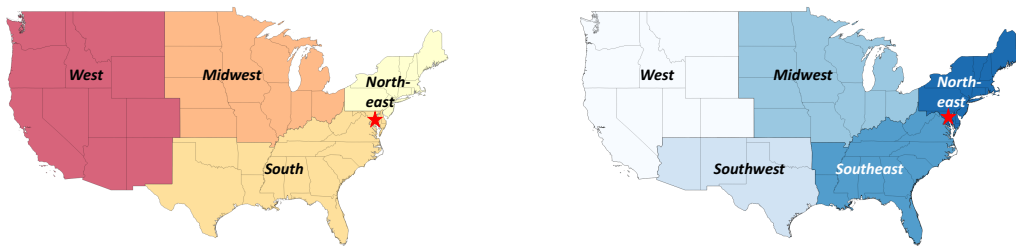


Figure 5.1: Different geo-taxonomies of the contiguous United States  
Census Bureau map  $T_{CEN}$  (left); National Diversity Council map  $T_{NDC}$  (right).

## 5.2 Related Work

Semantic interoperability is the 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, 2010).

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 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 (Chan & Zeng, 2006; de Andrade & de Lara, 2016; Zeng & Chan, 2010).

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, 2013a).

## 5.3 Taxonomy Alignment Problems

### 5.3.1 Step 1: Supply Input Taxonomies $T_1$ and $T_2$

Interpretations on the concepts and edges of the taxonomies is discussed in Chapter 4.5. 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 ...):

(1)  $T_{CEN}$

For example, in the  $T_{CEN}$  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)  $T_{NDC}$

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)

(3)  $T_{TZ}$

In  $T_{TZ}$ , 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 5.2) does not label the state-level concept names, therefore, we modelled the  $T_{TZ}$  with just one line:

(USA Pacific Mountain Central Eastern)

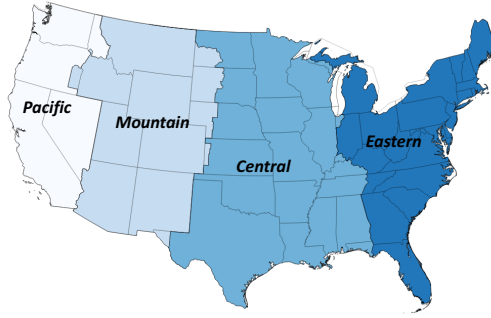


Figure 5.2: Time Zone map (TZ) for contiguous U.S.

### 5.3.2 Step 2: Formulate RCC-5 articulations

In our case studies, we are comparing (1)  $T_{CEN}$  with  $T_{NDC}$ ; and (2)  $T_{CEN}$  and  $T_{TZ}$  for the contiguous United States (48 states and D.C.). The comparisons are fairly straightforward from looking at the maps so that articulations can easily be formulated by any individuals.

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<sup>2</sup> equals (=), includes (>), is\_included\_in (<), overlaps (><) and disjoint (!), the following are the articulations we formulated:

(1)  $T_{CEN}$  vs.  $T_{NDC}$  (49 articulations, all with ‘equal’ relationships; only 5 shown here – the full input file can be found in our online repository (Cheng et al., 2017b):

```
[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)  $T_{CEN}$  vs.  $T_{TZ}$  (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]
```

---

<sup>2</sup>See details about RCC-5 in Chapter 2 and 4.

To demonstrate how we formulated the (2)  $T_{CEN}$ - $T_{TZ}$  articulations, we compare the Time Zone map (Figure 5.2) directly with the Census Regions map in Figure 5.1. We started by looking at one region from  $T_{TZ}$  first, and came up with all the articulations about that region, then move on to the next. For example, we can start with  $T_{TZ}$ .Eastern, and  $T_{TZ}$ .Eastern cannot geographically relate to  $T_{CEN}$ .West, given that one region is situated in the East of the U.S, the other is in the West. Therefore, we can determine that  $T_{CEN}$ .West is disjoint with  $T_{TZ}$ .Eastern.

Next, we can also see from the two maps that the area for  $T_{TZ}$ .Eastern is bigger than that of  $T_{CEN}$ .Northeast, and that the whole  $T_{TZ}$ .Eastern region subsumes  $T_{CEN}$ .Northeast. Therefore, we make the assumption that the whole area of  $T_{CEN}$ .Northeast is included in  $T_{TZ}$ .Eastern. However, not all  $T_{TZ}$ .Eastern is an inclusion of  $T_{CEN}$ .South, only the southern part of  $T_{TZ}$ .Eastern is included in  $T_{CEN}$ .South; likewise, only a few parts around the Michigan Lake Area for  $T_{TZ}$ .Eastern is included in the  $T_{CEN}$ .Midwest. Thus, we can formulate our articulations such that  $T_{CEN}$ .South and  $T_{CEN}$ .Midwest overlaps with  $T_{TZ}$ .Eastern. In the same manner, we can formulate articulations for the Central, Mountain, and Pacific Time Zone areas. In total, there are 12 articulations formed for  $T_{CEN}$  vs.  $T_{TZ}$ .

### 5.3.3 Step 3: Iteratively edit articulations within Euler/X

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  $T_{CEN}$  versus  $T_{NDC}$  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  $T_{CEN}$  versus  $T_{TZ}$  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 5.1 and 5.2, it may not be obvious that there is a

small area in Texas of  $T_{TZ}$ .Mountain that happens to be ‘overlapping’ with  $T_{CEN}$ .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 (whoever asserted the articulations) 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.

## 5.4 Results

The first part of the results is on the  $T_{CEN}$ - $T_{NDC}$  case, while the second part describes the  $T_{CEN}$ - $T_{TZ}$  case. All the files and figures we showed in this chapter is available in Cheng et al. (2017b).

### 5.4.1 Case 1: $T_{CEN}$ versus $T_{NDC}$

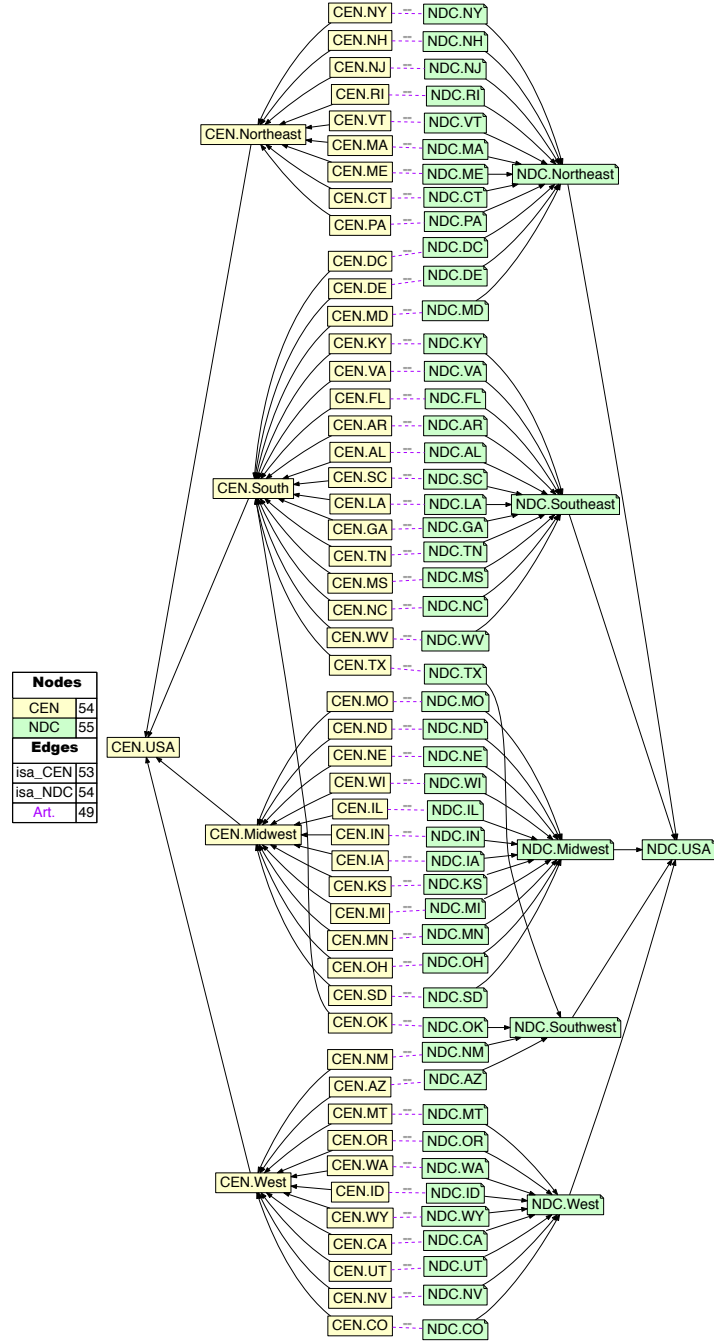


Figure 5.3: Taxonomy alignment problem  $T_{CEN}$  and  $T_{NDC}$

49 input articulations are made between state-level nodes of  $T_{CEN}$  and  $T_{NDC}$  with equivalence.

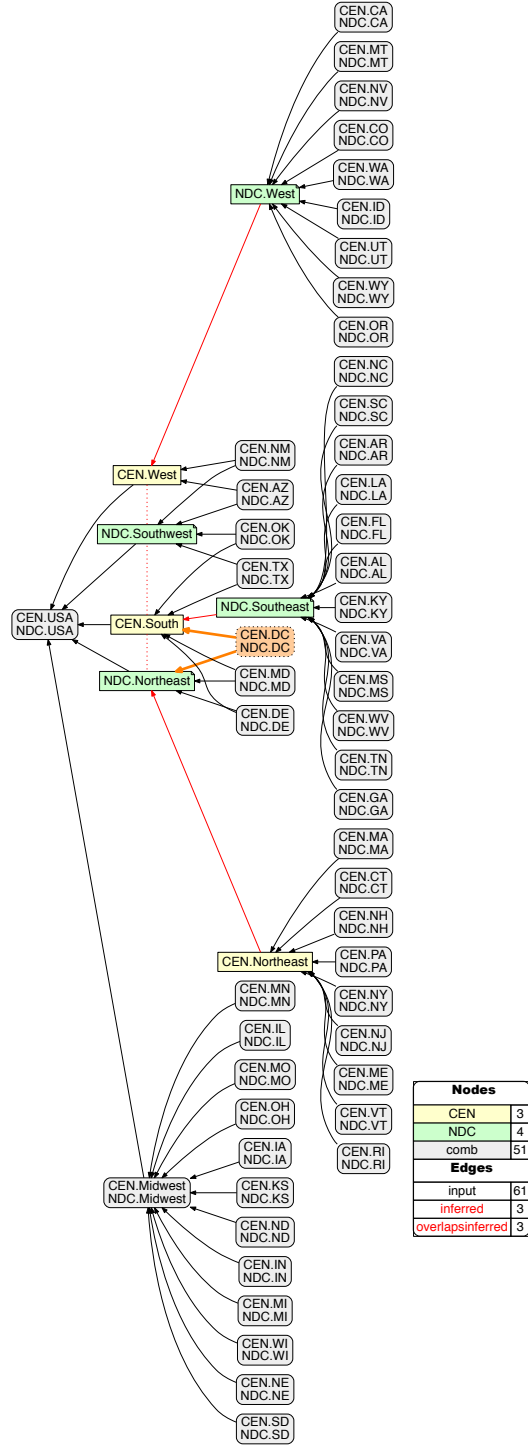


Figure 5.4: One possible world (PW)  $T_3$  reconciling  $T_{CEN}$  and  $T_{NDC}$

## 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 a bottom-up approach example of resolving taxonomy alignment problems, because all the articulations are at the lowest level nodes. Corresponding to Figure 5.1, this use case entails  $T_{CEN}$ ,  $T_{NDC}$ , and the 49 articulations formulated as the input file into the Euler/X tool. Euler/X then can produce the input visualization shown in Figure 5.3. We can easily see that the 48 states and Washington D.C. from either side of the taxonomy are mapped with ‘equal’ signs. The diagram shows the input articulations that we put in, Euler/X has not yet inferred anything for us in this input visualization.

In Figure 5.4, 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  $T_{CEN}$  or  $T_{NDC}$ . Interestingly, the two taxonomies also agree completely on the Midwest.

## Inferred new articulations for region-level alignments

On the upper-level classifications, we can see from the input visualization (Figure 5.3) that the big clusters of the  $T_{CEN}$  and  $T_{NDC}$  corresponding parents ( $T_{CEN}$ .West, South, Northeast, Midwest;  $T_{NDC}$ .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  $T_{NDC}$ .Northeast equal  $T_{CEN}$ .Northeast? Does  $T_{CEN}$ .West include  $T_{NDC}$ .Southwest?

The upper-level alignments were underspecified before Euler/X automatically inferred

the relationships for us. The newly inferred articulations (Figure 5.4) are depicted as red arrows, while the dotted lines tell us which concepts are overlapping. The inferred articulations about the parents ( $T_{CEN}$ .West, South, Northeast, Midwest;  $T_{NDC}$ .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 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 5.4). 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 5.4) 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, region-level. In Figure 5.4, marked in orange dotted boxes and bolded arrows, Washington D.C. is actually included in the South region in  $T_{CEN}$ , whereas in the  $T_{NDC}$ , D.C. is included in the Northeast region.

### 5.4.2 Case 2: $T_{CEN}$ versus $T_{TZ}$

#### **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 5.2 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 3,000 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 5.5 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 5.6. 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 5.6:  $T_{CEN}$  concepts are in yellow boxes, while  $T_{TZ}$  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 5.6. For example, the Pacific Time Zone is properly included in the West Census Region, as shown by the red arrow pointing from  $T_{TZ}.Pacific$  to  $T_{CEN}.West$ .

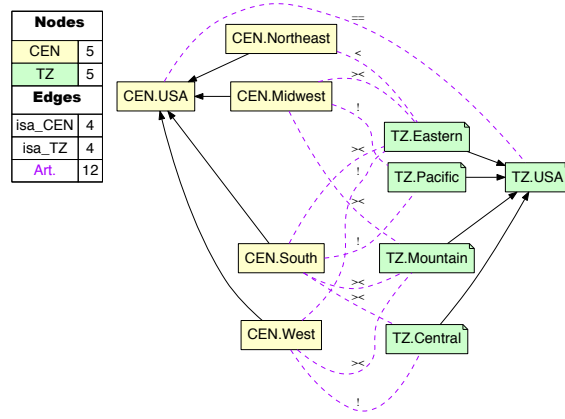


Figure 5.5: Region-level input alignments between  $T_{CEN}$  and  $T_{TZ}$

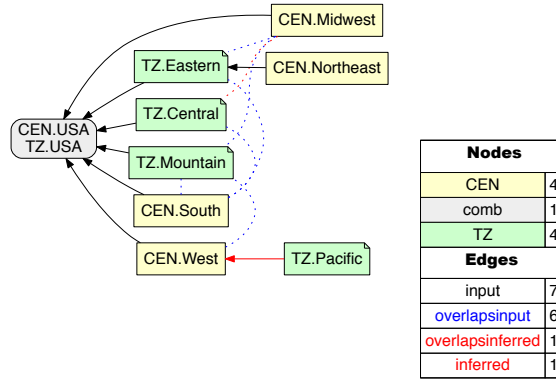


Figure 5.6: One PW for the  $T_{CEN}$  with  $T_{TZ}$  alignment

### Combined concepts solution for region-level alignments

In the  $T_{CEN}$  and  $T_{TZ}$  case, since there are a lot of overlapping relationships (Figure 5.6), 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 5.7), 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  $T_{CEN}.West * T_{TZ}.Mountain$ ) can be counted as properly

included in  $T_{CEN}.West$ . We can also infer from this graph that the intersection area of  $T_{TZ}.Central$  and  $T_{CEN}.Midwest$  ( $T_{CEN}.Midwest * T_{TZ}.Central$ ) is congruent with the Central Time Zone area excluding the Census Region's South ( $T_{TZ}.Central \setminus T_{CEN}.South$ ).

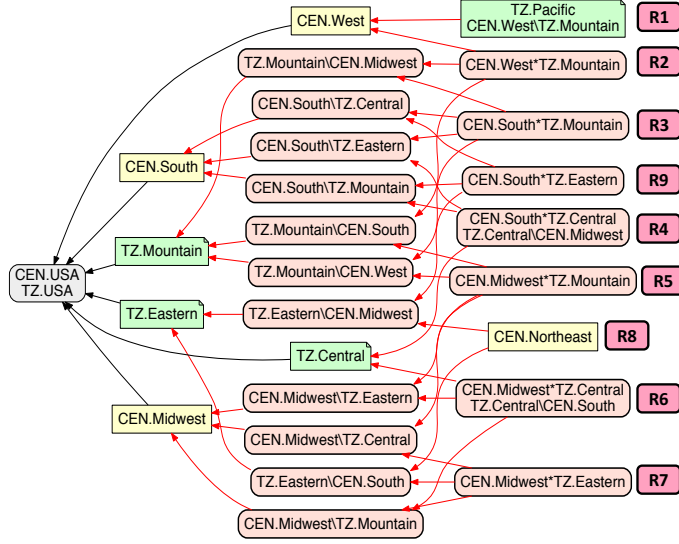


Figure 5.7: Combined concepts solution for  $T_{CEN}$  and  $T_{TZ}$

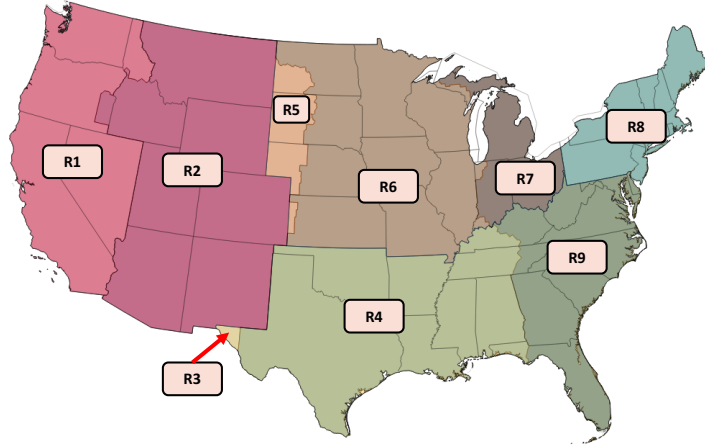


Figure 5.8: A GIS overlay of CEN and TZ maps

This confirms the logically derived solution with 9 new regions

## Comparing Euler solutions with a GIS solution

When overlaying the 4 regions in TZ and the 5 regions in CEN with a geographic information system (GIS), a new map with 9 regions is created (labeled R1-R9 on Figure 5.8). These correspond to the 9 leaf nodes produced by Euler/X (Figure 5.7). 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  $A \setminus B$ .

By using only the merged GIS results, we can see what the new 9 regions are, but there are no terms to describe each regions 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 5.7). We also labelled the corresponding 9 regions of the GIS view from R1 to R9 in Figure 10. 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  $A \setminus B$  (“\”denotes NOT). For instance, we see that R1 is simply  $T_{TZ}.\text{Pacific}$ ; alternatively we can say it is  $T_{CEN}.\text{West}$  without the  $T_{TZ}.\text{Mountain}$  area ( $T_{CEN}.\text{West} \setminus T_{TZ}.\text{Mountain}$ ). Similarly, we can describe R3 as the intersection between  $T_{TZ}.\text{Mountain}$  and  $T_{CEN}.\text{South}$  ( $T_{TZ}.\text{Mountain} * T_{CEN}.\text{South}$ ).

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. From the use cases in these chapter, 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”.

## 5.5 Chapter Summary

In this chapter, we have explored the feasibility of using logic-based RCC-5 articulations to align different regional classifications of the United States. Based on the findings of this chapter, we envision to extend the logic-based taxonomy alignment approach to be used in working with metadata crosswalking in 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 may be able to 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.

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.

In Case 1, we compare the National Diversity Council map with the Census Regions map in a bottom-up approach, in which 49 leaf-level articulations (the 48 contiguous states and Washington D.C.) can be congruently asserted, not in the “absolute crosswalking” sense but using the logic-based RCC-5 relation. On the other hand, in Case 2, we align 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 provide 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 are also able to 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 “ $T_{TZ}.\text{Eastern} * T_{CEN}.\text{South}$ ”). This extended or “hybrid” vocabulary preserves the provenance of input region naming while generating new and more granular label sets needed to identify 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.

# CHAPTER 6

## RECONCILING THE SOCIO-GEOGRAPHIC IN COUNTRY TAXONOMIES

This chapter is an example use case to apply the logic-based taxonomy alignment approach to socio-geographic contexts where the input taxonomies could encode latent assumptions. Inspired by a series of current events in the China-Taiwan conflict on the sovereignty of Taiwan, this chapter aims to examine three relevant taxonomies that list the world entities, which each taxonomy may have its own stance on which geographic entities constitute the world. The three taxonomies are: (1) ISO 3166 for country codes and subdivisions; (2) the geographic regions of the US Department of Homeland Security; (3) the Center Intelligence Agency’s World Fact Book. Our results highlight multiple merge solutions (or possible worlds) for situating Taiwan relative to other neighboring entities, and how we can compare these merged solutions.<sup>1</sup>

### 6.1 Introduction

In the face of heterogeneous standards and large-scale datasets, it has become increasingly important to understand the underlying knowledge structures within complex information systems. These structures may encode latent assumptions that could be susceptible to issues such as ghettoization, bias, erasure, or omission (Adler & Tennis, 2013). Over the years, scholars in the field of Knowledge Organization (KO) have examined these issues in a wide range of research topics, such as in the classification of gender (Olson, 2002), racial groups (Higgins, 2016), and indigenous populations (Littletree & Metoyer, 2015). These

---

<sup>1</sup>This chapter is adapted from Cheng, Y. Y., & Ludäscher, B. (2019). Exploring Geopolitical Realities through Taxonomies: The Case of Taiwan. *NASKO*, 7(1), 77-93 (Cheng & Ludäscher, 2019b).

analyses and others show how misrepresented library classifications tend to provide a one-sided perspective that favors a dominant world view (Adler, 2016; Olson, 2002).

Beyond the KO community, scholars from diverse disciplines have also taken interest in unveiling the presupposed structures in different contexts. In fact, the field of cartography and geographic information systems (GIS) have always been attentive to predetermined conformity behind the production of maps. Rundstrom (1995) has discussed the assimilation of western values on indigenous people’s knowledge in the setting of GIS software and products. Soeller, Karahalios, Sandvig, and Wilson (2016) have developed a tool to detect opaque personalization of online maps that showed different geopolitical country views based on locations. Stewart, Piburn, Sorokine, Myers, and White (2015) have developed a World Spatiotemporal analysis tool that incorporates a myriad of data sources and aims at representing world entities changes over time.

Inspired by a series of current events in the China-Taiwan conflict (Buckley & Horton, 2019; Horton, 2019) on the sovereignty of Taiwan, our research aims to provide a comprehensive and holistic view of multiple ‘truths’. Similar to our prior work (Cheng et al., 2017a; Cheng & Ludäscher, 2018), we apply the logic-based taxonomy alignment approach to reconcile distinct but overlapping taxonomies. We specifically examine three relevant taxonomies that enumerate countries in the world: (1) ISO 3166 country code<sup>2</sup>; (2) the United States Homeland Security geographic regions<sup>3</sup>; and (3) the Center Intelligence Agency’s World Fact Book<sup>4</sup>. We hope that this work can be a first step to demonstrate and reconcile different geopolitical realities through taxonomies.

---

<sup>2</sup>ISO 3166: <https://www.iso.org/iso-3166-country-codes.html>

<sup>3</sup>DHS: <https://www.dhs.gov/geographic-regions>

<sup>4</sup>CIA Fact Book: <https://www.cia.gov/library/publications/the-world-factbook/>

## 6.2 Related Work

### 6.2.1 Embedded Structure and Knowledge Organization Systems

The impact of embedded beliefs within knowledge organization systems (such as taxonomies, library classifications, thesaurus, ontologies, etc.) can be seen from two dimensions: misrepresentation of subjects; and social, cultural, or political influences.

Works by Olson (2002, 2017) demonstrate how marginalization and systematic biases are manifested in classification systems by examining cases on how females are often wrongly classified in racial groups and occupations in library classifications. Likewise, discussions on how racial representation in Dewey Decimal Classifications (DDC) (Higgins, 2016), or how indigenous peoples knowledge and cultures are classified in major classifications (Green, 2015; Hajibayova, Buente, Quiroga, & Valeho-Novikoff, 2016; Littletree & Metoyer, 2015; Webster & Doyle, 2008) have all reached similar conclusions that the unseen system biases are already entrenched in society (Mai, 2016).

Moreover, classification is heavily influenced by social, cultural, or political matters. As described in Bowker and Star (2000), this was manifested in the apartheid South Africa in the 1950s, where everyone was classified into racial groups. Consequently, this classification system segregated where people can live, work, or even go to school.

Suggestions have been made to improve the status quo: for instance, Bowker and Star (2000) suggest that we should uncover the invisible infrastructures and retrace the original design intentions behind classifications, while Tennis (2012) proposes that the designer of a bibliographic classification should anticipate future changes and ‘design for change’. While most of these studies aim to raise awareness of the entrenched, hidden structure within KOS from a critical theory perspective, advocating changes can be difficult without empirical analysis or realistic actions.

### 6.2.2 Geopolitical Realities

Behind the creation of maps lies a number of social, political, and power assumptions that a nation, or a group of nations make about their boundaries and bordering territories. Geopolitical mapping is not simply mathematical or scientific reflection of the “real world”, but rather, “technologies of power” (Callahan, 2009) that can shape a certain discourse about a nation. The subtle exhibitions of political messages on maps show its social influence as well as the power dynamics: omissions, or silences on maps only amplify the fact that elite groups exist and they are using maps to ‘promote an uneven dialogue’ to the weak (Crampton, 2001; Harley, 1988, 1992).

Extended empirical works confirm that maps, or modern days GIS systems are cultural and political products that overlook minority values, and advocate the geopolitical realities of those in power. Rundstrom (1995) examines the development of GIS tools and concludes that knowledge of the indigenous people may have been undermined due to the tacit social influences of the tool. Soeller et al. (2016) have developed the MapWatch tool to detect personalized views on Google Maps based on the users’ location. The personalized views happen especially at places where there are border conflicts (e.g., Russia, Ukraine and Crimea, etc.). Soeller et al. (2016) concluded that their scrutiny to the hidden structure behind maps is not enough; a call for transparency on the border conflicts is needed. Further, Stewart et al. (2015) have developed a World Spatiotemporal analysis project to aggregate all different data sources of world entities with an ontology-aided data model documenting entity name changes over time. Their intention to integrate various data sources to disclose disparate features in these sources also entails a cellophane-like, permeable structure.

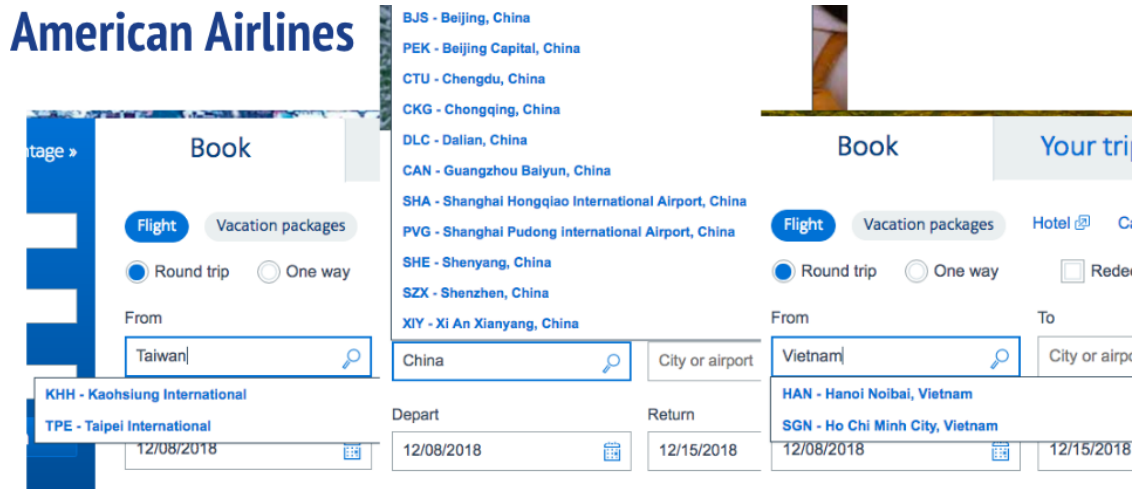


Figure 6.1: Major U.S. Airline web interface.

These empirical studies have examined controversial cases such as the Indo-Pakistan border conflict, South China Sea, or how eastern Europe split from Yugoslavia to Croatia, Slovenia, and more (Soeller et al., 2016; Stewart et al., 2015), but empirical studies on other disputed regions such as Taiwan-China are yet to be discussed. Historically, Taiwan and China have been in a longstanding dispute over sovereignties and there is yet to be a formal arrangement between the two geobodies. Recent decisions from several global companies' websites to drop the name "Taiwan" but keep only the capital city "Taipei" (Figure 6.1) as an indicator of the territory suggests there may be renegotiation of power between the two regions (News, 2018; Wee, 2018). Shown in Figure 6.1, a search on the keyword "Taiwan" on American Airlines' website would only return the cities Kaohsiung and Taipei; a search on China will return major cities in China but not Taipei; a search on other countries such as Vietnam will return "Hanoi, Vietnam" with both the city and country names included. Moreover, most of the empirical studies mentioned above have explore these geopolitical issues at the actual map-level, but a close inspection on how KO systems such as taxonomies can make a impact is not yet seen in these studies.

## 6.3 Taxonomy Alignment Problems

To demonstrate how logic-based taxonomy alignment can aid with reconciling conflicting boundary perspectives, we aim at aligning the vocabularies used for representing the world's entities, specifically zooming into the areas in Asia. Adopting the terminology used in Stewart et al. (2015), we also use the term *entities* to avoid the disagreement among different data sources on what constitutes a *nation*, *country*, or *state*.

The figure consists of four screenshots of the ISO Online Browsing Platform (OBP) interface, each showing search results for a specific entity. The interface includes a search bar, a sidebar with navigation links (Standards, Collections, Publications, Graphical symbols, Terms & Definitions), and a table of results.

**Search 1: Macao**

2 results for Macao

| English short name | French short name | Alpha-2 code | Alpha-3 code | Numeric |
|--------------------|-------------------|--------------|--------------|---------|
| China              | Chine (la)        | CN           | CHN          | 156     |
| Macao              | Macao             | MO           | MAC          | 446     |

**Search 2: Hong Kong**

2 results for Hong Kong

| English short name | French short name | Alpha-2 code | Alpha-3 code | Numeric |
|--------------------|-------------------|--------------|--------------|---------|
| China              | Chine (la)        | CN           | CHN          | 156     |
| Hong Kong          | Hong Kong         | HK           | HKG          | 344     |

**Search 3: Japan**

1 result for Japan

| English short name | French short name | Alpha-2 code | Alpha-3 code | Numeric |
|--------------------|-------------------|--------------|--------------|---------|
| Japan              | Japon (le)        | JP           | JPN          | 392     |

**Search 4: Taiwan**

2 results for Taiwan

| English short name         | French short name          | Alpha-2 code | Alpha-3 code | Numeric |
|----------------------------|----------------------------|--------------|--------------|---------|
| China                      | Chine (la)                 | CN           | CHN          | 156     |
| Taiwan (Province of China) | Taiwan (Province de Chine) | TW           | TWN          | 158     |

Figure 6.2: ISO's presupposed hierarchical structure for the entities

Interpretations on the concepts and edges of the taxonomies is discussed in Chapter 4.5.

The three taxonomies we examined in this chapter are:

- (1) **ISO:** ISO 3166 Country Codes<sup>5</sup> is maintained by 14 different organizations. ISO 3166 reflects the United Nation's (UN) perspective of its member states. For an entity that is not a UN member state, it is still searchable on the ISO online platform, but it is not be listed as a *country*. Alternatively, it may be listed as a sub-entity of a particular country. For example, performing a search on the ISO platform on *Macao* (or Macau), *Hong Kong*, or *Taiwan* will result in two records for each, indicating that *China* is the *parent* of these entities, whereas performing a search on other entities such as *Japan* will only result in one record. Further, another layer of inconsistency is seen for the name shown for *Taiwan*. Unlike the naming for *Macao* directly shown as *Macao*, ISO has to re-amplify that *Taiwan* is a "Province of China". See Figure 6.2 for ISO's implicit assumptions on sub-entities.
- (2) **DHS:** This is the taxonomy used in the United States Department of Homeland Security to identify the world's geographic regions<sup>6</sup>. The taxonomy itself is a flatter structure, meaning that there are no nested entities or sub-entities. Even for obvious cases where there might exist a parent country (e.g., *Macao*), DHS considers considered it as a disjoint entity from China. It is not explicitly stated whether this 'flat' structure means that each entity is its own *independent state*, or it merely means that the *geographic area* of these entities are disjoint from each other.
- (3) **CIA World Fact Book:** the Central Intelligence Agency World Fact Book<sup>7</sup> consists of a comprehensive list, history, and facts of the world's entities. The World Fact Book also contains a hierarchy of entities. When a certain entity is part of an independent state, that entity will contain one note on its *Dependency Status*<sup>8</sup> and its *Independence*

---

<sup>5</sup>ISO 3166: <https://www.iso.org/iso-3166-country-codes.html>

<sup>6</sup>DHS: <https://www.dhs.gov/geographic-regions>

<sup>7</sup>CIA Fact Book: <https://www.cia.gov/library/publications/the-world-factbook/>

<sup>8</sup>Dependency Status: "This entry describes the formal relationship between a particular nonindependent entity and an independent state."

Table 6.1: Comparison of  $T_{ISO}$ ,  $T_{DHS}$ , and  $T_{CIA}$ 

| Source                     | Created/maintained by      | # of regions | # of entities |
|----------------------------|----------------------------|--------------|---------------|
| ISO 3166                   | 14 different organizations | 5            | 250           |
| Dept. of Homeland Security | United States              | 8            | 232           |
| CIA World Factbook         | United States              | 11           | 267           |

\*Data collected May, 2019.

status will be listed as *none*. For example, **Macau** is noted as “special administrative region of the People’s Republic of China” on its *Dependency Status* with the section on *Independence* noted as “*none*”. As a counter example, **Japan** does not have an extra section on *Dependency Status*, and for its *Independence* section, “3 May 1947” is listed as the independence date. As for the case of **Taiwan**, there are neither sections on *Dependency Status* nor sections on *Independence*, though *Administrative divisions* and *National Holiday* different from China, Hong Kong, and Macau are described.

A comparison of the three taxonomies is shown in Table 6.1.

In this chapter, we employed pairwise alignments with three taxonomies ( $T_{ISO}$  vs.  $T_{DHS}$ ,  $T_{ISO}$  vs.  $T_{CIA}$ , and  $T_{CIA}$  vs.  $T_{DHS}$ ). Two layers of interpretations are specified and explicitly explained here:

(1) *The taxonomies* – we model each data source into a tree-like structure to the best of our knowledge or as reflected in the literature. This is demonstrated in the prior section describing the three taxonomies.

(2) *the articulations*: the articulations given between each concept in each pair of taxonomies are specified by us. For every use case, since the relation between Taiwan and China are undefined (and should not be defined by us), we do not specify any articulation between the two entities. Specifically, we left the nodes  $T_{ISO}.China$ ,  $T_{ISO}.Taiwan$ ,  $T_{DHS}.China$ ,  $T_{DHS}.Taiwan$ ,  $T_{CIA}.China$ ,  $T_{CIA}.Taiwan$  **unmapped** from  $T_1$  to  $T_2$ . Other regions in Asia are marked as equivalent in  $T_1$  and  $T_2$  if they share same or similar names (e.g.,  $T_{ISO}.VietNam$  equals  $T_{DHS}.Vietnam$ ). To mark all other entities in Asia as equivalent is a bold assumption, however, this is one way to aid us in focusing solely on the China-

Taiwan entities. Further, we are keeping an open stance on the geopolitical issue of these entities by leaving Taiwan and China unmapped. This can also help us see the different modeling situations unfolded by the logic-based alignment approaches.

## 6.4 Results

### 6.4.1 Use Case 1: $T_{ISO}$ - $T_{DHS}$

#### **Input Taxonomies**

For  $T_{ISO}$ , because ISO has inconsistent naming and a presupposed hierarchical structure in which they give Taiwan the name “Taiwan (Province of China)” (see Figure 6.2), in this modeling situation, we then assume that China has four children: HongKong, Taiwan, Macao, and ChinaOther (to represent mainland China and other regions that may be of China’s). As for  $T_{DHS}$ , since it is a flat structure with every entity being its own independent state, we listed several entities in Asia and described them all as siblings instead of children of a super-entity. Figure 6.3 is the visualization graph of the input alignment.

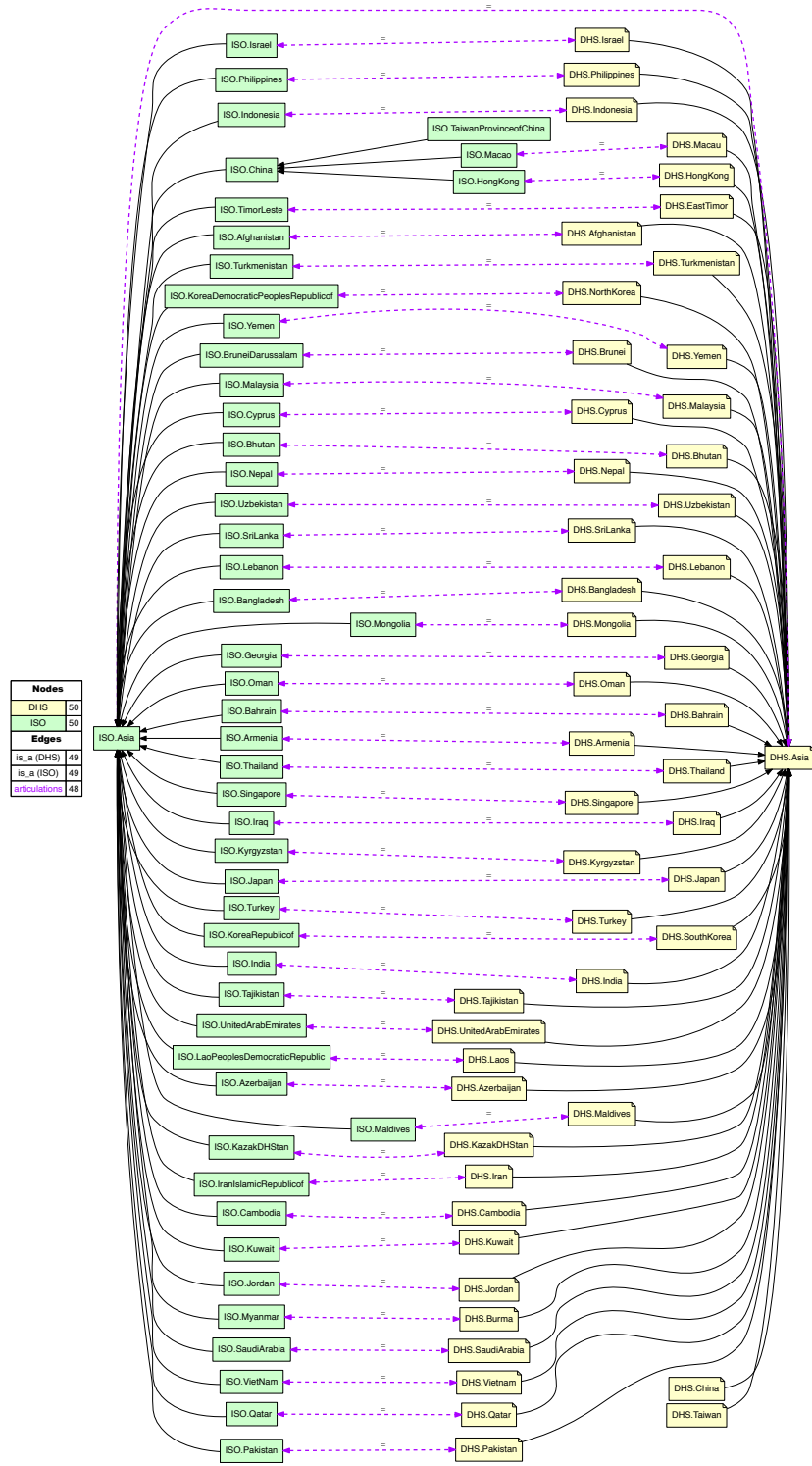


Figure 6.3: Input alignment for  $T_{ISO}-T_{DHS}$

Since there are many geographic entities in Asia, it can potentially obscure the viewing of the visualizations of the taxonomies. Here we also provide an shortened version of the entities in Asia as an aid. An abridged example of the Euler/X (or LeanEuler) input text file is as follows:

```
taxonomy ISO ISO3166
(Asia China VietNam AsiaOther)
(China HongKong Macao Taiwan ChinaOther)

taxonomy DHS HomelandSecurity
(Asia China ChinaOther HongKong Macau Taiwan Vietnam AsiaOther)

articulation ISO-DHS ISO-Homeland
[ISO.Asia equals DHS.Asia]
[ISO.HongKong equals DHS.HongKong]
[ISO.Macao equals DHS.Macao]
[ISO.VietNam equals DHS.Vietnam]
[ISO.ChinaOther equals DHS.ChinaOther]
[ISO.AsiaOther equals DHS.AsiaOther]
```

The shortened version of the input taxonomy can be visualized by Figure 6.4.

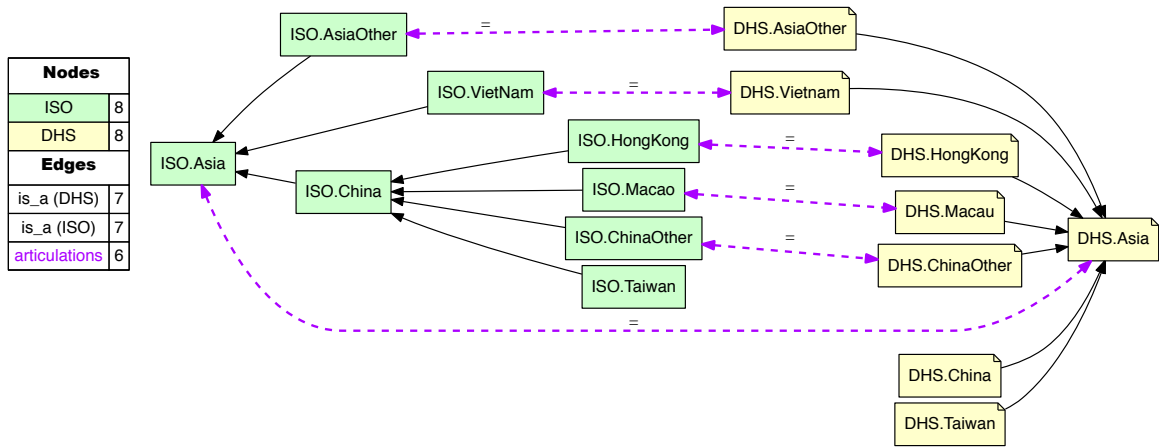


Figure 6.4: Abridged input alignment for  $T_{ISO}-T_{DHS}$

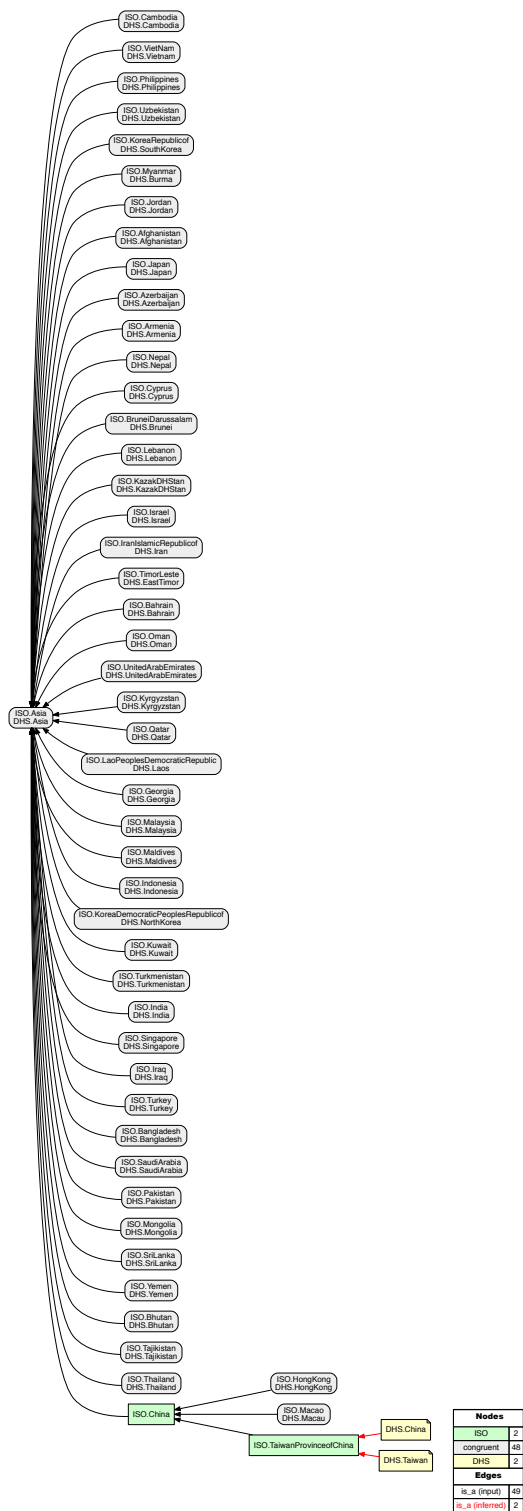


Figure 6.5: One possible world for  $T_{ISO}-T_{DHS}$  (full)

### $T_{ISO}$ - $T_{DHS}$ : merged $T_3$ possible world

The TAP of  $T_{ISO}$  and  $T_{DHS}$  resulted in only one merged, unique possible world. The full lattice of the possible world can be seen in Figure 6.5. Equivalent nodes (marked in grey round boxes) meant that those regions are reconciled and viewed as congruent; red arrows are the inferred relations from Euler/X. The merged possible world preserves the information from both taxonomies with extra information on how to make the Taiwan-China nodes consistent and logical in a unique world. In this world, both  $T_{DHS}$ .China and  $T_{DHS}$ .Taiwan are children of  $T_{ISO}$ .Taiwan, but  $T_{ISO}$ .Taiwan is a child of  $T_{ISO}$ .China (Figure 6.6). This created an interesting PW where one taxonomy  $T_{DHS}$  has become sub-taxonomy of  $T_{ISO}$  when the nodes are not congruent (equivalent)(grey boxes). This means that there are parts ( $T_{DHS}$ .China and  $T_{DHS}$ .Taiwan) of  $T_{DHS}$  that should be viewed as children of  $T_{ISO}$ .

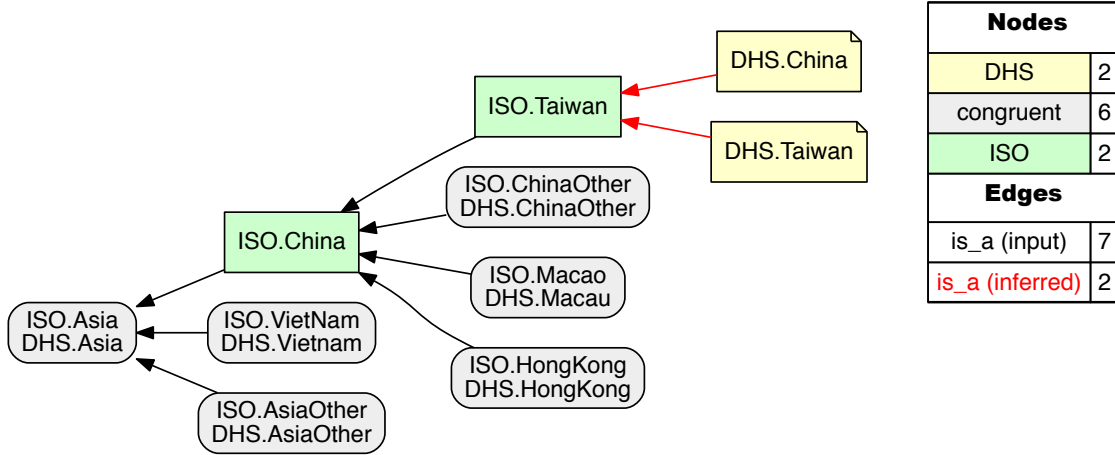


Figure 6.6: One possible world for  $T_{ISO}$ - $T_{DHS}$  (abridged)

#### 6.4.2 Use Case 2: $T_{ISO}$ - $T_{CIA}$

Similar modeling situation has been devised for comparing  $T_{ISO}$  and  $T_{CIA}$  World Fact Book taxonomies. Aligning all the entities in Asia is sometimes illegible, therefore for the com-

parison from now on, we will focus only in parts of Asia with fewer entities.

## Input Taxonomies

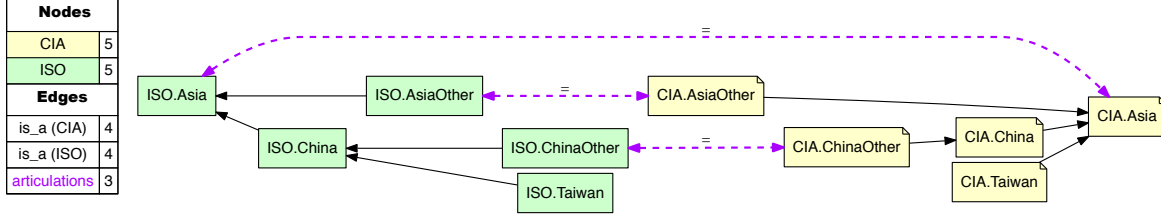


Figure 6.7: Input alignment for  $T_{ISO}-T_{CIA}$ .

For use case 2 we have  $T_{ISO}$  and  $T_{CIA}$ . Again, we have left the articulations for specifying either China or Taiwan unmapped, while marking the rest as equivalent. For this use case we pinpointed solely on Taiwan and China, and the other entities are all combined to become one entity *AsiaOther* (Figure 6.7).

### $T_{ISO}-T_{CIA}$ : merged $T_3$ possible world

The merged possible world looked as if they are a smaller subset of the result for the  $T_{ISO}-T_{CIA}$  alignment (Figure 6.6). However, because both  $T_{ISO}$  and  $T_{CIA}$  have implicit hierarchical structures to specify entities, the merged  $T_3$  PW had subtle differences from the previous use case.  $T_{CIA}.China$ ,  $T_{CIA}.ChinaOther$  and  $T_{ISO}.ChinaOther$  are considered the same; while  $T_{ISO}.Taiwan$  is equivalent to  $T_{CIA}.Taiwan$ .  $T_{ISO}.China$  is the super-entity of both  $T_{CIA}.China$  and  $CIA.Taiwan$  (or  $T_{ISO}.Taiwan$ ). The two children are at equal level rather than one being superior than the other (See Figure 6.8).

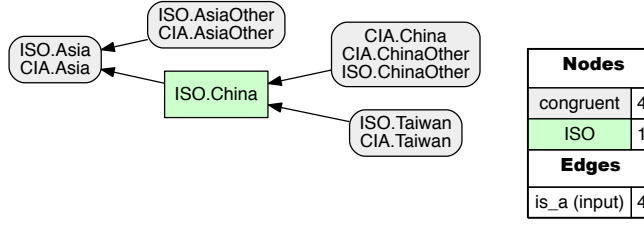


Figure 6.8: One possible world for  $T_{ISO}-T_{CIA}$

#### 6.4.3 Use Case 3: $T_{CIA}-T_{DHS}$

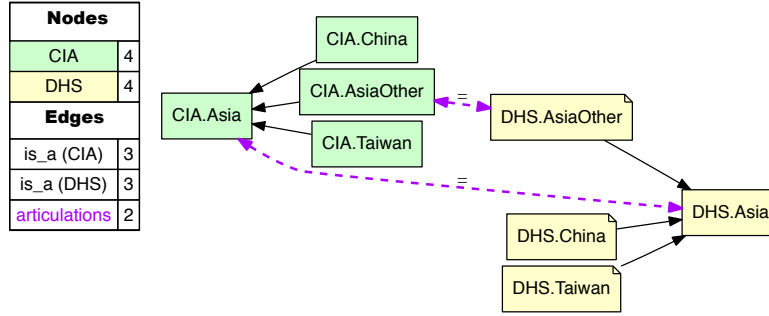


Figure 6.9: Input alignment for  $T_{CIA}-T_{DHS}$

### Input taxonomies

For use case 3, we compare  $T_{CIA}$  with  $T_{DHS}$ . We have once again focused only on the two entities Taiwan and China, with all other entities combined as AsiaOther. Articulations between entities from  $T_{CIA}$  and  $T_{DHS}$  are marked as equivalent except for Taiwan and China. The interesting part about this use case is that the two taxonomies look identical in terms of structure, having a two-level structure with a root entity *Asia* and three children *Taiwan*, *China*, *AsiaOther*. Figure 6.9 shows the symmetry of the two taxonomies.

$T_{CIA-T_{DHS}}$ : merged  $T_3$  possible worlds

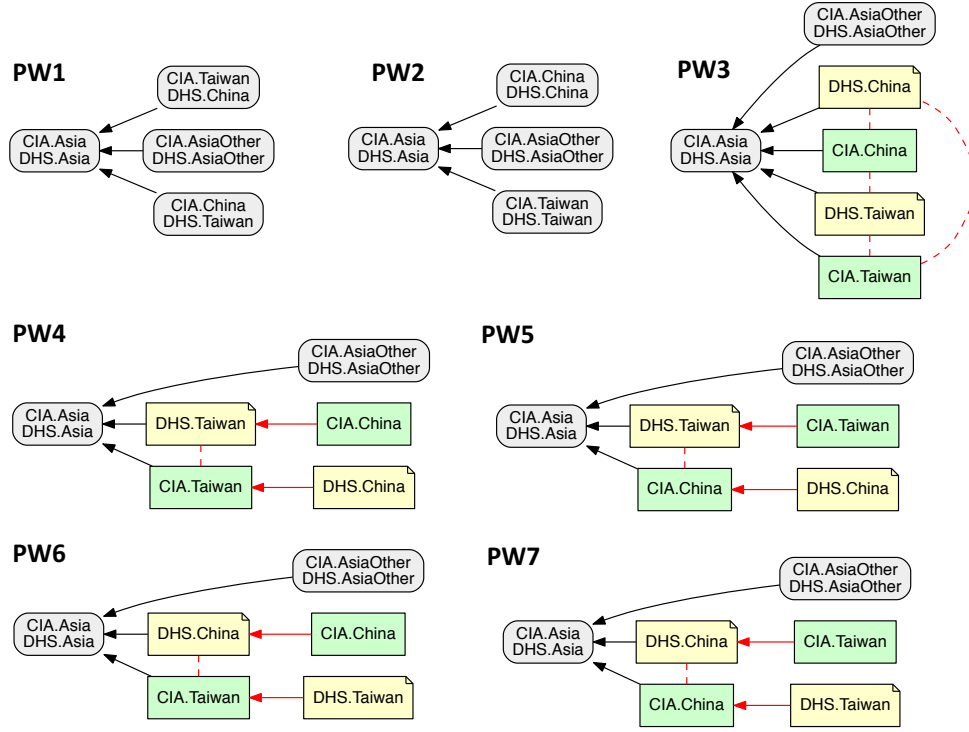


Figure 6.10: Seven possible worlds for  $T_{CIA-T_{DHS}}$

The seven worlds possibly represent different perspectives on the China-Taiwan relationship.

This particular use cases presented different results than the previous use cases. Due to open, *missing* articulations among  $T_{CIA.China}$ ,  $T_{CIA.Taiwan}$ ,  $T_{DHS.China}$ ,  $T_{DHS.Taiwan}$ , the results held room for ambiguities. This has resulted in seven possible worlds – 7 different ways to merge the two taxonomies. Interestingly, these PWs may coincides with different social worlds that comprise different governmental (or political party's) perspectives (See Figure 6.10). Based on our interpretations from BBC (2021); Brown (2022); Charney and Prescott (2000); Ross (2002), the PWs correspond to:

**PW1** may be how the rest of the world sees Taiwan-China (T-C) relationship – they might not even be aware of the conflicts.

**PW2** may be how the Democratic Progressive Party (DPP) views the T-C relationship – Taiwan and China are separate countries.

**PW3** may be how neighboring governments view the T-C relationship – the two regions overlap in multiple ways and it is difficult to state what is true (red dotted lines means overlaps).

**PW4** may correspond to political extremists' view in Taiwan.

**PW5 and PW6** may be how the Kumintang (KMT) party view the situation – Taiwan has its own individual identity but still overlap partially with China.

**PW7** may correspond to how the China government view this relationship – Taiwan is part of China.

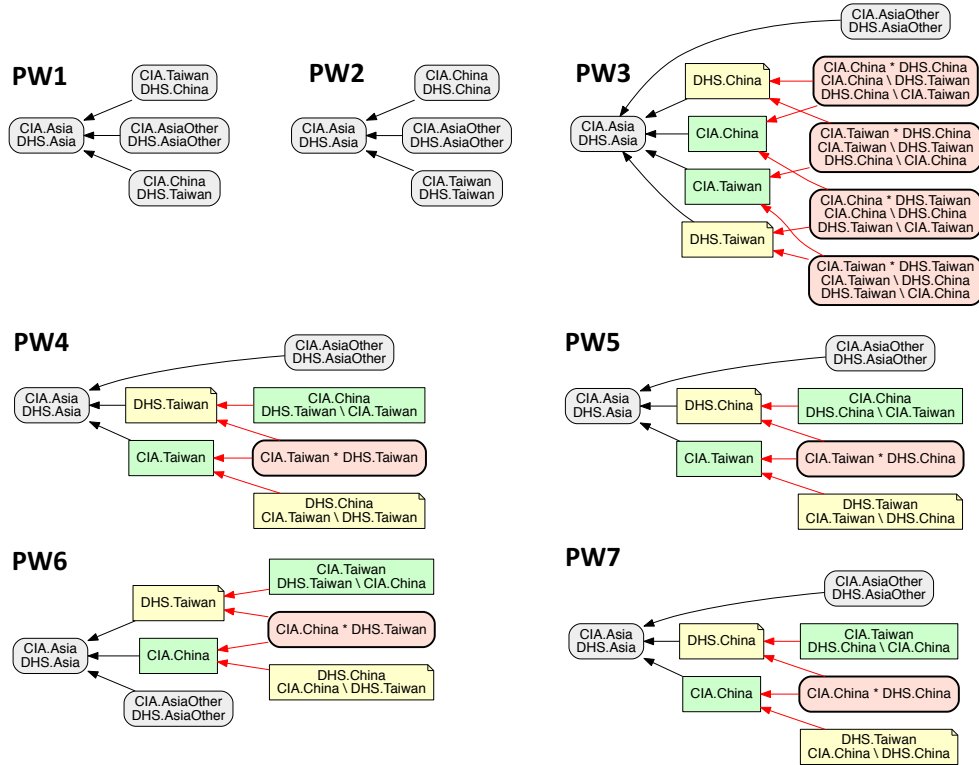


Figure 6.11: Combined concept solution for  $T_{CIA}-T_{DHS}$

We can further examine the overlapping relations (red dotted lines) in Figure 6.10 and amplify those relations. Figure 6.11 shows this further dissection on how new regions (new entities) can be created in these 7 PWs. Pink boxes represent new names for the entities; The asterisk sign means *intersection* of two entities; the slash sign means *minus*. For example, in PW6,  $T_{DHS}.China$  can also be expressed by  $T_{CIA}.China$  *minus* the entity  $T_{DHS}.Taiwan$ ; similar expressions can be said for  $T_{CIA}.Taiwan$ . Further, there exists a new entity  $T_{CIA}.China$  *intersects*  $T_{DHS}.Taiwan$  that is included in *both*  $T_{CIA}.China$  and  $T_{DHS}.Taiwan$  (Pink box in PW6 of Figure 6.11). More complex situations are seen in PW3 with four new entities, each has two parent nodes. To make sense of these new entities, for instance, the pink node in the bottom has three different ways to express the same entity: “ $CIA * DHS.Taiwan$ ” OR “ $CIA.Taiwan \setminus DHS.China$ ” OR “ $DHS.Taiwan \setminus CIA.China$ ”.

These newly created entities (pink boxes) might look complicated, but it is informative to specify which part is *truly* overlapping with which other entities. Moreover, this also shows that in every PW, the modeling situations for Taiwan and China perhaps cannot be as incisive given the historical ties and entangled relationships. All the artifacts, including the input taxonomies as well as the output visualizations, used and produced by this chapter can be freely accessed from Cheng and Ludäscher (2019a).

## 6.5 Chapter Summary

This chapter is a first attempt at looking into the disputed geographic regions and sovereignties in the world, focusing on the case of Taiwan. The three taxonomies chosen in this chapter serve different purposes: the ISO standard (ISO) conforms with the United Nation’s viewpoints of the world with a more reserved stance on its member stances; the U.S. Homeland Security (DHS) taxonomy provides a more liberal, flat worldview; and the CIA World Fact Book (CIA) strikes a balance of the former two taxonomies.

We have employed the logic-based, taxonomy alignment approach to align three use

cases –  $T_{ISO}$  vs.  $T_{DHS}$ ,  $T_{ISO}$  vs.  $T_{CIA}$ , and  $T_{DHS}$  vs  $T_{CIA}$ . The results of these three use cases provide different modeling situations for the two entities (Taiwan and China) and our taxonomy alignment results (specifically the use case on  $T_{DHS}$  vs.  $T_{CIA}$ ) interestingly coincides with different perspectives in the real-world. The merged possible worlds  $T_3$  in these use cases may be viewed as a *new taxonomy* that integrated multiple sources, and the *PWs* may then later become a new input of a new taxonomy alignment problem.

Latent assumptions within information systems is not unusual. Despite that there are multitudinous taxonomies about the world’s entities, a copious existing information systems endorse the ISO 3166 country codes. For instance, the Library of Congress (LOC) <sup>9</sup> adopts the ISO country code for both the MARC 21 standard <sup>10</sup> and the LOC classifications <sup>11</sup>. The MARC 21 code list makes subtle distinction between *China* and *China (Republic: 1949–)*, where the former is specifically pointing to the mainland, the latter is meant for Taiwan. On the other hand, the LOC classification conforms entirely to the ISO standard and lists Taiwan as a sub-division of China. The reason LOC adopts the ISO country code is not explicitly manifested, and how such choice may lead to further unexpected, unforeseen actions or consequences in information systems is out of the scope of this research.

However, through the lens of taxonomies and taxonomy alignment, it is exemplified that geopolitical realities exist in knowledge structures, and we believe making these assumptions and choices explicit is crucial in this information age. Moreover, to *raise awareness* that there are hidden structures within a KOS is only the first step, to *present a more holistic view* on the possible worlds of these geopolitical realities simultaneously will be inevitable.

Whether Taiwan is an independent nation or part of China still remains a moot point. Despite the close ties we might have with Taiwan, we have limited any personal opinions on the Taiwan and China relation in this research. Yet, we have attempted on making any interpretations or modeling situations of the use cases unequivocal in order to keep a open

---

<sup>9</sup>LOC country code statement: <https://www.loc.gov/standards/sourcelist/country.html>

<sup>10</sup>MARC Code list for countries: [https://www.loc.gov/marc/countries/countries\\_name.html#c](https://www.loc.gov/marc/countries/countries_name.html#c)

<sup>11</sup>LOC classifications: [https://www.loc.gov/aba/cataloging/classification/lcco/lcco\\_d.pdf](https://www.loc.gov/aba/cataloging/classification/lcco/lcco_d.pdf)

stance and embrace different opinions on this issue.

In a longer term, future work of this study on looking at geopolitical realities in taxonomies or other knowledge organization systems will attempt to incorporate more data sources, align other entities in the world, and discuss perceived boundaries between minority groups and the authorities. We believe this line of work can be of great contributions to demonstrate different geopolitical realities in taxonomies and KOS as well as to give voices to the silenced values.

## CHAPTER 7

# RECONCILING THE SOCIO-GEOGRAPHIC IN INDIGENOUS PEOPLES' TRIBES

In this chapter, we apply the logic-based taxonomy approach to socio-geographic contexts pertaining to groupings of people based on geographic location. Specifically, we examine the use case of Taiwanese indigenous peoples' tribes classifications and the indigenous constituencies (voting groups) of the legislature. We compare four finer-grained taxonomies for indigenous people with the coarse-grained indigenous peoples' electoral constituencies that only recognize two regions (Lowland, Highland). The results of the alignment address our question on how to preserve the information on the input taxonomies. This further demonstrates that the preserved, co-existing information from both taxonomies in the merged solution can help with representing different perspectives in unity.<sup>1</sup>

### 7.1 Introduction

Over the years, information science professionals have been studying biases in bibliographic classifications and knowledge organization systems (KOS). The robustness of classifications has been examined in diverse measures, ranging from the representation of race (Higgins, 2016), gender (Olson, 2002), ethnic minorities (Hajibayova et al., 2016), to indigenous peoples (Littletree & Metoyer, 2015). Many have concluded that bibliographic classifications such as the Dewey Decimal Classification, Library of Congress Subject headings, or similar, are susceptible to a systematic bias of a dominant, euro-centric perspective (Adler, 2016; Kam,

---

<sup>1</sup>This chapter is adapted from 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) (Cheng & Ludäscher, 2020b).

2007; Lilley, 2015; Olson, 2002; Webster & Doyle, 2008).

While efforts have been made in uncovering latent assumptions about minorities in bibliographic classifications, institutionalized knowledge of geopolitical taxonomies are difficult to unwind. For instance, the public’s knowledge on the categorization of indigenous people’s tribes is highly informed by the authorities (Jarvis, 2017; McLaughlin, 2019). This knowledge is deeply entrenched and may be woven into political discourses about indigenous peoples’ territories, and even, voting rights (Biho, n.d.; Bowman, 2021). Moreover, classification and organization of indigenous knowledge have always been challenging. Many resolved to create a classification serving specifically for indigenous knowledge (Kam, 2007; Lilley, 2015; Littletree & Metoyer, 2015), but linking the lesser-known classifications to mainstream classifications usually warrants further investigations.

Drawing inspirations from the 2020 Presidential and Legislature Election in Taiwan, this chapter examines the taxonomies of the Taiwanese indigenous peoples’ tribes and the indigenous constituencies (voting groups) (Templeman, 2018) of the legislature representatives from the period of 2000 to 2020. This includes comparing (1) the division of multiple tribes into dichotomous voting groups, historically created during the Japanese colonization period; and (2) the evolving classification of the federally recognized indigenous tribes.

The current design for the indigenous legislature representatives is based on a reserved seat system. Out of the 113 legislators, 6 are reserved to be indigenous candidates. Out of the 6 indigenous legislators, 3 are reserved for “Highland” candidates, the other three are for “Lowland” candidates (Templeman, 2018). Whereas the government may favor the Lowland-Highland dichotomous indigenous constituencies, the 16 indigenous tribes may each want to have their own legislator. The goals of this chapter therefore are: first, to unravel the implicit assumptions of KOS in more ubiquitous, commonsense taxonomies; second, to reconcile these different perspectives by using a logic-based approach to provide co-existing, pluralistic opinions.

The four taxonomy alignment problems we examine in this chapter are: (1) binary

constituencies (Lowland-Highland (LH)) versus recognized tribes over time; (2) LH versus the recognized tribes in present time; (3) LH versus the possible other tribes in Taiwan; and (4) LH versus a re-scaled modeling of the tribes based on population. The attempt of this chapter is not to discuss voting rights, but rather, to demonstrate the hidden, entrenched biases that hinders the expansion of these taxonomies. We hope to extend conversations regarding how the importance of considering multiple perspectives prior to policy-making.

## 7.2 Related Work

### 7.2.1 Bias in Knowledge Organization Systems

Prior literature states that biases are prone to be found in knowledge organization systems (KOS) such as taxonomies, bibliographic classifications, or ontologies (Adler & Tennis, 2013; Bowker & Star, 2000; Clarke & Schoonmaker, 2019; Fox, 2016; Mai, 2016; Zhitomirsky-Geffet, 2019). Stemming from Olson (2002)’s seminal work *The Power to Name* and many other studies by the author, Mai (2016) discusses the systematic biases in Dewey Decimal Classifications (DDC) and Library of Congress Subject Headings (LCSH), and how Olson’s work unraveled the incompetency of these KOS in terms of describing marginalized groups. Expanding from Olson’s research landscape, Fox (2016) further explores cases of intersectionality in KOS, and concludes that the linearity and mutual exclusiveness of classes (or subjects) in bibliographic classification make it difficult to portray identities belonging to more than one marginalized groups.

Foundational work by Bowker and Star (2000) explains how hidden infrastructure in KOS can be socially, culturally, or politically influenced. According to Adler and Tennis (2013), these latent assumptions, whether “intentional or accidental”, can cause problems such as ghettoization, exceptionalism, erasure, or omission to not only individuals, but also communities and nations at large. As exemplified in Higgins (2016), the representation of

Asian Americans in DDC as “perpetual foreigners” rather than Americans, suggest that the Asian American community is invisible to the collective memories of all Americans. Further, when temporal factors are considered, impact of biases in KOS on a nation is more apparent. For instance, geopolitical taxonomies such as maps may be susceptible to silencing voices of the weak (Crampton, 2001). Attempts to mitigate geopolitical biases of country name changes over time are seen in Stewart et al. (2015), in which the authors developed an ontology-based model to formally describe the joint or split event of a nation. Similarly, in a panel discussion by Adler et al. (2013), the temporal changes of KOS are discussed in terms of schema expanding, evolving, and reparations.

Some scholars offer solutions to incorporate more diverse perspectives in KOS. For example, Clarke and Schoonmaker (2019) systematically examine the accessibility of metadata standards pertaining to minorities or marginalized groups; and Zhitomirsky-Geffet (2019) develops an ontology-based model that interlinks different perspectives. Most of the studies to date have called for inclusiveness of marginalized groups and voices in KOS. However, more approachable cases outside of the information science domain are needed to raise true awareness of these embedded biases in KOS and complex systems to wider audiences.

### 7.2.2 Indigenous Knowledge Organization

There is a growing body of literature on indigenous knowledge organization (Adler, 2016; Doyle, 2013; Duarte & Belarde-Lewis, 2015; Green, 2015; Hajibayova & Buente, 2017; Hajibayova et al., 2016; Kam, 2007; D. Lee, 2011; Lilley, 2015; Littletree & Metoyer, 2015; Montenegro, 2019; Webster & Doyle, 2008). Green (2015) argues that the complicated relationship between mainstream perspectives and indigenous perspectives are difficult to untangle, resulting in KOS such as DDC to be misunderstood or truly biased. Despite efforts in expanding DDC as described in Green (2015), how mainstream KOS such as LCSH, DCC, or the Library of Congress Classifications (LCC) can be skewed in describing minority

groups have been manifested in Hajibayova et al. (2016) and Hajibayova and Buente (2017). The authors investigate the topic of Hawaiian Hula and found that these KOS depict Hula dance as a strand of folk dance, rather than what indigenous peoples consider to be a “sacred celebration”.

Furthermore, Webster and Doyle (2008) claim that LCSH treats the indigenous peoples’ knowledge as a diaspora, where the knowledge is ghettoized in one single section. For example, all the subject headings about indigenous peoples belongs in class E (History), implicitly denoting that the indigenous peoples are “people in the past”. There is also a prevalent use in subject headings on “Indian wars”, “captivities”, and “massacre” relating to the indigenous peoples, hinting perpetual stereotypes of a combative nature. Similarly, the continued use of the word “Indian” also portrays a colonialism, euro-centric perspective Kam (2007); D. Lee (2011); Webster and Doyle (2008).

Awaiting mainstream KOS to incorporate indigenous knowledge may be futile, given that consensus from working groups and reflection of new changes in updated versions usually happens gradually. To mitigate this issue, the Maori subject headings are developed to enhance the “intellectual access to Maori materials” in New Zealand (Lilley, 2015). The Xwi7xwa Library at the University of British Columbia in Canada has adopted the Brian Deer Classification scheme (BDC) to organize their collections (Kam, 2007). The Maschan-tucket Pequot Thesaurus has also been constructed to reflect Native American’s philosophies (Littletree & Metoyer, 2015). In a more recent paper, Montenegro (2019) introduces embedding “TK labels” in Dublin Core metadata to explicitly show indigenous people’s traditional knowledge. Improving the visibility of indigenous knowledge organization via the creation of alternative standards or KOS may be a current best practice. However, how to align the mainstream KOS with the lesser-known KOS is pending an answer.

We explore the hidden assumptions of knowledge organization systems (mainly taxonomies), in a more approachable, common-knowledge use case; second, reconciling these taxonomies using a logic-based approach to provide co-existing perspectives. Specifically,

Table 7.1: Taiwan’s Recognized tribes over time

| Year      | Number | Tribes                                                                  |
|-----------|--------|-------------------------------------------------------------------------|
| 1948-2000 | 9      | Amis, Atayal, Paiwan<br>Bunun, Rukai, Pinuyumayn<br>Tsou, Saisyat, Yami |
| 2001      | 10     | +Thao (org. Tsou)                                                       |
| 2002      | 11     | +Kavalan (org.Amis)                                                     |
| 2004      | 12     | +Truku (org.Atayal)                                                     |
| 2007      | 13     | +Sakizaya (org. Amis)                                                   |
| 2008      | 14     | +Seedqi (org.Atayal)                                                    |
| 2014-     | 16     | +Hla’alua, Kanakanavu<br>(both org. Tsou)                               |

“+” sign indicates that the tribe(s) is an addition to the previous row(s).

Source: Templeman (2018) and Tsai (2017)

this chapter draws on the use case of Taiwanese indigenous peoples’ tribe classification and the indigenous constituencies of the legislature electoral representation from the period of 2000 to 2020 (as reported in (Templeman, 2018) and (Tsai, 2017)). In the remainder of the chapter, we examine and reconcile the intertwined taxonomies of the Japanese colonial period classification of two indigenous constituencies (Lowland People and Highland People) (Figure 7.1), and the recognized indigenous groups from 9 to 16 groups (Table 7.1).

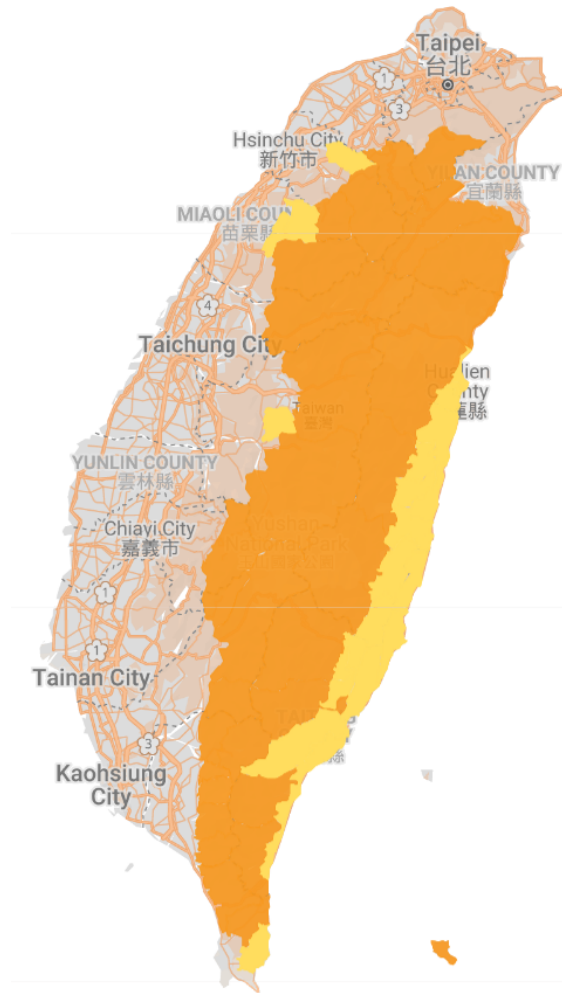


Figure 7.1: Taiwan's Indigenous electoral constituencies map

Lowland and Highland electoral constituencies in Taiwan (LH). Orange: Highland; yellow: Lowland.

Source of the approximate geographic distribution of LH is from the Council of Indigenous Peoples (CIP) population dataset.

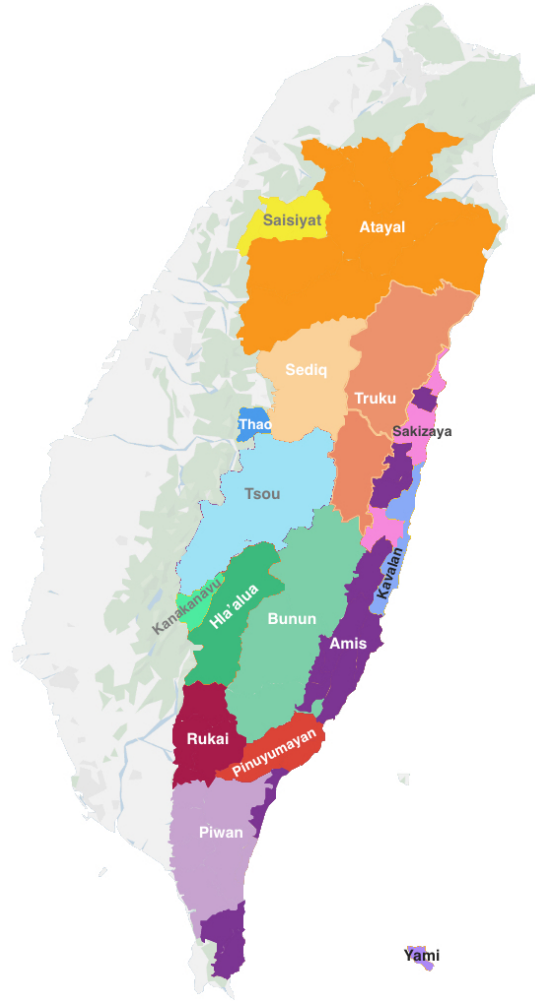


Figure 7.2: Taiwan’s current 16 recognized tribes and their distribution map

Source of data: CIP.

## 7.3 Taxonomy Alignment Problems

### 7.3.1 Background

#### **Taiwan Indigenous Classification and Electoral Constituencies**

Closely related to other ethnic groups in Asia-Pacific, indigenous Taiwan is viewed as an integral part of the Austronesian peoples and their languages (Blust et al., 2013; Chiu &

Chiang, 2012). Similar to many other native peoples all over the world, Taiwan's indigenous peoples have experienced hardships of colonialism, oppression, and cultural assimilation. To date, they are continuously marginalized, deprived of their home and resources, and has become a minority of the nation (Chiu & Chiang, 2012; Ericsson, 2004; Simon, 2010; Templeman, 2018; Tsai, 2017).

Taiwan indigenous peoples' identities are gradually recognized (Chiu & Chiang, 2012; Ericsson, 2004; Templeman, 2018). Over the years, indigenous peoples' name rectification movements include (1) banning the use of pejorative terms such as "cultivated barbarians", "mountain compatriots" to the use of neutral term such as "yuanzhumin" (aborigines); (2) allowing indigenous peoples to change their personal names to their mother tongue rather than retaining a Mandarin name; (3) recognizing more tribal names; and (4) discussing the restoration of Austronesian naming conventions for geographical features such as land and water bodies (Chiu & Chiang, 2012; Ericsson, 2004). While some of these movements are still undergoing in Taiwan, tribes have been federally recognized and added since the year 2000, evolving from only 9 groups to the now 16 tribes (Figure 2). The 16 tribes are: Amis, Atayal, Paiwan, Bunun, Puyuma, Rukai, Tsou, Saisiyat, Yami, Thao, Kavalan, Truku, Sakizaya, Seediq, Hla'alua and Kanakanavu (Templeman, 2018; Tsai, 2017).

Significant efforts were made by name rectification activists, tribal people, and allies to get the tribes officially recognized. For instance, when Thao peoples were arguing a split from the Tsou peoples, they have to leverage measures such as scientific DNA data to prove they have different ancestral roots from the Tsou people (Tsai, 2017). However, in Tsai (2017)'s ethnographic study, it is also revealed that although the Thao people eventually regained their tribal name, they grew up with identity crisis, not knowing where they truly belong to. To quote a Thao interviewee in Tsai's study, "Why do we speak different languages and cannot understand each other if we are both from Tsou?".

Further, to ensure equitable rights for the indigenous peoples in Taiwan, an arbitrary reserved seats system has been introduced to elect indigenous legislatures (Ericsson, 2004;

Templeman, 2018). However, the “reserved seats” system is often questionable given that the constituencies (voting groups) are divided in a dichotomy that was developed during the 1940s Japanese colonization period and still used in present day (Simon, 2010; Templeman, 2018). In the most recent 2020 presidential and legislatures election in Taiwan, out of the 113 legislatures, six seats were reserved for indigenous candidates (Biho, n.d.; Bowman, 2021; Templeman, 2018). In these six seats, three are for “Lowlanders”, the other three for the “Highlanders”. In other words, those who are geographically classified by their ancestral roots in the “mountains” are the electoral area of the “highland” legislatures, while their counterparts presumably in the “plain” areas can only vote for the “lowland” candidates. This is particularly problematic when in fact more than half of the indigenous peoples no longer live in their designated voting areas (Lu, 2019). Despite the outdated bipartite taxonomy of the constituencies, indigenous leaders believe that having representatives in the government is the only way to have their voices heard (Biho, n.d.). This shows there is a pressing need to update the indigenous constituencies.

Similar socio-political discourses on indigenous tribes’ recognition and politics involvement is prevalent on different continents in the world. For instance, in the United States, complicated laws and processes are enforced for the federal and state recognition of a tribe (Koenig & Stein, 2008; McCulloch & Wilkins, 1995), and who qualifies as a member of a tribe (Jarvis, 2017). In New Zealand, Maori voices are represented in the Parliament in similar reserved seats measures as in Taiwan (Humpage, 2008; Xanthaki & O’Sullivan, 2009). Further, Bird (2014) enumerates 19 nations worldwide and sorts them into three “families” of different mechanisms to provide ethnic quotas in their electoral systems. This gives evidence that indigenous peoples worldwide face tribe identity crises; in addition, how their electoral constituencies are formed is also an ubiquitous issue.

### 7.3.2 Data collection

**Tribal background information.** Data about the now 16 recognized tribes in Taiwan are collected from the Council of Indigenous Peoples in Taiwan’s website (CIP)<sup>2</sup>. This includes information on: the 16 tribes’ names, the approximate residencies of indigenous peoples, and the 55 indigenous counties in Taiwan. We also verify this information with two other well-known sources: Taiwan’s Indigenous Peoples Portal (TIPP)<sup>3</sup> and Taiwan Indigenous People’s knowledge Economic Development Association website<sup>4</sup>. Data about the evolving number of recognized tribes from 9 to 16 and the tribes’ names is gathered from Templeman (2018) and Tsai (2017) (See Table 7.1). By utilizing these data sources, we approximate the geographical distribution of the tribes as shown in Figure 7.2. These maps are simplified versions of where each tribe is. Some tribes are scattered within a county rather than populating all of the county, while some others include more than one tribes (but the map layering can only show one tribe). These maps also do not take into account the tribes’ traditional territories (perceived boundaries of location rather than the municipal division of cities and counties).

**Indigenous electoral constituencies information.** The classification of lowlanders and highlanders as well as every legislative election results can be found in Taiwan’s Central Election Commission database (CEC)<sup>5</sup>, as well as CIP’s website. Approximation of the geographic distribution of Lowland-Highland is shown in Figure 7.1.

**Indigenous population dataset.** The indigenous people population census dataset is also collected from CIP’s website, with updated census information every month. The date of data collection for this chapter is April 11th, 2020. The dataset contains specific sections on ‘lowlander’ and ‘highlander’, and how many people in each tribes counts as highlanders or lowlanders. As shown in Table 7.2, most tribes have members counted in both constituencies,

---

<sup>2</sup>CIP:<https://www.cip.gov.tw/>

<sup>3</sup>TIPP:<http://www.tipp.org.tw/>

<sup>4</sup>Taiwan Indigenous People’s knowledge Economic Development Association:<http://www.twedance.org/>

<sup>5</sup>CEC:<https://www.cec.gov.tw/english>

Table 7.2: Taiwan indigenous peoples population dataset

| <b>Tribes</b> | <b>Lowland</b>  | <b>Highland</b> | <b>Population</b> | <b>Percentage</b> |
|---------------|-----------------|-----------------|-------------------|-------------------|
| Amis          | 212611 (99.44%) | 1199 (0.56%)    | 213810            | 38.06             |
| Paiwan        | 21419 (20.81%)  | 81512 (79.19%)  | 102931            | 18.32             |
| Atayal        | 1871 (2.03%)    | 90364 (97.97%)  | 92235             | 16.42             |
| Bunun         | 364 (0.61%)     | 59248 (99.39%)  | 59612             | 10.61             |
| Truku         | 147 (0.45%)     | 32263 (99.55%)  | 32410             | 5.77              |
| Pinyumayan    | 14446 (99.23%)  | 112 (0.77%)     | 14558             | 2.59              |
| Rukai         | 2644 (19.60%)   | 10847 (80.40%)  | 13491             | 2.40              |
| Sediq         | 15 (0.14%)      | 10455 (99.86%)  | 10470             | 1.86              |
| Saisyat       | 4506 (66.89%)   | 2230 (33.11%)   | 6736              | 1.20              |
| Tsou          | 15 (0.22%)      | 6686 (99.78%)   | 6701              | 1.19              |
| Yami          | 12 (0.26%)      | 4681 (99.74%)   | 4693              | 0.84              |
| Kavalan       | 1499 (99.87%)   | 2 (0.13%)       | 1501              | 0.27              |
| Sakizaya      | 986 (99.80%)    | 2 (0.20%)       | 988               | 0.18              |
| Thao          | 811 (99.14%)    | 7 (0.86%)       | 818               | 0.15              |
| Hlaalua       | 0               | 420 (100%)      | 420               | 0.07              |
| Kanakanavu    | 0               | 366 (100%)      | 366               | 0.07              |
| Total         | 261346          | 300394          | 561740            | 100               |

Dataset collected from CIP, with our annotations on percentage of lowlander-highlander by each tribe's population. Our threshold for counting the citizens registered to be Lowland or Highland is over 20%. If it exceeds 20% in either Lowland or Highland, we consider that particular tribe overlaps in both regions.

Total: total population of each tribe; Percentage: percentage of each tribe's population over total indigenous population.

but some with very small population (e.g., only two people out of 1,499 from Kavalan are highlanders).

We adjust the classification of lowlanders and highlanders given some registered population of lowland or highland is minuscule. As shown in the percentages in the Lowland and Highland columns in Table 7.2, if the percentage of either Lowland or Highland do not exceed 20%, we adjust the tribe to only belonging in one category. For instance, Tsou’s percentage of Lowland is only 0.22%, then we consider Tsou’s population only in highland; Saisiyat’s percentage of Highland is 33.11%, then we consider Saisiyat’s population to be both in highland and lowland. We use this dataset and the 20% threshold as the basis to form the articulations in our TAPs. This study is a first attempt to compare the classifications of the indigenous voting groups with the actual number of tribes in Taiwan, we choose 20% as a threshold to roughly reflect the 20% minority migration pattern studies such as in (Hervitz, 1985). The percentage threshold can be modified in terms of different modeling situations and interpretations. We do not use the geospatial maps in Figure 7.1 or 7.2 to form articulations because the maps are only an approximation of where the tribes are.

### 7.3.3 Modeling and interpretation

Taxonomies sometimes may not be readily available in a tree-like structure. Thereby, conscious human decisions are involved to model the taxonomies. For instance, based on the map shown in Figure 2, we make the assumption that these 16 tribes are all geographically located in Taiwan, therefore considering these tribes as children nodes of Taiwan in a taxonomy. We then transpose the map into a tree, or taxonomy (as shown in Figure 5). Besides modeling taxonomies  $T_1$  or  $T_2$ , human interpretations are also needed when there are no evidence or ground truth for how to relate articulations A from  $T_1$  to  $T_2$ . For instance, experts’ opinions are given in domain-specific TAPs such as the species concepts alignments in Franz, Chen, et al. (2016); Franz, Pier, et al. (2016). Other example of modeling and

interpretations of taxonomy alignment problems can be found in our prior work (Cheng et al., 2017a; Cheng & Ludäscher, 2019b). Interpretations on the concepts and edges of the taxonomies is discussed in Chapter 4.5.

### 7.3.4 Four pairs of input taxonomies

In this chapter, we compare different taxonomies against  $T_1$  – the bipartite electoral constituencies of the indigenous groups, which divides the indigenous voting groups into ‘Lowland’ and ‘Highland’. We transcribed the maps into trees, with the syntax (Parent Child1 Child2... ChildN):

$T_1$ : The Bipartite Electoral Constituencies, Lowland and Highland (LH)

The input taxonomy  $T_1$  for highland and lowland is simply as the following, in which both of the groups are situated in Taiwan:

(Taiwan Highland Lowland)

$T_2$ : Recognized tribes in the past – 9 tribes

We investigate the input taxonomy  $T_2$  for the recognized tribes over time, mainly from 9 tribes to the now 16 tribes. Here we only report the input taxonomy for the 9 tribes.

(Taiwan Amis Atayal Paiwan Bunun Tsou Rukai Pinuyumayan Saisiyat Yami)

$T_3$ : Recognized tribes in the present – 16 tribes

We also compare the input taxonomy  $T_3$  of the recognized tribes now (as of year 2020) to  $T_1$ . The input taxonomy is:

(Taiwan Amis Atayal Paiwan Bunun Tsou Rukai  
Pinuyumayan Saisiyat Yami Thao Kavalan Truku  
Sakizaya Sediq Hlaalua Kanakanavu)

$T_4$ : Other tribes – 26 tribes

In the digital archive of the Institute of Ethnology Academia Sinica there are 10 other groups that are not officially recognized yet. We take this into account and add it to the 16 tribes in the present to form a taxonomy of 26 tribes:

(Taiwan Amis Atayal Paiwan Bunun Tsou Rukai Pinuyumayan  
Saisiyat Yami Thao Kavalan Truku Sakizaya Sediq Hlaalua  
Kanakanavu Ketagalan Kulon Taokas Papora Babuza Hoanya  
Pazih Makatao Siraya Taivoan)

$T_5$ : Re-scaled groups based on population percentage – 5 groups

Based on the indigenous peoples census population data from CIP, the population of some tribes do not exceed 10% of the total population ( $n=561,740$ ). For instance, the 5th biggest tribe Truku accounts for only 5.66% ( $n=32,410$ ) of the total indigenous population. For this taxonomy, we re-scaled the tribes based on their weighted population to provide yet another modeling situation for this study. We set the threshold of 10% of the total population – and groups the smaller tribes in a “combined group”. This re-scaled taxonomy has 5 groups in total:

(Taiwan Amis Atayal Paiwan Bunun Combined)

As shown in Table 7.3, Amis, Atayal, Paiwan, Bunun, and Combined have population, respectively, 38.06% ( $n=213,810$ ), 16.42% ( $n=92,235$ ), 18.32% ( $n=102,931$ ), 10.61% ( $n=59,612$ ), and 16.58% ( $n=93,152$ ).

All datasets and use case artefacts are available at Cheng and Ludäscher (2020a).

Table 7.3: Taiwan indigenous peoples re-scaled population

|                 | Total          | Percentage |
|-----------------|----------------|------------|
| Amis            | 213,810        | 38.06      |
| Paiwan          | 102,931        | 18.32      |
| <b>Combined</b> | 93,152         | 16.58      |
| Atayal          | 92,235         | 16.42      |
| Bunun           | 59,612         | 10.61      |
| <b>Total</b>    | <b>561,740</b> | <b>100</b> |

Re-scaled population calculated by our study that combines the tribes that do not exceed 10% of the total indigenous population into a “Combined” category. Note that the Combined group overlaps in both Lowland and Highland regions.

## 7.4 Results

### 7.4.1 TAP 1: LH versus 9 tribes

**Articulations.** Based on the population data and our exclusion threshold, only Paiwan and Saisiyat are overlapping in both lowland and highland, all other tribes are either one or the other. Therefore, our articulations A for this TAP is as the following:

[T1.Taiwan equals T2.Taiwan]  
 [T1.Lowland includes T2.Amis]  
 [T1.Lowland includes T2.Pinuyumayan]  
 [T1.Highland includes T2.Atayal]  
 [T1.Highland overlaps T2.Paiwan]  
 [T1.Lowland overlaps T2.Paiwan]  
 [T1.Highland includes T2.Bunun]  
 [T1.Highland includes T2.Tsou]  
 [T1.Lowland overlaps T2.Saisiyat]  
 [T1.Highland overlaps T2.Saisiyat]  
 [T1.Highland includes T2.Rukai]  
 [T1.Highland includes T2.Yami]

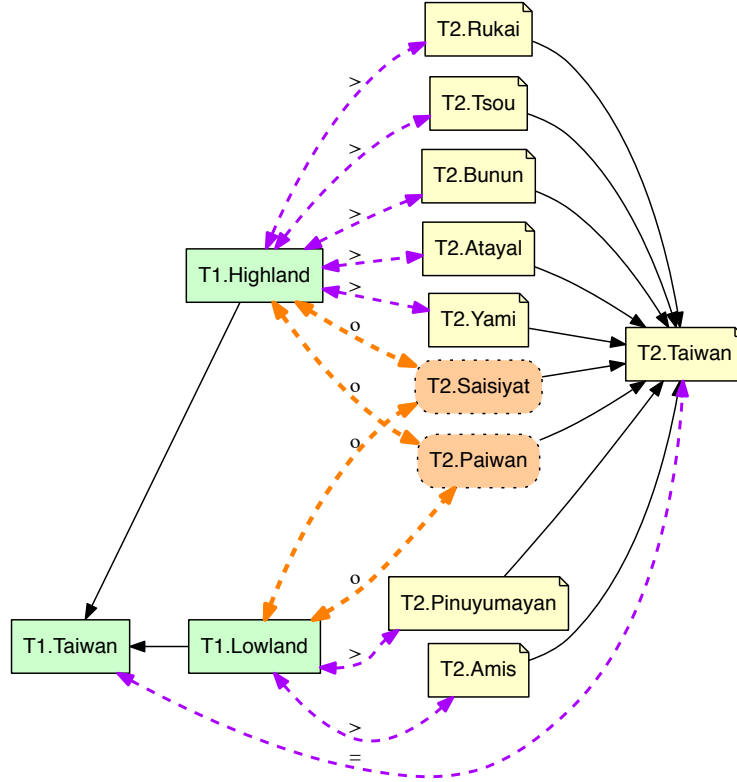


Figure 7.3: Input alignment for  $T_1$ .LH and  $T_2$ .9 tribes

**Input visualization.** As shown in Figure 7.3, the purple lines are the articulation  $A$  that connects  $T_1$  and  $T_2$ . We highlight the nodes and links of  $T_2$ .Siasiyat and  $T_2$ .Paiwan to orange to stress the overlapping nature of these two tribes.

**Interpretation of the PW.** Generated by the Euler tools, Figure 7.4 is the visualization of the output unique possible world for TAP 1. Grey round box nodes are congruent nodes from both taxonomies. All information from  $T_1$  (green) and  $T_2$ 's (yellow) are still presented in the PW  $T_3$ . Figure 7.4 shows that almost all of the nodes in  $T_2$  are now children of either  $T_1$ .Lowland or  $T_1$ .Highland, as expected. For  $T_2$ .Saisiyat and  $T_2$ .Paiwan, given that they overlap with both lowland and highland, the Euler tool inferred them to be direct child on the same level with  $T_1$ .Lowland and  $T_1$ .Highland. To resolve these overlaps, Figure 7.4 presents eight new pink round nodes that are the “combined concepts solution”, along with the inferred relations in red arrows. To read these newly generated nodes, for instance,

$T_1.$ Lowland \*  $T_2.$ Saisiyat is considered congruent with the region of  $T_2.$ Saisiyat  $\setminus T_1.$ Highland. Asterisk (\*) can be read as ‘intersection of’, while the backslash (\) means ‘excluding the region of’.

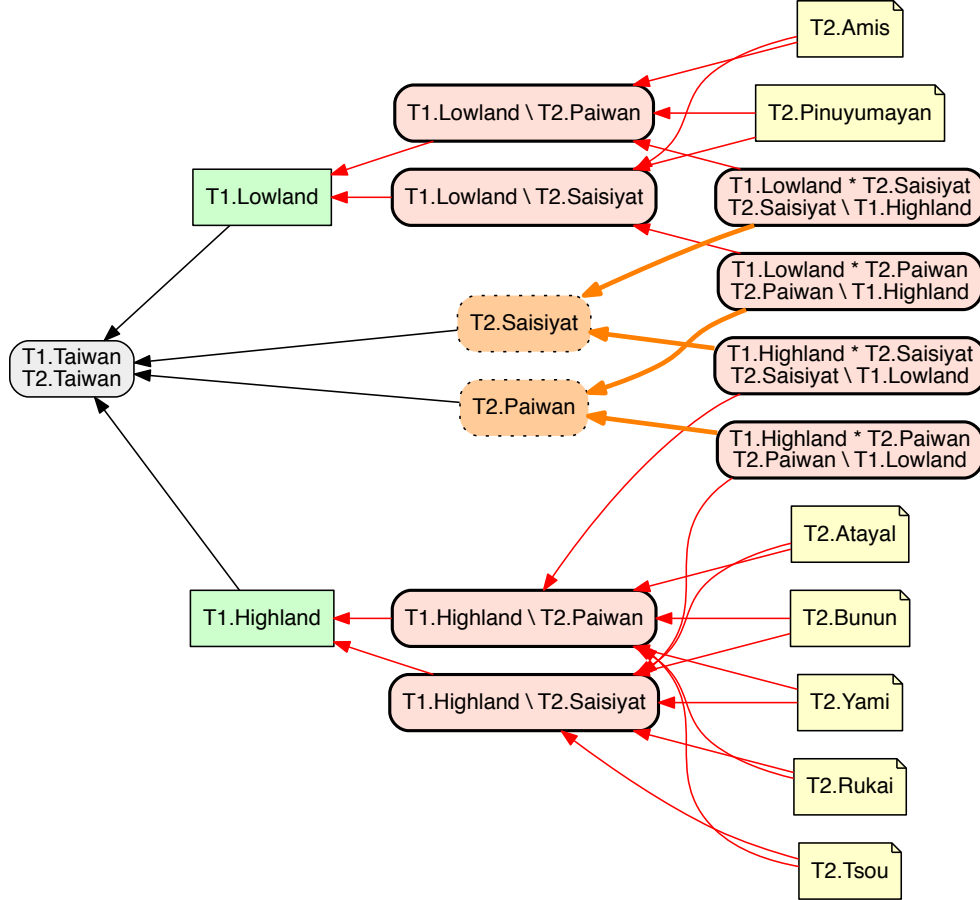


Figure 7.4: One PW for  $T_1.$ LH and  $T_2.$ 9 tribes

Examining closer at the PW in Figure 7.4, it is clear that all the nodes in  $T_2$  are children of the Lowland-Highland dichotomy. The new generated pink regions resolve the overlaps, in which both Siasiyat and Paiwan are further broken down with two children each: Saisiyat-Lowlanders, Saisiyat-Highlanders, Paiwan-Lowlanders, and Paiwan-Highlanders. This suggests that if we were to alter the indigenous electoral constituencies from the “archaic” bipartite taxonomy (Templeman, 2018) to a more granular taxonomy, we should have at least 11 voting groups (the lowest level nodes): the 7 recognized tribes plus the four new

pink nodes splitting from Saisiyat and Paiwan. The TAPs for other numbers of tribes are very similar to Figure 7.4, so here we only illustrate the PW for 9 tribes to symbolize the recognized tribes in the past.

#### 7.4.2 TAP 2: LH versus 16 tribes

**Articulations.** The articulations for the 16 recognized tribes are similar to that of TAP 1, but with seven more children to be sorted out in  $T_3$ . These articulations are a simplified version of which tribe counts in which electoral constituencies; as mentioned in the Data Collection section, we set a threshold of exceeding 20% population percentage to be considered in both groups. Therefore, in this TAP, the only two groups that overlaps in both lowland and highland are still Saisiyat and Paiwan.

[T1.Taiwan equals T3.Taiwan]  
[T1.Lowland includes T3.Amis]  
[T1.Lowland includes T3.Pinuyumayan]  
[T1.Lowland includes T3.Sakizaya]  
[T1.Lowland includes T3.Kavalan]  
[T1.Lowland includes T3.Thao]  
[T1.Highland includes T3.Truku]  
[T1.Highland includes T3.Atayal]  
[T1.Lowland overlaps T3.Paiwan]  
[T1.Highland overlaps T3.Paiwan]  
[T1.Highland includes T3.Bunun]  
[T1.Highland includes T3.Tsou]  
[T1.Highland includes T3.Hlaalua]  
[T1.Highland includes T3.Kanakanavu]  
[T1.Lowland overlaps T3.Saisiyat]

[T1.Highland overlaps T3.Saisiyat]

[T1.Highland includes T3.Rukai]

[T1.Highland includes T3.Sediq]

[T1.Highland includes T3.Yami]

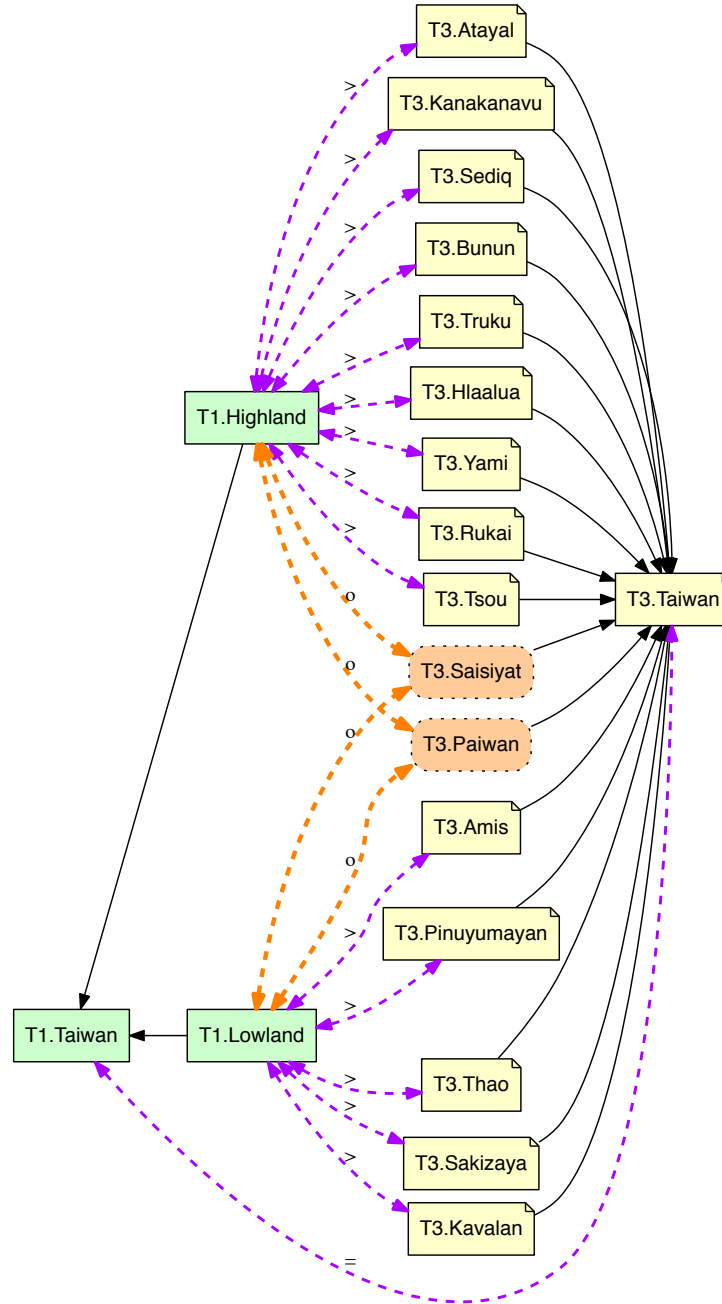


Figure 7.5: Input alignment for  $T_1$ .LH and  $T_3$ .16 tribes

**Input visualizations.** Figure 7.5 demonstrates the  $T_1$ ,  $T_3$  and A in this TAP. Again, the nodes where overlapping regions happens are marked with orange, while the articulations for the other nodes are marked to be either included in the highland region, or the lowland region.

**Interpretation of the PW.** Figure 7.6 also depicts similar features as Figure 7.4. To explain further the combined concept regions in the middle, (from now on, “middle-layer”), there are four new pink regions in the second-level nodes of the PW graph, mainly: (1) Lowlanders exclude Paiwan’s region; (2) Lowlanders exclude Saisiyat’s region; (3) Highlander exclude Paiwan’s region; and (4) (3) Highlander exclude Saisiyat’s region. Notably,  $T_3$ .Saisiyat and  $T_3$ .Paiwan are intertwined to be on the same level as  $T_1$ .Lowland and  $T_1$ .Highland, as well as the same level with the four middle-layer pink nodes.

Comparing Figure 7.6 to Figure 7.4 we can see that not much has progressed for the voting groups over time until now. More groups are lumped into the  $T_1$ .Highland than  $T_1$ .Lowland, creating an imbalance for the two voting groups. It may look like since there are a lot more tribes considered as Highland voters, they can potentially gain more voices. However, the simple dichotomy of Lowland versus Highland is actually diluting possible voices for all indigenous tribes. For instance, there may be two candidates from Amis, one from Pinuyumayan to run for legislative representatives in the Lowland area. However, those from Sakizaya, Kavalan, and Thao, considered as Lowland voters, can only then vote for candidates from tribes that are not theirs.

Based on Figure 7.6, ideally, we should have at least 18 voting groups rather than only two groups, as explicated in the lowest level children nodes of this figure, as well as the GIS map overlaying solution in Figure 7.7. In reality, it may be difficult to divide the indigenous reserved seats to 18 portions, given that the overall indigenous population accounts for only 0.02% of the total Taiwanese population (about 23 million), and the government officials may object to this idea. What the GIS solution in Figure 7.7 cannot show, is how to suggest a more fine-grained level than the Lowland-Highland division, but a more coarse-grained

level than the 18 different groups. To agree to disagree, the Lowland-Highland division (government’s perspective on the voting groups) and the at least 16 tribes (each tribe may want at least one candidate in the legislative) may be compromised in the “middle-layer” of the logic-based, taxonomy alignment approach. Rather than two or 16 groups, perhaps a more granular choice that both sides can agree on is the six groups in the middle layer of Figure 7.6.

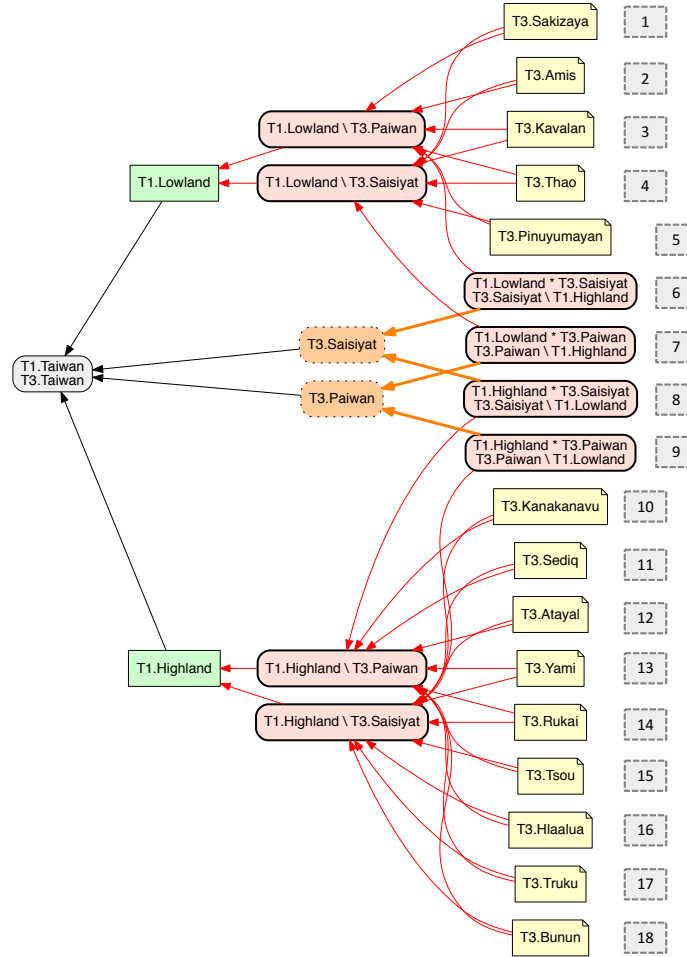


Figure 7.6: One PW for  $T_1$ .LH and  $T_3$ .16 tribes.

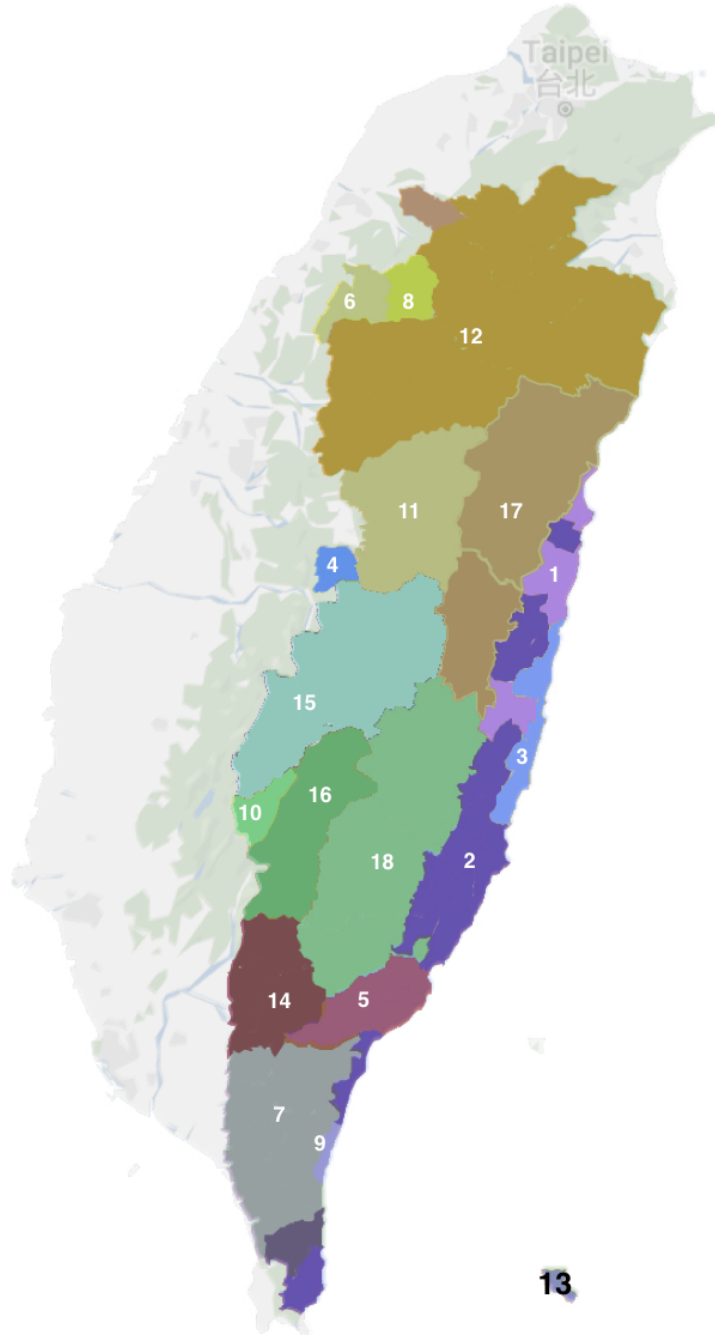


Figure 7.7: GIS map overlaying solution for  $T_1$ .LH and  $T_3$ .16 tribes

**Comparisons on using different thresholds.** We experiment here if using different thresholds in considering which tribes overlap in both group would yield any differences. In Table 7.2, 19.60% is in the Lowland, while 80.40% is in the Highland. Our original

alignment for  $T_1$ .LH and  $T_3$ .16 tribes uses the 20% threshold, based on literature about migration patterns. With this threshold, Rukai will not be considered to be overlapping in both groups. If we lowered the threshold to 19%, in the figure on the right on Figure 7.8, Rukai would then be considered to be overlapping in both regions. The resulting PW comparison is shown in Figure 7.9. We can see changes from the original 20% threshold, yielding six groups in the middle-layer, to the new 19% threshold, yielding nine groups in the middle-layer. This additional comparison is to demonstrate how different modeling situations (e.g. by simply choosing a difference threshold number) may influence the output PWs.

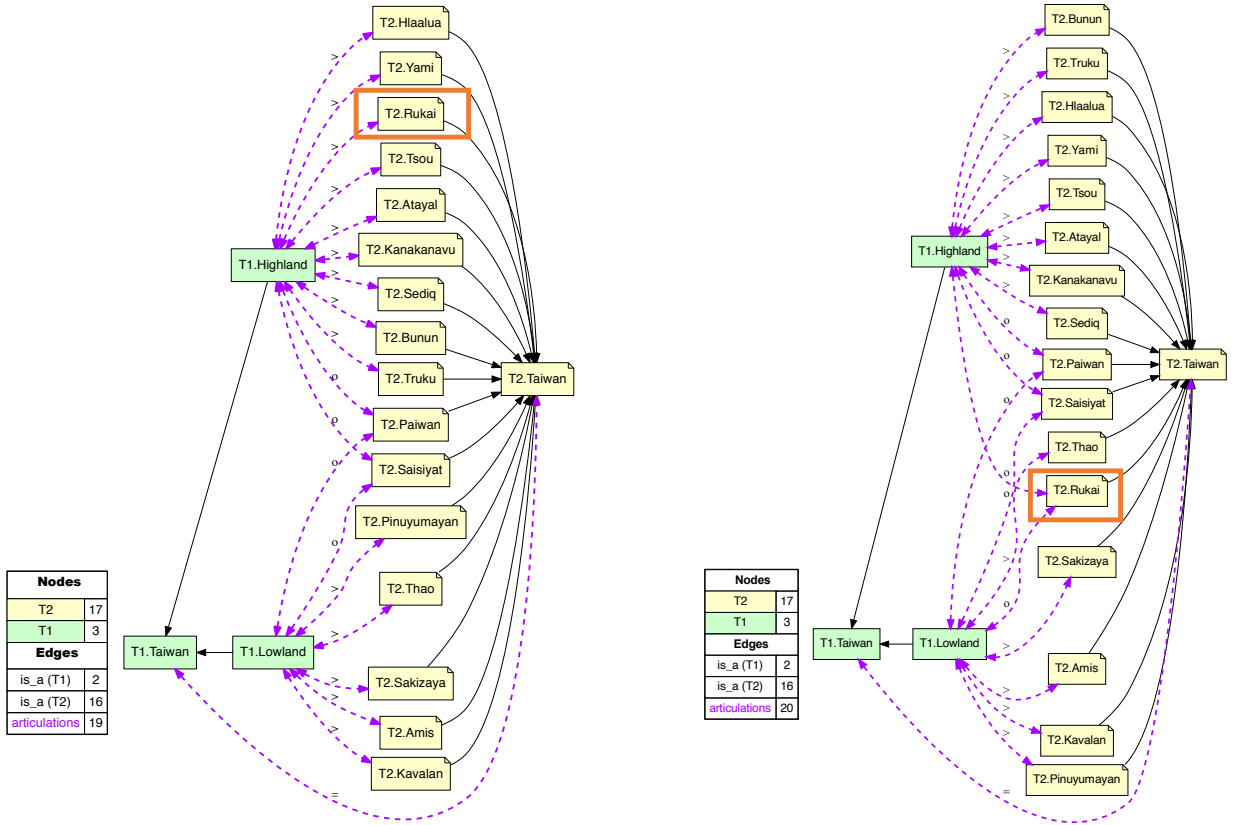


Figure 7.8: Input visualizations comparison based on two different thresholds

Left: 20% threshold input visualization; Right: 19% threshold input visualization

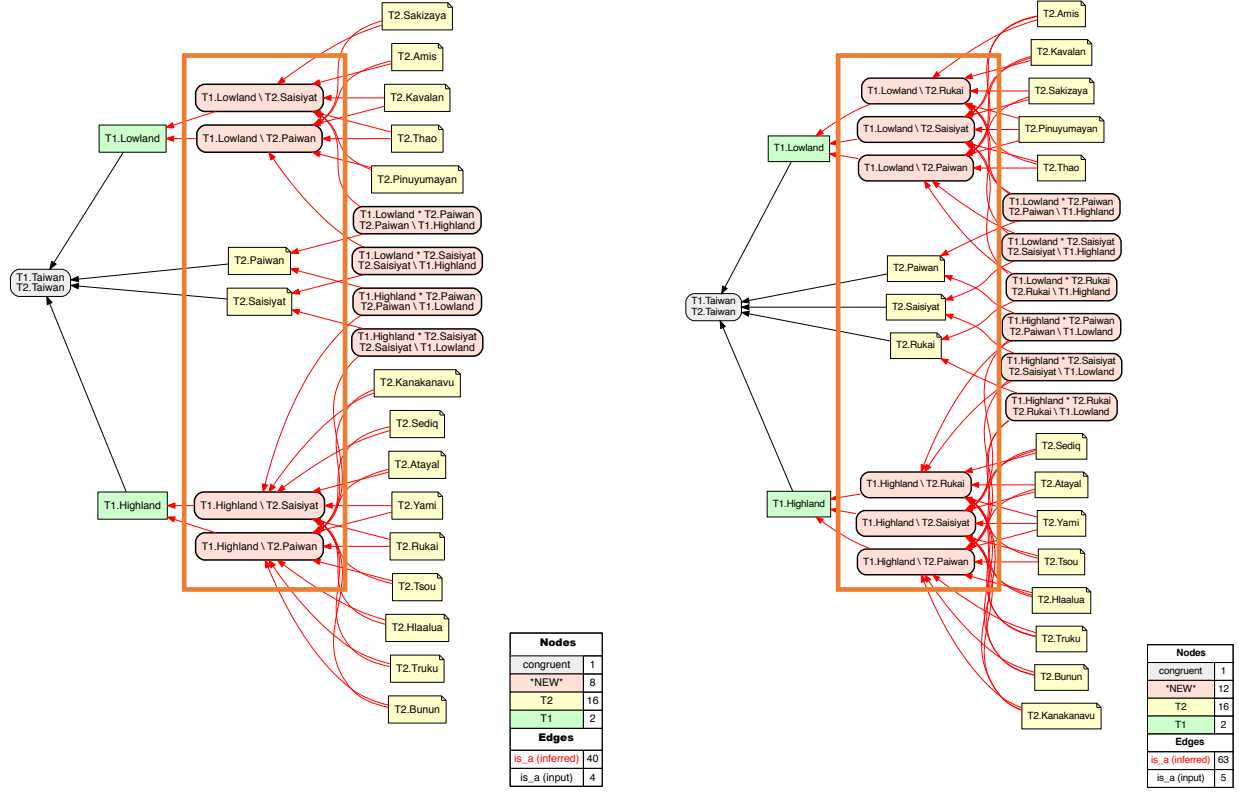


Figure 7.9: PW comparison based on two different thresholds  
Left: 20% threshold output PW; Right: 19% threshold output PW

### 7.4.3 TAP 3: LH versus 26 tribes

**Articulations.** To examine possible 10 other tribes that are not recognized and how they fit into the picture of the electoral constituencies, we make the following adjustments to the articulations. For  $T_1$ , we add a new child node “Other” to suggest there may be other electoral areas than the highland-lowland. Also, the data source has shown that these 10 other tribes do not live in the highland nor the lowland. They are scattered around the west coast and the north of Taiwan. For  $T_4$ , we added these other articulations to the existing articulations in TAP 2 with 16 groups:

[T1.Other includes T4.Ketagalan]

[T1.Other includes T4.Kulon]

[T1.Other includes T4.Taokas]  
[T1.Other includes T4.Papora]  
[T1.Other includes T4.Babuza]  
[T1.Other includes T4.Hoanya]  
[T1.Other includes T4.Pazih]  
[T1.Other includes T4.Makatao]  
[T1.Other includes T4.Siraya]  
[T1.Other includes T4.Taivoan]  
[T1.Other disjoint T4.Atayal]  
[T1.Other disjoint T4.Paiwan]  
[T1.Other disjoint T4.Rukai]  
[T1.Other disjoint T4.Saisiyat]

**Input visualizations.** Figure 7.10 depicts how it will look like if we include these other 10 tribes into the picture. The newly added nodes are marked in orange round nodes, making the input  $T_1$ ,  $T_4$ , and A an elongated figure.

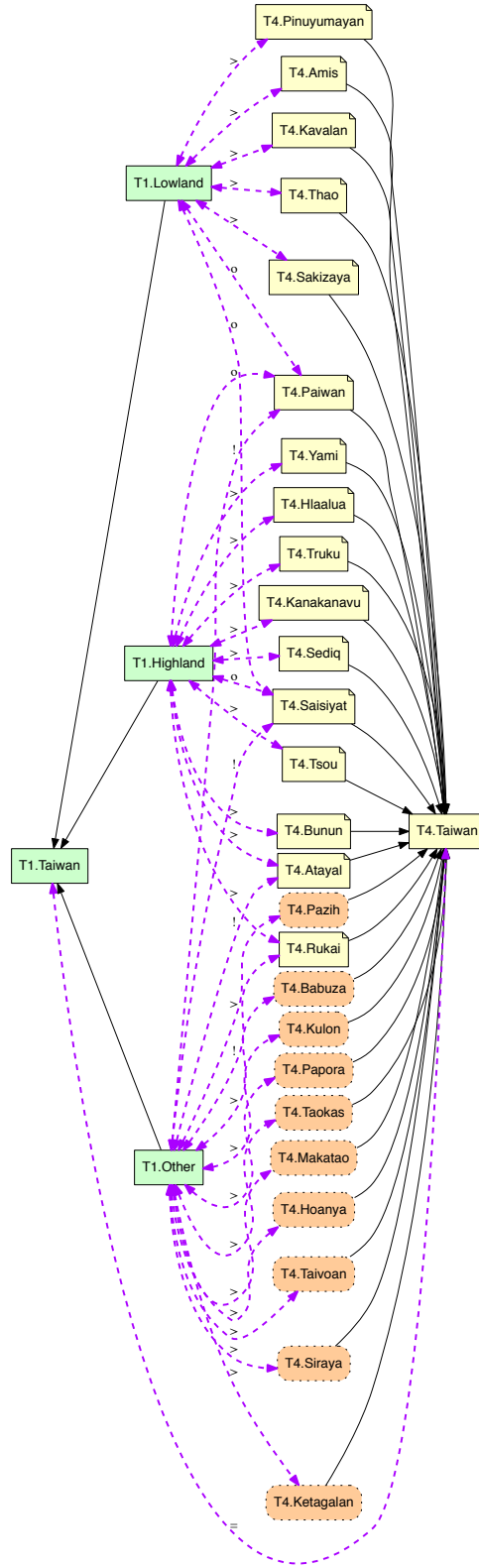


Figure 7.10: Input alignment for  $T_1$ .LH and  $T_4$ .26 tribes

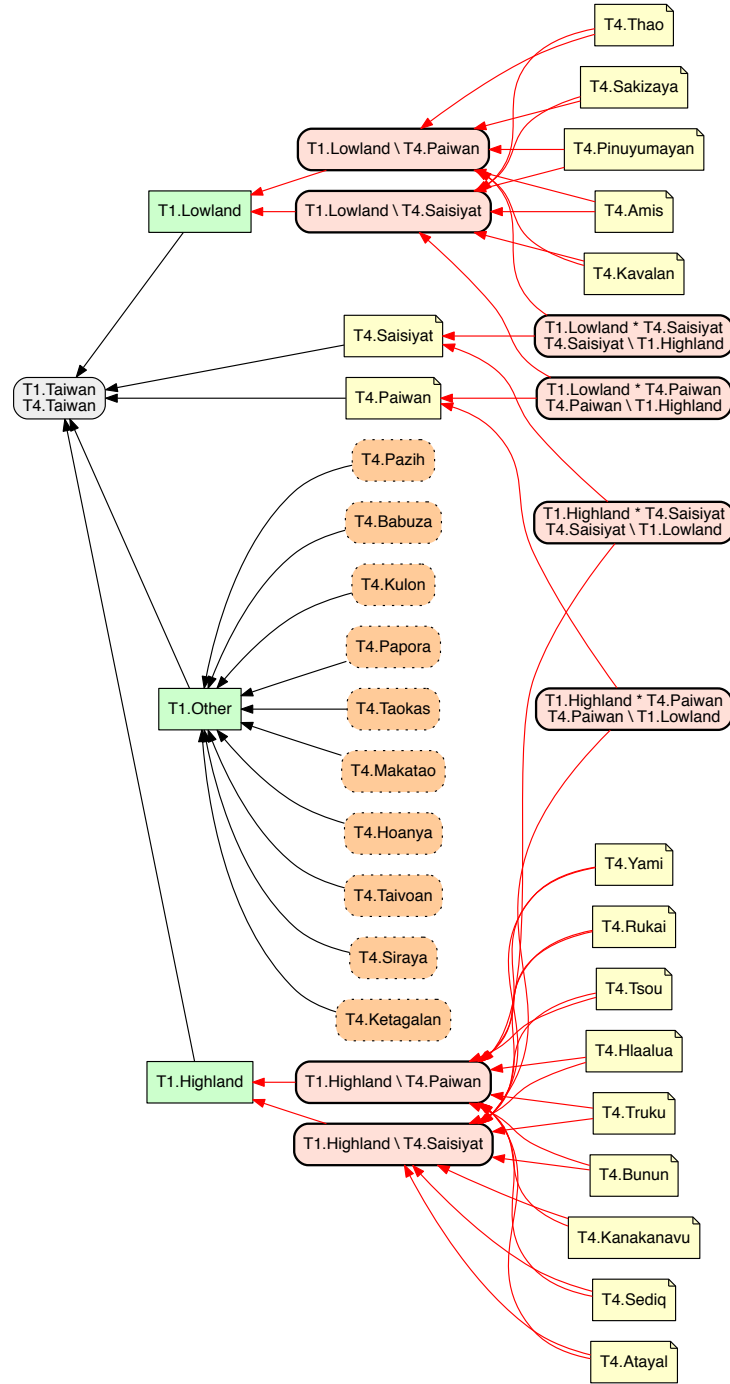


Figure 7.11: One PW for  $T_1$ .LH and  $T_4$ .26 tribes

**Interpretation of the PW.** The dynamic between the original 16 tribes and how they are merged with the Lowland-highland taxonomy is similar in Figure 7.11 as in Figure 7.6.

The only additions are the new 10 tribes (in orange round nodes). However, this graph suggests that it will be more difficult to migrate to a more granular taxonomy when other tribes might be recognized in the future. The “middle-layer” now not only has six nodes, but 16 altogether if the new tribes are added. This shows that the combined concepts pink nodes may work well in TAP 1 and TAP 2 as an agreement from two sides (Figure 7.4, 7.6); but it may not work so well when the number of recognized tribes becomes bigger and more complex (Figure 7.11). Further, even when the 6-groups division can function, there may be imbalance distribution of population among some groups (e.g., Saisiyat’s population is only 1.20% (n=6,636), but in our 6-groups middle-layer division they are count as one voting group).

#### 7.4.4 TAP 4: LH versus 5 re-scaled groups

**Articulations.** The four tribes that surpass the 10% threshold of the total indigenous population are Amis, Atayal, Bunun, and Paiwan. The rest of the 12 tribes are aggregated together as a ‘combined’ group to demonstrate the capability of re-scaled modeling of the logic-based taxonomy alignment approach. Given that some of the tribes in the Combined node are in the lowland, some in the highland, we consider the combined node to be overlapping with both constituencies. The revised articulations for the re-scaled modeling alignment are:

```
[T1.Taiwan equals T5.Taiwan]
[T1.Lowland includes T5.Amis]
[T1.Highland includes T5.Atayal]
[T1.Lowland overlaps T5.Paiwan]
[T1.Highland overlaps T5.Paiwan]
[T1.Highland includes T5.Bunun]
[T1.Lowland overlaps T5.Combined]
```

[T1.Highland overlaps T5.Combined]

**Input visualizations.** The layout and structure of Figure 7.12 is a much more simplified view than the other TAPs. Still, there are two regions that overlaps with both highland and lowland: Paiwan and the Combined node. Here we only mark the Combined in orange to stress that this is a new node based on the re-scaled population modeling.

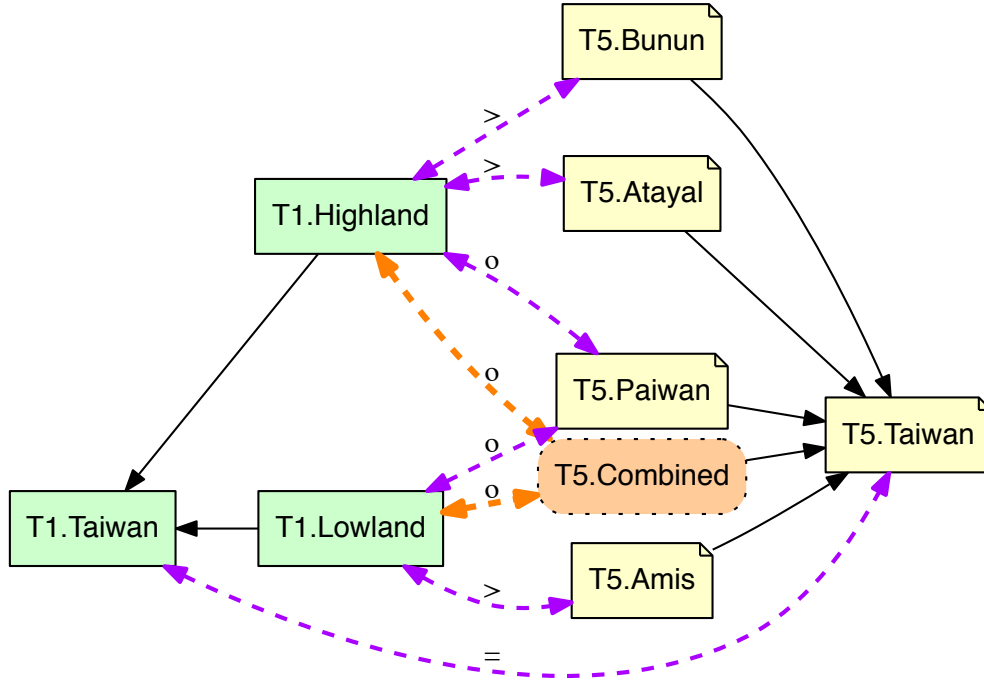


Figure 7.12: Input alignment for  $T_1$  and  $T_5$

**Interpretation of the PW.** Based on Figure 7.13, at the lowest level nodes, the number of tribes significantly dropped from 16 to only seven groups. Further, the “middle-layer” combined concept nodes now only have six groups: (1) Lowland excludes all those groups in the combined regions; (2) Lowland excludes Paiwan; (3) Paiwan; (4) Combined; (5) Highland excludes the combined regions; (6) Highland excludes Paiwan. We label the population of each group in the figure, starting from the lowest level nodes. Each of these nodes have two edges, so we evenly divide its population to distribute to the middle layer nodes. For instance, the Combined node has one edge from  $T_1$ .Lowland \*  $T_5$ .Combined (3.98%, n=22,371), and

another from  $T_1.$ Highland \*  $T_5.$ Combined (12.60%, n=70,781), which divide by two, the edges are 1.99% (n=11,185.50) and 6.30% (n=35,390.50) respectively. This adds up to 8.29% (n=46,576) of the total population for the Combined node.

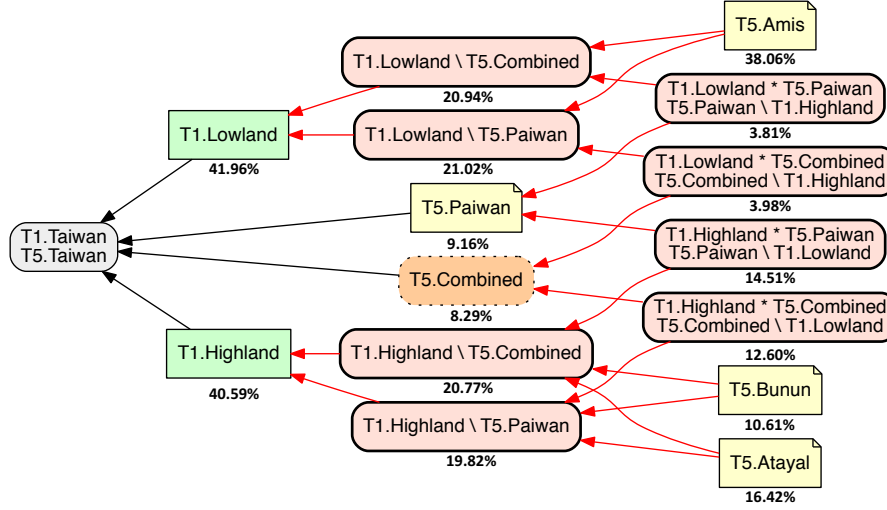


Figure 7.13: One PW for the re-scaled modeling TAP  $T_1$  and  $T_5$

Visualization shows the population percentage of each groups over the total indigenous population underneath each node (n=561,740).

The resulting middle-layer has population ranging from 8.29% (n = 46,576) to 21.02% (n = 118,091). Though the re-scaled population modeling alignment is still not evenly distributed in the middle layer nodes, it is higher than the former mentioned 1.19% of Saisiyat being its own voting group and can hopefully mitigate the imbalance in the electoral constituencies.

## 7.5 Chapter Summary

Through the use case of comparing the taxonomies of recognized tribes and the electoral constituencies of legislative representatives of indigenous Taiwan, we exemplify how embedded assumptions can be indicative not only in bibliographic classifications, but also in

geopolitical taxonomies. In our case, we raise the question of why the number of recognized indigenous tribes is increasing, but the indigenous electoral constituencies (voting group) remain static in a dichotomy from the 1940s until now.

The four taxonomy alignment problems (TAP) we present in this chapter show that multiple viewpoints can coexist in the resulting merged taxonomy, or possible world(s) (PW). Our first TAP showcases how the recognized indigenous tribes in the past (with 9 groups), when aligned with the bipartite Lowland-Highland voting groups using the logic-based approach, preserves both taxonomies in the PW (in green and yellow nodes). The resulting PW (Figure 7.4) also depicts an ideal situation in the lowest-level nodes to classify a more granular electoral constituency with at least 11 groups rather than only two group.

With the progression of time, not much has changed for the voting constituencies as of now. This is shown in TAP 2, where we align the now 16 recognized indigenous peoples' tribes with the Lowland-Highland division. In the resulting PW (Figure 7.6), there is an imbalance of more tribes being included in the Highland than the Lowland, but the tribes being lumped in either Lowland or Highland are diluting the true voices of each single tribe. The elected candidate may not reflect the true tribal representative since they may be from only one tribe, and based on the PW, perhaps at least 18 groups needed to be represented in the electoral constituencies.

However, given that the indigenous peoples population are only about 0.02% of the total population in Taiwan (23 million people), claiming 18 reserved electoral seats (out of 118 seats) to indigenous peoples is highly unlikely. This is when our logic-based taxonomy alignment approach demonstrates potential in solving the issue by inferring a middle-layer "combined concepts" to reach a consensus on how to divide up the voting groups. Instead of 18 groups, we can perhaps utilize the six groups in the middle layer of Figure 7.6. This middle layer inference has also shown a novel approach of reconciliation of taxonomies that GIS map layering result cannot show (Figure 7.7).

Agreeing to disagree with the combined concept alignment may help both parties to

reach a consensus in the middle ground. However, with more possible groups added into the picture in TAP 3, the middle layer concepts may also grow exponentially (Figure 7.11). Further, if population is taken into account, for tribes that are inferred in the middle layer, some of them account for only 1.20% of the overall indigenous peoples' population. To mitigate the skewed population distribution in the mid-level concepts, in TAP 4: we discuss the possibility of a re-scaled taxonomy alignment modeling. We group those tribes that have smaller population into a Combined group and re-align the weighted groups with the Lowland-Highland division.

The resulting PW (Figure 7.13) demonstrates a more proportionate classification of tribes. This TAP may have shown that the logic-based alignment approach can be flexible to re-scale and model the taxonomies to the granularity of our liking, but it is not to say that the PW of this TAP is the most accurate out of the four TAPs we showed in this chapter. We are cautious that all TAPs demonstrated in this chapter serve as options rather than optimal solutions to compare these taxonomies. More factors still need to be considered (e.g., relation of the tribes to one another, parties, etc) to form more comprehensive modeling situations in aligning the taxonomies.

This chapter is our first attempt on a re-scaled taxonomy alignment problem. The limitation of this study is that the threshold for the cut-offs or combination we set is uniquely catered towards our population dataset, so it may not be applicable to other datasets. Further, we use the indigenous peoples population dataset rather than the actual indigenous voter population (age 20+), in which the latter would be minuscule in scale to model the taxonomies.

The contributions of this study is to present different perspectives of a taxonomy alignment problem. This approach also may have implications for schema migration or KOSs versioning updates – given that the Lowland-Highland taxonomy is not only a taxonomy for voting groups, but it is also how the government recognized the indigenous groups in the 1940s. The practical implication, on the other hand, is manifested in our use cases with how

the results may be taken into account for policy-making in the future. Nevertheless, the discussion on policy-making, regaining voting rights, or how our results have implication for partisan, racial, or ethnic gerrymandering is beyond the scope of this study.

Cognizant of the fact that both taxonomies are technically through the lens of the government (the government decided on the dichotomous constituencies; the “government” recognized the tribes), this chapter endeavors to incorporate the many possible other tribes (26 groups) to be inclusive of possible different perspectives. Through these alignments, we exhibit the feasibility of modeling and interpretation of the logic-based taxonomy alignment approach. In the near future, we plan to consider other dimensions relating to the electoral constituencies, such as incorporating factors of political parties; and examine other facets on the perceived boundaries of indigenous peoples’ land.

## CHAPTER 8

# RECONCILING THE SOCIO-GEOGRAPHIC IN BIODIVERSITY DATA AND INFORMATION SYSTEMS

Biodiversity data has been used extensively to approximate species population, track species status, or make conservation plans. However, complex information embedded in biodiversity data and information systems often makes it difficult to gain a holistic picture about the species, especially when geographic and historical information are involved. For instance, geographic information may be constantly changing, while in biodiversity data only the most recent geographic entities are documented. In legacy biodiversity literature, place-names that were used in the past may no longer be the same name used now. These challenges create a perpetual gap from linking historical texts that could be useful to understand species' habitats and the causes of population reduction, to the actual species occurrence datasets that are used to inform decision-making.

In this chapter, we address these socio-geographic concerns in biodiversity data by proposing a conceptual model for modeling historical sovereignties over time. The conceptual model involves the construction of geo-taxonomies across time periods and alignment of these taxonomies. Given the ambiguity lies in modeling historical sovereignties, we anticipate that the taxonomy alignment process will yield multiple merged solutions. This chapter presents a systematic comparison of these multiple solutions by categorizing them into 'possible world types'.

## 8.1 Introduction

At around 5:45 P.M., October 10th, 2021, at the Cigu Salt Pan Wetland (23.152643, 120.076005) in Tainan City, an ornithologist witnessed a flock of birds preening their snowy feathers, and each other’s feathers, while occasionally stuffing their long, spatula-like bills into the muddy wetland to probe for a mudskipper feast. Excited as she was, the ornithologist, a wading-bird expert her whole life, knew almost immediately that these are a rare kind of Spoonbill birds that can only be spotted in East Asia during certain seasons. She quickly took detailed notes on the physical description, habitat, locality, and the number of birds observed. She then submitted here observations to iNaturalist <sup>1</sup> to share her discoveries with her fellow citizen scientists.

| TABLE | GALLERY                                         | MAP             | TAXONOMY      | METRICS       | ⬇️ DOWNLOAD        |
|-------|-------------------------------------------------|-----------------|---------------|---------------|--------------------|
| ⋮     | Scientific name                                 | Country or area | Coordinates   | Month & year  | Basis of record    |
|       | <i>Platalea minor</i> Temminck & Schlegel, 1849 | Chinese Taipei  | 24.8N, 120.9E | 1985 December | Human observation  |
|       | <i>Platalea minor</i> Temminck & Schlegel, 1849 | Chinese Taipei  | 24.2N, 120.5E | 1985 December | Human observation  |
|       | <i>Platalea minor</i> Temminck & Schlegel, 1849 | Chinese Taipei  | 24.2N, 120.5E | 1985 December | Human observation  |
|       | <i>Platalea minor</i> Temminck & Schlegel, 1849 | Chinese Taipei  |               | 1923 January  | Preserved specimen |

Figure 8.1: GBIF’s interface of the occurrence records of *Platalea minor*

Note that the “country or area” of these occurrence records are in the “Chinese Taipei”.

Obtaining geographic information is crucial in understanding biodiversity data for approximating species population, or tracking status changes, but geographic information in biodiversity data is often multi-layered and complex (Boakes et al., 2010; Bowker, 2000b; Sung, Tse, & Yu, 2018; Yeung, Yao, Hsu, Wang, & Li, 2006). For example, monitoring and gathering information about migratory endangered birds, such as the Black-faced Spoonbills, can be difficult because the information can span across multiple places, nations, or geographic entities (Sung et al., 2018).

<sup>1</sup>iNaturalist:<https://www.inaturalist.org/>

One complexity is when biodiversity aggregated databases adopt institutional-recognized geographic standards, or country taxonomies. In Figure 8.1, the occurrence records on the GBIF interface from 1923 of the Black-faced Spoonbills, or *Platlea minor* Tenminck & Schlegel 1849, are from the “Chinese Taipei”, a country list used only in the International Olympic Committee <sup>2</sup>, whereas other geographic information sources may call it “Taiwan” or the “Republic of China”.

| I.—PLATALEA MINOR (Korea and Japan). |              |                |              |       |                |                 |                        |         |                        |               |                         |
|--------------------------------------|--------------|----------------|--------------|-------|----------------|-----------------|------------------------|---------|------------------------|---------------|-------------------------|
| Collector and No.                    | Sex and age. | Locality.      | Date.        | Wing. | Tail feathers. | Exposed culmen. | Greatest width at tip. | Tarsus. | Middle toe, with claw. | Total length. | Naked portion of tibia. |
| Jouy, 1470                           | ♂ jun.       | Fusan, Korea.  | Dec. 7, 1884 | 380   | 109            | 168             | 52                     | 133     | 87                     | 775           | 66                      |
| Petersen, 81                         | ♂ jun.       | Nagasaki, Jap. | Dec., 1866   | 350   | 104            | 188             | 51                     | 129     | 80                     | .....         | 80                      |

| II.—PLATALEA MINOR (Formosa; <i>fide</i> Swinhoe). |              |                      |          |       |       |       |         |               |        |
|----------------------------------------------------|--------------|----------------------|----------|-------|-------|-------|---------|---------------|--------|
| Collector and No.                                  | Sex and age. | Locality.            | Date.    | Wing. | Tail. | Bill. | Tarsus. | Total length. | Tibia. |
| Swinhoe, 4                                         | ♂ ad.        | Tamsuy, Formosa..... | March 17 | 361   | 127   | 198   | 127     | 800           | 89     |
| Swinhoe, 2                                         | ♂ jun.       | .....do.....         | March 7  | 330   | 114   | 183   | 119     | 787           | 61     |
| Swinhoe, 3                                         | ♀ jun.       | .....do.....         | March 17 | 356   | 114   | 190   | 127     | 800           | 76     |

Figure 8.2: A table in historical text comparing different occurrences of *Platalea Minor*

Source: Proceedings of the United States National Museum (1887)

The level of complexity to examine geographic information rises even higher when there is a discrepancy between *contemporary* place-names in biodiversity data and *historical* place-names in biodiversity literature (Cheng, Dilliplane, & Ludäscher, 2020). Biodiversity occurrence datasets (Darwin Core metadata formatted) are usually based on contemporary country lists, or a certain perspective of the world’s geographic entities (Cheng, Dilliplane, & Ludäscher, 2020), but these place-names may have changed drastically over time (Bowker, 2000b; Buckland & Lancaster, 2004). For instance, the aforementioned Taiwan is referred

<sup>2</sup>International Olympic Committee’s nation lists: <https://olympics.com/ioc/national-olympic-committees>

to by its contemporary place-names in species occurrence datasets (Figure 8.1). As a contrast, Figure 8.2 is taken from the Smithsonian’s *Proceedings of the United States National Museum* in 1887, but only Taiwan’s historical name, *Formosa*, was mentioned in this piece of historical literature. In fact, in 1923 when the Black-faced Spoonbills are spotted in Taiwan in Figure 8.1, Taiwan was under the reign of the *Empire of Japan* and was referred to as *Formosa* at that point in time. If historical place-names are not considered, detailed text descriptions and table comparison such as the one in Figure 8.2 to compare Temminck & Schlegel’s *Platalea minor* spotted in Japan and Korea versus Swinhoe’s *Platalea minor* spotted in *Formosa* may have been lost.

|                                                                                                                                                 |                                  |
|-------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------|
| 268                                                                                                                                             | CHECK-LIST OF BIRDS OF THE WORLD |
| <b>PLATALEA MINOR</b>                                                                                                                           |                                  |
| <b>Platalea minor</b> Temminck and Schlegel                                                                                                     |                                  |
| <i>Platalea minor</i> Temminck and Schlegel, 1849, in Siebold, Fauna Japonica, Aves, p. 120, pl. 76—Japan.                                      |                                  |
| Korea, central Manchuria, probably eastern China south to Fukien. Winters Kwangtung, Hainan, Indochina, formerly Japan. Accidental Philippines. |                                  |

Figure 8.3: Possible sightings and occurrences of *Platalea Minor* in the world

Source: Check-list of Birds of the World (1979)

Furthermore, simulating historical geographic information (e.g., events and habitats) could be potentially helpful to understand the cause of species population reduction, like how the Korean War is believed to be one of the major causes of the decline in *Platalea minor*’s population (Sung et al., 2018; Yeung et al., 2006), but historical geographic information is difficult to trace. Information about historical events and habitats are reliant heavily in historical texts, which based these events in historical place-names. For instance, Figure 8.3 is taken from the *Check-lists of birds of the world* in 1979. It is mentioned that *Platalea Minor* occurs in “Indochina” in winter – a sovereignty that ceased to exist anymore. If accessing information relating to historical contexts are based entirely on contemporary names and sovereignties, information such as “Indochina” will not be found. Similarly, without an

understanding of where “Indochina” was and which places it covers, it will be difficult to trace causes of biodiversity loss events that corresponds to the current geographic entities (or nations).

The proliferation of these complex issues results from the perpetual disconnection of biodiversity species data to its historical records or literature (Franz & Sterner, 2018; Hobern et al., 2019). Darwin Core’s most recent update (version 2021-07-15) aims at bridging different resources by using the controlled vocabulary *Material Citation*<sup>3</sup> to document the provenance of an occurrence that is from publications or field notes. This mitigation would have been a great addition to tracking provenance of a species record in relation to its historical texts, but most of the occurrences have not been updated to reflect this label in their *Basis of Record* field. What exacerbates the linkage between biodiversity data and different databases is when the chain of custody on historical specimen is unknown. Here, we view the data not just as the digital species occurrence data, but also the physical specimen housed in Natural History Museums. For instance, the holotype of Black-faced Spoonbills (ID Number RMNH.AVES.87107), first discovered in Japan, is now owned by the Naturalis Biodiversity Center in the Netherlands; but there is no provenance to support how the holotype moved from Japan to the Netherlands. We can speculate that the people who first named the species (Temminck and Schlegel, both from the Netherlands) may have brought it home with them. The cause of this opaque provenance information could also be due to improper historical geographic name searches. On top of this, the species record for ID RMNH.AVES.87107 is also missing in occurrence datasets in aggregated databases such as the iNaturalist or in the Global Biodiversity Information Facility (GBIF). Overall, the linkage between biodiversity data, databases, and natural history literature still needs improvement to trace the provenance of biodiversity artifacts.

Valuable descriptions on species in historical texts and data would be lost if geographic name changes and historical sovereignties are not considered. The goal of this chapter thus

---

<sup>3</sup>DwC:MaterialCitation:[https://dwc.tdwg.org/list/#dwc\\_MaterialCitation](https://dwc.tdwg.org/list/#dwc_MaterialCitation)

is to unravel the intertwined relations of geopolitical and time (or historical) information in biodiversity data through conceptual modeling. We focus specifically in biodiversity species data that allegedly occur in the East Asia region. Biodiversity data is usually annotated with geographic location and information of contemporary nations (as of 2022), but each data entry could have been from a time that the contemporary nations had different place-names, or had not existed (Cheng, Dilliplane, & Ludäscher, 2020). We choose East Asia as our geographic point of interest because it is supposedly a lower “biodiversity data diversity” region compare to North America, with supposedly more richness in biodiversity but much fewer systematic information about the species (Bowker, 2000a, 2000b). To this extent, we aim to propose a conceptual model to examine how historical sovereignties of places change over time in the East Asia region, to uncover the hidden layer of geographic information that were previously unnoticed in biodiversity data.

The contributions of this chapter is three-fold: (1) we hope this work will acknowledge the implicit infrastructures (Bowker, 2000b; Lampland & Star, 2009) that are embedded in geographic information, taxonomies, or standards, and make these assumptions explicit; (2) through the use of taxonomic views to represent data (as opposed to timelines, or geographic maps), we hope to incorporate multiple perspectives and be inclusive to different dimensions of the data in critical cartography (Bishop, Moulaison, & Burwell, 2015); (3) we hope this work can shed lights on the intricacies of constant-evolving geographic information one need to be aware of when attempting to integrate biodiversity data, or make any decisions based on these data.

## 8.2 Related Work

### 8.2.1 Organizing Geographic Information

Geographic information is central to how we conceptualize the world, or even just what our surroundings are. Whether it be a thematic map of the electoral college in the 2020 U.S. presidential election, or a forecasting map showing the trajectory of a next winter storm, geographic information helps us make sense of the reality (Purves, Winter, & Kuhn, 2019). In fact, geographic information is said to represent an “abstraction of reality” that approximate a “multidimensional” complex reality in a flat, digital world (Bishop & Grubestic, 2016a). This near reality is not only represented in the form of maps, texts, structured data, knowledge organization systems (e.g., gazeteer, metadata) also often contain geographical components such as place-names, latitude-longitude, nations, etc (Bishop & Grubestic, 2016b; Buckland & Lancaster, 2004).

Efforts have been made in organizing geographic information. Generic Geographic knowledge organization systems (KOS) such as Geonames<sup>4</sup> and Getty Thesaurus of Geographic Names (TGN)<sup>5</sup> have been useful to disambiguate place-names and geographic entities. Standards and metadata schemes such as ISO 3166 Country Codes<sup>6</sup>, CIA World FactBook<sup>7</sup>, or domain-specific geographic distribution such as the World Geographical Scheme for Recording Plant Distributions (WGSRPD)<sup>8</sup> are used as controlled vocabularies for filling out place-name values.

Geographic information is not simply a reflection of the reality, it could be a skewed reality that depicts a certain image of a nation (Callahan, 2009). Harley (1988, 1992) drew from the Foucauldian ideas of the unintentional and the intentional and suggests that maps could be deliberately used as tools for “concealment, censorship and falsification” that leads

---

<sup>4</sup>Geonames: <http://www.geonames.org/>

<sup>5</sup>TGN: <https://www.getty.edu/research/tools/vocabularies/tgn/>

<sup>6</sup>ISO: <https://www.iso.org/iso-3166-country-codes.html>

<sup>7</sup>CIA World FactBook: <https://www.cia.gov/the-world-factbook/>

<sup>8</sup>WGSRPD: <https://www.tdwg.org/standards/wgsrpd/>

to the silence of certain discourses and the promotion of an “uneven record of the dialogue” between the strong and the weak. Crampton (2001) expanded Harley’s work by discussing how maps and cartography are used as communication devices to exhibit power relations, and how geographers should be aware of the power dynamics in geographic information.

The relative reality in geographic information makes it susceptible to inherent biases and assumptions. Bishop et al. (2015) introduced critical cartography cataloging and summarize three types of biases that could make geographic knowledge organization go awry: people naming places, cartographers mapping locations on maps, and information professionals organizing geographic information. All three processes, involving human-decision-making, introduce individual (or group) interpretations, hence, they could be prone to biases. The authors explained that one way to balance these biases could be place-name standardization in codes. However, codes such as country codes could also display biases towards certain perspectives (Cheng & Ludäscher, 2019b).

Extant empirical work endeavors to detect biases in geographic information. Quinn and Tucker (2017) examined the geopolitical entities in crowd-sourced online maps and confirm that even seemingly neutral maps like OpenStreetMap cannot be entirely objective in their projection of the world. Soeller et al. (2016) developed a tool (MapWatch) to detect customized map boundaries on Google map when it shifts borders based on users’ geolocation, especially at regions that are contested or have border conflicts. Our prior work on geopolitical country taxonomies and the geographic coverage of the Taiwanese indigenous peoples’ tribes aimed at raising awareness of the embedded assumptions in knowledge infrastructures, with reconciliation of biased taxonomies as an action to support multiple perspectives (Cheng & Ludäscher, 2019b, 2020b). These studies echoed Bishop et al. (2015)’s stance on how standardization and customization of geographic information could strip away the diversity and “discourages multiplicity”.

What complicates the problem with geographic information further is the presence of time. Bowker (2000b), in discussing specifically in the field of biodiversity, stressed that

the challenges with organizing geographic information are brought towards an even greater degree when time information is involved. In the next section we will talk about how time is generally organized, and how both time and space are organized in combination.

### 8.2.2 Organizing Time and Changes

Time, fluid and continuous as it is, has always be used to measure changes. Previous research has examined how the unit of analysis of time is paramount to capture changes in bibliographic classifications (Adler et al., 2013; Tennis, 2007, 2010). Tennis (2010) categorized the units to long-time, short-time, and micro-time to address revisions in classification schemes, the processes of cataloging and indexing, and the decision-making moment of indexing, respectively. Adler et al. (2013) further described in a panel discussion how knowledge organization is a non-linear, dynamic process, resonating Tennis (2010)’s viewpoint that the momentarily act of cataloging could “reverberate through time”. In other words, we could usually only capture a certain worldview at a particular point in time, but that worldview may ripple through time and introduce different interpretations in a future point in time.

To track changes over time in bibliographic classifications, in a follow-up related study, Tennis, Thornton, and Filer (2012) discussed versioning of classification schemes and categorize the changes as structural changes (differences in classification structures), word-use changes (new concepts or classes added), and textual changes (literary warrant differences, document-set differences). They observed the changes of subjects in Dewey Decimal Classification in a 10-year time interval and concluded that changes are largely human (cataloger) decision-making process over time. Tennis (2012) further examined the subject term “Eugenics” in a case study, and defines the bibliographic subject changes over time as “subject ontogeny”.

Analogously, in biodiversity classification systems, systematists define the position of a taxon or a taxon concept in the tree of life to their best judgements at that certain

moment in time; and that judgement could alter over time by other systematists, or even by themselves (Franz, Chen, et al., 2016). Bowker (2000b), in describing this phenomenon in biodiversity classifications, concluded that though taxon names in a phylogeny are unstable over time, there is always a tendency to keep the names the same. The seemingly stability in taxon name entails that museum curators or specialists will not have to re-classify or relabel the specimens and taxidermy in their museums. As much as the strong pressure in keeping taxon names the same, Franz, Chen, et al. (2016) examined 11 classifications spanning across 126 years that shares the same name of a type of grass *Andropogon glomeratus-virginicus* ‘complex’, and concluded that names are unreliable. The authors further explained that there is a dissociation between name and meaning across time: Taxon names, while staying same across years, could be referring to different species.

What amplifies the uncertainties, is when subject name changes or taxon name changes are accompanied with geographic name changes over time. Buckland and Lancaster (2004) explained the instabilities of place-names and geographic information, and how it is important to reference place-names in conjunction with “time-qualified gazetteer entries”. However, Bishop et al. (2015) mentioned that even time related resources can be problematic. Formal time documentation on specific dates can have confusion when date formats are ambiguous, or when different calendar systems are used; informal time documentation such as time periods could also be useful, but it can only go so far to approximate a certain time frame. It is all more challenging when place-names are used with multiple historical maps (or in Bowker (2000b)’s case, biodiversity maps), where no one *time* classification structure could cover all granularity of different eras, empires, or periods of times.

### 8.2.3 Sociotechnical Aspects in Biodiversity Data

Extant literature on mapping biodiversity recognizes the interconnection between the social construct of geography and the technical systems of biodiversity data management (Boakes

et al., 2010; Bowker, 2000b; Malavasi, 2020).

The social construct of geography is rooted in Foucault’s notion of governmentality and biopolitics, in which he asserts that the government defines the borders, and that “neutral knowledge simply represents the governing reality” (Lemke, 2001). Biologists draw on this lens and note that scientific knowledge is not apolitical (Biermann & Mansfield, 2014; Oksanen & Vuorisalo, 2019). Oksanen and Vuorisalo (2019) discuss how geopolitical borders and sovereign states play important parts in conservation policy-making to determine historic range and conservation status of species. Biermann and Mansfield (2014), extending Foucault’s view on biopolitics to consider human to nonhuman(nature) relations, raise questions on whether truths exist in human’s decision on determining borders in biodiversity conservation. Based on these prior literature, it is important to understand how these human decisions influence the design and usage of biodiversity information systems.

Existing biodiversity information systems such as the Global Biodiversity Information Facility<sup>9</sup> (GBIF), the Biodiversity Heritage Library<sup>10</sup> (BHL), and others, with GBIF being the largest, contain biodiversity data and literature that can be based on different geopolitical assumptions. For instance, Smith, Benson, Chen, Yamada, and Mims (2020) find discrepancies in the freshwater fish occurrences based on GBIF and NatureServe maps because the watershed boundaries are different in the two maps. Ewers-Saucedo et al. (2021), in examining 200 years of faunal change in natural history museum collections, posit that some species were not well-represented due to the shifting borders at the North Sea regions during the World Wars. Girardello et al. (2019)’s study discovers that GBIF underrepresented samples in Central Africa because geopolitical events hinder researchers to collect data. Altogether, these studies exhibit that the technical aspects of biodiversity are not void of social and political concerns.

Multiple challenges arise when geographic information is coupled with time and changes

---

<sup>9</sup>GBIF:<https://www.gbif.org/>

<sup>10</sup>BHL:<https://www.biodiversitylibrary.org/>

in the context of biodiversity data (Bowker, 2000b; Buckland & Lancaster, 2004). The inherent assumptions and biases in geographic information are weaved into biodiversity data and information systems, and these biases persist through time when species names, places names, classification systems, and times undergo changes (Bowker, 2000b; Cragin, 2003). Biodiversity data, though seemingly objective, is not void of sociotechnical and political problems, resonating Bowker (2000b)’s view that “politically charged decisions” are always involved in scientific data, and that “the political empire writ large on the natural world”.

## 8.3 Proposed Conceptual Model

### 8.3.1 Overview

Conceptual models and frameworks are graphical representations used to understand real-world phenomena (Jabareen, 2009; Wand & Weber, 2002). From static objects to dynamic events and processes, the goal of constructing a conceptual model is to offer possible interpretations to social realities rather than to deliver outcome-oriented explanations (Jabareen, 2009; Wand & Weber, 2002). In surveying conceptual models researchers and businesses used in practice, Fettke (Fettke (2009) found that ER (Entity-Relation) diagrams are used the most, followed by UML (Unified Modeling Language) diagrams and workflow diagrams. In the field of Information Sciences, there are a variety of research on conceptual modeling for specific topics, some used ER diagrams, like conducting online design (K. J. Lee et al., 2021); some worked with UML diagrams, such as Jett, Sacchi, Lee, and Clarke’s model for video games (Jett, Sacchi, Lee, & Clarke, 2016). Some others depicted their conceptual model for librarianship design partnerships in a comparison chart (Yip, Lee, & Lee, 2020). Yet some others provided multi-level diagram, such as Littletree, Belarde-Lewis, and Duarte (2020)’s Indigenous systems of knowledge conceptual model. These studies demonstrated that different graphical representations are equally effective in perceiving the world’s entities.

The conceptual model proposed in this chapter uses graphical representations similar to workflow diagrams, presenting the different phases in modeling the historical sovereignty within biodiversity data. We demonstrate the feasibility of this model by using a real-world species use case and its occurrences in a specific region in the world. In the following sections, we first describe each phase of our proposed conceptual model for aligning historical sovereignties taxonomies (Figure 8.4), then we provide a use case that implements this conceptual model.

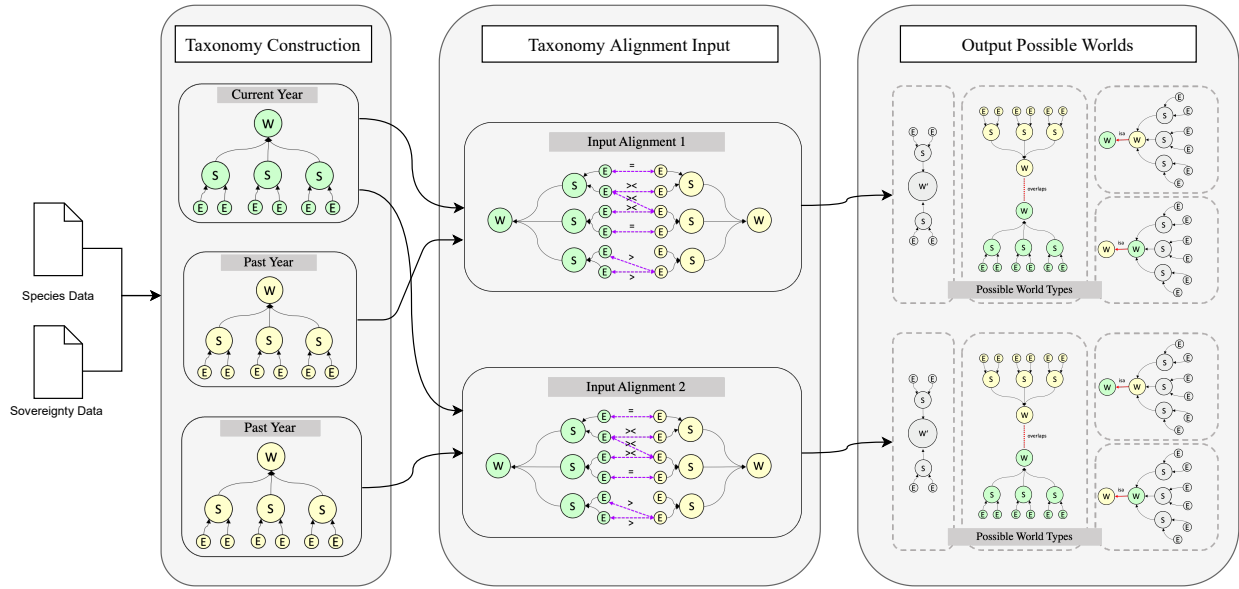


Figure 8.4: Proposed conceptual model for aligning historical sovereignties taxonomies  
Four phases of this conceptual mode: (1) Species and Sovereignty Data; (2) Taxonomy Construction; (3) Taxonomy Alignment Input; (3) Output Possible Worlds

### 8.3.2 Phase 1: Species and Sovereignty Data

#### Species Data

Given that biologists usually work with one species at a time, this phase of the conceptual model focuses on collecting data of one particular species. Species are usually endemic to a particular range of geographic locations, so species data in this phase is used to narrow

down the geographic points of interest.

Species data can be collected from individual institutions, or from aggregated databases such as the Global Biodiversity Information Facility (GBIF). Most species datasets use Darwin Core metadata standard<sup>11</sup> to annotate and document the species occurrence datasets. The datasets should include the following core metadata elements: the scientific names of the species; country code and locality where the species were spotted; the year of the collection (sighting); institution code specifying who owns the occurrence records.

### Sovereignty Data

Historical sovereignty data can be collected via existing taxonomies, entry lists, or other knowledge organization systems relating to geographic areas and place-names. Examples of these include the CIA World Factbook<sup>12</sup>, Encyclopedia Britannica<sup>13</sup>, Getty Thesaurus of Geographic Names<sup>14</sup>, Geonames<sup>15</sup>, and place-specific historical timeline page<sup>16</sup>. There are also other efforts to incorporate time changes in geographic entities: the World Spatiotemporal analysis project that utilizes ontology to document geographic entity name changes over time (Stewart et al., 2015) and the PeriodO gazetteer<sup>17</sup>(Rabinowitz, Shaw, Buchanan, Golden, & Kansa, 2016).

However, no one source could systematically and chronologically document all geographic entities, especially when the geographic entities are in a conflicting state or given shifting geopolitical disposition. For instance, in the PeriodO project, given that the majority of its data comes from the Library of Congress Subject Headings (LCSH), some geographic entities may be unsearchable because the LCSH is based on the ISO-3166 standard. The ISO-3166

---

<sup>11</sup>Darwin Core:<https://dwc.tdwg.org/>

<sup>12</sup>CIA World Factbook: <https://www.cia.gov/the-world-factbook/>

<sup>13</sup>Britannica:<https://www.britannica.com/>

<sup>14</sup>Getty Thesaurus of Geographic Names: <https://www.getty.edu/research/tools/vocabularies/tgn/>

<sup>15</sup>Geonames:<http://www.geonames.org/>

<sup>16</sup>Taiwan government website on history: [https://www.taiwan.gov.tw/content\\_3.php](https://www.taiwan.gov.tw/content_3.php)

<sup>17</sup>PeriodO:<https://client.perio.do/?page=About>

standard has its implicit assumptions on which areas or geographic entities are contested in the world, as explained in Chapter 6. Suggested by Bishop et al. (2015) and Buckland and Lancaster (2004), time-period directories such as Wikipedia’s sovereignty themed lists may be the most useful to view changes in place-names and collect sovereignty data.

### 8.3.3 Phase 2: Taxonomy Construction

The species and sovereignty data collected in Phase 1 are usually in a tabular format. Tabular data model is an intuitive form that allows the viewers to take in all the information at once; viewers tend to read the tables from left-to-right, as the table designers intended them to (Van Hooland & Verborgh, 2014). In our prior work (Cheng, 2021), we discuss the differences between tabular and tree models in mapping projects. Tree models (e.g., taxonomies) are useful to structure hierarchical information and permits more flexible views. For instance, depending on how the taxonomy is presented, it can be read from left-to-right, right-to-left, top-down, or bottom-up. When comparing multiple taxonomies, a tree model representation of mapping usually does not assume source-target comparisons, unless there are explicit information labeled on the taxonomies. A tabular model, on the other hand, by presenting information side-by-side in columns, assumes equivalency between terms; and the first column is usually the “anchor” metadata standard. These reasons, along with the hierarchical information and relational data embedded in multiple tables associated with species data and sovereignty data, explain why a taxonomy (tree model) representation is needed.

#### The Definition of a Taxonomy

Recall in this dissertation, 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. Interpretations on the concepts and edges of the taxonomies is discussed in Chapter 4.5.

## Constructing Historical Sovereignty Taxonomies

This phase requires a transformation from the tabular sovereignty data into taxonomies for a given year that relates the geographic entity(areas) and the political sovereignty of that entity. In Figure 8.5, the root node of the taxonomy is the entire area “the World” where every node is a part of. *Sovereignty* means the political entity that governs certain coverage of geographic areas. *Geographic entity* is the physical land mass that a sovereignty covers, excluding areas in maritime zones. In a given year, the world may consist of many political sovereignties, in which each sovereignty oversees certain geographic entities. Each is-a edge in this taxonomy can be interpreted as “part of” relation, and each node in this taxonomy can be interpreted as a class.

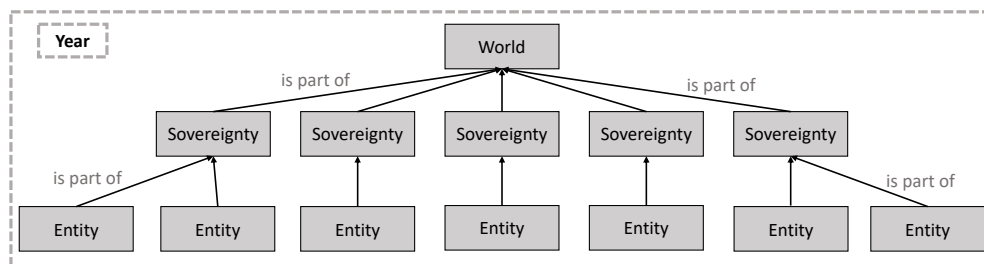


Figure 8.5: The taxonomy of sovereignty and geographic entities in a given year

Taxonomies can be extracted from datasets, country code lists, and/or metadata standards

### 8.3.4 Phase 3: Taxonomy Alignment Input

Because the focus of the conceptual model is to find the historical sovereignties of a modern state, this phase is to relate the modern to the historical by comparing the taxonomies of the world’s sovereignties of now to the past. The process of comparing the taxonomies is called *taxonomy alignment*.

Given two (or more) taxonomies, here we relate the taxonomies using a logic-based approach, where  $T_1$ ,  $T_2$  are connected through a set of input RCC-5 relations to yield a “merged” taxonomy  $T_3$ . RCC-5 (Region Connection Calculus) are five base relations: equivalent ( $==$ ),

disjointness (!), overlap ( $><$ ), proper part ( $<$ ), and proper part inverse ( $>$ ). Examples of taxonomy alignment using RCC-5 in different applications are demonstrated in our prior work (Cheng et al., 2017a; Cheng & Ludäscher, 2019b, 2020b; Thomer, Cheng, Schneider, Twidale, & Ludäscher, 2017). Figure 8.6 is an illustration of how the alignment process looks like.

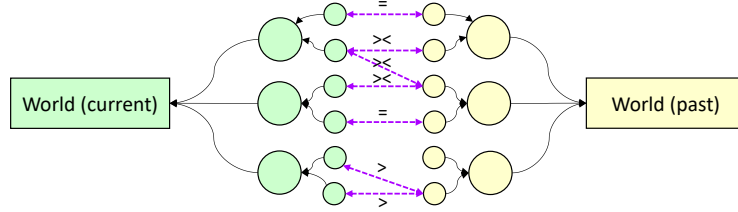


Figure 8.6: Input alignment for  $T_{Current}$  and  $T_{Past}$

The green nodes are in the taxonomy of the current world, and the yellow nodes are in the taxonomy of the past world of a given year. The taxonomies are connected via purple dotted lines, marked by one of the five RCC-5 relations.

### 8.3.5 Phase 4: Output Possible Worlds

Possible worlds (PW) are the output merged solution(s)  $T_3$  of a taxonomy alignment problem. Any solution taxonomy  $T_3$  must satisfy all logic constraints implied by the input taxonomies  $T_1$ ,  $T_2$ , and must reveal all pairwise relationships between concepts in  $T_1$  and  $T_2$ . The results of the alignment may have zero, one, or many PWs. In the case of zero PW, it suggests that there are some logic inconsistencies in the input alignments, or the relations needed to be resolved to perform the alignment. In the case of one PW, there is only one unique logical combination to merge the input taxonomies. In the case of multiple PWs, it means that more than one merged solutions are valid to logically relate the taxonomies.

In relating the taxonomies of historical sovereignties and modern sovereignties, there are four types of possible worlds, shown and explained in Figure 8.7.

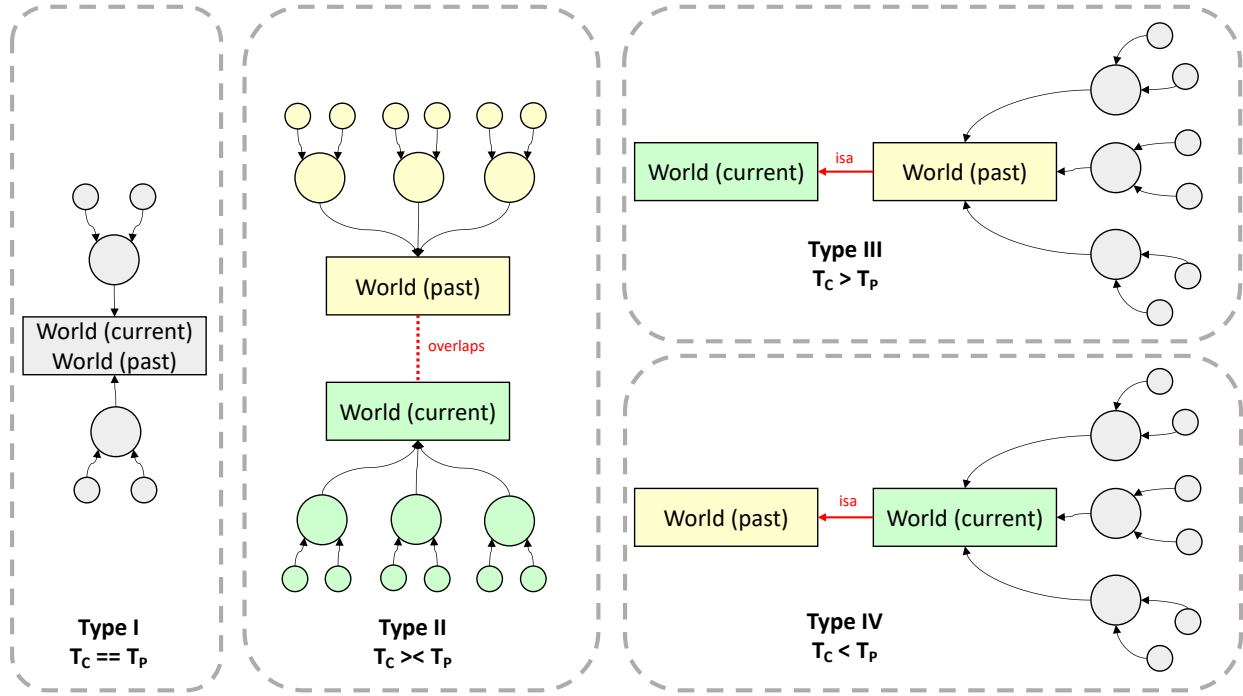


Figure 8.7: Four types of possible worlds

Type I ( $T_C == T_P$ ): The Current World is congruent with a Past World; Type II ( $T_C >> T_P$ ): The Current World is overlapping with a Past World; Type III ( $T_C > T_P$ ): The Current World contains a Past World; Type IV ( $T_C < T_P$ ): The Current World is a part of a Past World. Note: these are illustrations of how the possible worlds look like; the real possible worlds may look different on the lower and leaf nodes.

## 8.4 Use Case

We demonstrate the use of the proposed conceptual model by using a real-world species use case and its occurrences in a specific region in the world. All data used in this study can be found in Cheng (2022).

### 8.4.1 Phase 1: Species and Sovereignty Data

#### Species Data

We focus on one species as a use case in this research – a type of endangered wading bird in the spoonbill family that is endemic to East Asia.

Table 8.1: Abridged species occurrence dataset of *Platalea minor*

| Dataset Key | Scientific Name                          | Country Code | Year | Institution Code |
|-------------|------------------------------------------|--------------|------|------------------|
| 1           | Platalea minor Temminck & Schlegel, 1849 | CN           | 1884 | AMNH             |
| 2           | Platalea minor Temminck & Schlegel, 1849 | CN           | 1895 | MCZ              |
| 3           | Platalea minor Temminck & Schlegel, 1849 | HK           | 1982 | CLO              |
| 4           | Platalea minor Temminck & Schlegel, 1849 | HK           | 1983 | CLO              |
| 5           | Platalea minor Temminck & Schlegel, 1849 | JP           | 1916 | MNHAH            |
| 6           | Platalea minor Temminck & Schlegel, 1849 | JP           | 1957 | HCSM             |
| 7           | Platalea minor Temminck & Schlegel, 1849 | KH           | 2010 | CLO              |
| 8           | Platalea minor Temminck & Schlegel, 1849 | KH           | 2014 | CLO              |
| 9           | Platalea minor Temminck & Schlegel, 1849 | KP           | 2016 | CLO              |
| 10          | Platalea minor Temminck & Schlegel, 1849 | KP           | 2017 | CLO              |
| 11          | Platalea minor Temminck & Schlegel, 1849 | KR           | 1884 | USNM             |
| 12          | Platalea minor Temminck & Schlegel, 1849 | KR           | 1910 | YIO              |
| 13          | Platalea minor Temminck & Schlegel, 1849 | MO           | 2002 | CLO              |
| 14          | Platalea minor Temminck & Schlegel, 1849 | MO           | 2014 | CLO              |
| 15          | Platalea minor Temminck & Schlegel, 1849 | PH           | 1914 | USNM             |
| 16          | Platalea minor Temminck & Schlegel, 1849 | PH           | 1927 | USNM             |
| 17          | Platalea minor Temminck & Schlegel, 1849 | TH           | 2003 | CLO              |
| 18          | Platalea minor Temminck & Schlegel, 1849 | TH           | 2003 | CLO              |
| 19          | Platalea minor Temminck & Schlegel, 1849 | TW           | 1923 | NSMT             |
| 20          | Platalea minor Temminck & Schlegel, 1849 | TW           | 1985 | CLO              |
| 21          | Platalea minor Temminck & Schlegel, 1849 | VN           | 1970 | CLO              |
| 22          | Platalea minor Temminck & Schlegel, 1849 | VN           | 1996 | CLO              |

The dataset lists the two earliest occurrences of each country. Data collected from GBIF in January 2022.

**Data Collection.** We conduct a scientific name search on GBIF of *Platalea minor* Temminck & Schlegel, 1849. At time of data collection in January 2022, GBIF returned one aggregated dataset with 27,086 occurrences from 50 different GBIF partnering institu-

tions. The occurrences of *Platalea minor* ranges from 1884-present, and from 12 different geographic entities (GBIF.org, 2022). Table 8.1 shows an abridged version of the dataset listing the earliest occurrences by country code.

**Data Cleaning.** We use the open-source software OpenRefine<sup>18</sup> (version 3.4.1) for preliminary data exploration and data cleaning. To address our research goals, in which the time and the geographic entities of the species sightings are crucial, we focus data cleaning on the *year* and *country code* columns in the dataset. We eliminate 661 rows from the dataset because they were missing year values and 93 rows with blank country code values. We eliminate the two occurrences outside our geographic area of interest, East Asia. The cleaned dataset contains 26,410 occurrences of *Platalea minor* Temminck & Schlegel, 1849. Table 8.2 shows the number of occurrences of *Platalea minor* per geographic entity.

Table 8.2: Number of occurrences per geographic entity

| Entity      | Country code | Number of Occurrences |
|-------------|--------------|-----------------------|
| Taiwan      | TW           | 19,402                |
| Hong Kong   | HK           | 3,033                 |
| South Korea | KR           | 1,546                 |
| China       | CN           | 1,268                 |
| Japan       | JP           | 883                   |
| Thailand    | TH           | 86                    |
| Philippines | PH           | 76                    |
| Vietnam     | VN           | 70                    |
| Macau       | MO           | 41                    |
| North Korea | KP           | 3                     |
| Cambodia    | KH           | 2                     |

Entity listed by the order of the most to the least number of occurrences.

## Historical Sovereignties Data

**Data Collection.** For this use case, we elect to using Wikipedia’s List of predecessors of sovereign states in Asia <sup>19</sup> to collect historical sovereignty data about East Asia. We collect

<sup>18</sup>OpenRefine: <https://openrefine.org/>

<sup>19</sup>Wikipedia’s sovereign states in Asia: [https://en.wikipedia.org/wiki/List\\_of\\_predecessors\\_of\\_sovereign\\_states\\_in\\_Asia](https://en.wikipedia.org/wiki/List_of_predecessors_of_sovereign_states_in_Asia)

place-name and sovereignty changes information from 11 entities in East Asia from the year 1880 to 2022. These entities corresponds to the geographic entities appeared in the species data, including: Cambodia, China, Hong Kong, Japan, Macau, North Korea, Philippines, South Korea, Taiwan, Thailand, Vietnam. For each of the entities, we read its entry from the Wikipedia’s sovereign states list to determine its historical sovereignties and the geographic boundary the sovereignty covers at a certain point of time. We present here one example of this data in Table 8.3 and 8.4.

Table 8.3: Modern day Japan

| Year | Sovereignty     | Geographic Areas Covered                  |
|------|-----------------|-------------------------------------------|
| 1868 | Empire of Japan | Japan                                     |
| 1895 | Empire of Japan | Japan, Taiwan                             |
| 1910 | Empire of Japan | Japan, Taiwan, Korea                      |
| 1941 | Empire of Japan | Japan, Taiwan, Korea, Hong Kong           |
| 1945 | Empire of Japan | Japan, Taiwan, Korea, Hong Kong, Cambodia |
| 1947 | Nihonkoku       | Japan                                     |

Data includes Japan’s sovereignty and geographic areas from 1868-1947.

**Data Modeling.** We organize the 11 geographic entities in a tabular data model. Given that each entity has its own historical timeline and differences in years in terms of sovereignty change, we model this all-in-one view with time slices of 10 year intervals. The species occurrence dataset we collected spans from 1884 to now, so the time slices begin at 1880. For instance, in the year 1940, Cambodia, Hong Kong, Macau and Taiwan were under the sovereignties of French Indochina, United Kingdom of Great Britain and Northern

Table 8.4: Modern day Vietnam

| Year      | Sovereignty                    | Geographic Areas Covered                        |
|-----------|--------------------------------|-------------------------------------------------|
| 1802      | Nguyen Dynasty                 | Vietnam                                         |
| 1887-1945 | French Indochina               | Vietnam                                         |
| 1862-1949 | Colony of Cochinchina          | Southern Vietnam                                |
| 1884-1949 | Annam Protectorate             | Central Vietnam                                 |
| 1884-1949 | Tonkin Protectorate            | Northern Vietnam                                |
| 1949-1955 | State of Vietnam               | Overlapping with Democratic Republic of Vietnam |
| 1945-1976 | Democratic Republic of Vietnam | North Vietnam                                   |
| 1955-1975 | Republic of Vietnam            | South Vietnam                                   |
| 1976      | Socialist Republic of Vietnam  | Vietnam                                         |

Data includes Vietnam’s sovereignty and geographic areas change from 1802-1976.

Ireland, Portuguese Republic, and Empire of Japan respectively. In 1950, the same four entities' sovereignties were Kingdom of Cambodia, United Kingdom of Great Britain and Northern Ireland, Portuguese Republic, and Republic of China, respectively, with Cambodia and Taiwan experiencing sovereignty change after the World War II, while Hong Kong and Macau stayed under the same sovereignty. Table 8.5 is one example table illustrating each geographic entity (the actual areas) with the sovereignty covers it every 10 years from 1880.

#### 8.4.2 Phase 2: Taxonomy Construction

We relate information in Tables 8.3, 8.4 8.5, and 8.6, and transform them into 14 sovereignty taxonomies by a given year. The time slices of these taxonomies start from 1880 and up until 2022, with 10-year intervals in between each taxonomy. The most current year is 2022, and the taxonomies of 2010, 2020 are exactly the same as the year 2000, so we leave 2010 and 2020 out of the analysis. We also model the taxonomies with placeholder “non-coverage” nodes, describing at the sovereignty-level to indicate a sovereignty may have other unclaimed territories or areas not considered by this study. For example, Macau was part of the Portuguese Republic in 1940, but Portuguese Republic also has other geographic entity in the world that is not in East Asia (e.g. Portugal mainland). We hence add a placeholder “nc\_Portugal” node to describe these areas for keeping consistencies in the taxonomies. In total, we construct 14 taxonomies for this use case (model taxonomies shown in Figure 8.8. The taxonomies are shown in Figure 8.9 and 8.10, where the 2022 taxonomy is marked in green, and the 1940 taxonomy is marked in yellow.

Table 8.5: East Asia Sovereignties from 1880 to 1950

| Geographic Area    | Year                                        |                                                                                 |                                                                                 |                                                                                 |                                                                                 |                                                      |
|--------------------|---------------------------------------------|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------|---------------------------------------------------------------------------------|------------------------------------------------------|
|                    | 1880                                        | 1890                                                                            | 1900                                                                            | 1910                                                                            | 1920                                                                            | 1930                                                 |
| <b>Cambodia</b>    | French Indochina                            | French Indochina                                                                | French Indochina                                                                | French Indochina                                                                | French Indochina                                                                | French Indochina                                     |
| <b>China</b>       | Qing Dynasty China                          | Qing Dynasty China                                                              | Qing Dynasty China                                                              | Qing Dynasty China                                                              | Republic of China                                                               | Nationalist government of Republic of China          |
| <b>Hong Kong</b>   | United Kingdom of Great Britain and Ireland | United Kingdom of Great Britain and Ireland                                     | United Kingdom of Great Britain and Ireland                                     | United Kingdom of Great Britain and Ireland                                     | United Kingdom of Great Britain and Northern Ireland                            | United Kingdom of Great Britain and Northern Ireland |
| <b>Japan</b>       | Empire of Japan                             | Empire of Japan                                                                 | Empire of Japan                                                                 | Empire of Japan                                                                 | Empire of Japan                                                                 | Empire of Japan                                      |
| <b>Macau</b>       | Portuguese Republic                         | Portuguese Republic                                                             | Portuguese Republic                                                             | Portuguese Republic                                                             | Portuguese Republic                                                             | Portuguese Republic                                  |
| <b>North Korea</b> | Kingdom of Great Joseon                     | Kingdom of Great Joseon                                                         | Korean Empire                                                                   | Empire of Japan                                                                 | Empire of Japan                                                                 | Democratic People's Republic of Korea                |
| <b>Philippines</b> | Spanish Empire                              | Spanish Empire                                                                  | United States Military Government                                               | Insular Government of the Philippine Islands                                    | Insular Government of the Philippine Islands                                    | Commonwealth of the Philippines                      |
| <b>South Korea</b> | Kingdom of Great Joseon                     | Kingdom of Great Joseon                                                         | Korean Empire                                                                   | Empire of Japan                                                                 | Empire of Japan                                                                 | Republic of Korea                                    |
| <b>Taiwan</b>      | Qing Dynasty China                          | Qing Dynasty China                                                              | Empire of Japan                                                                 | Empire of Japan                                                                 | Empire of Japan                                                                 | Empire of Japan                                      |
| <b>Thailand</b>    | Rattanakosin Kingdom of Siam                | Rattanakosin Kingdom of Siam                                                    | Rattanakosin Kingdom of Siam                                                    | Rattanakosin Kingdom of Siam                                                    | Rattanakosin Kingdom of Siam                                                    | Kingdom of Thailand                                  |
| <b>Vietnam</b>     | Nguyen Dynasty                              | French Indochina, Colony of Cochinchina, Annam Protectorate, Tokin Protectorate | French Indochina, Colony of Cochinchina, Annam Protectorate, Tokin Protectorate | French Indochina, Colony of Cochinchina, Annam Protectorate, Tokin Protectorate | French Indochina, Colony of Cochinchina, Annam Protectorate, Tokin Protectorate | State of Vietnam, Democratic Republic of Vietnam     |

Table 8.6: East Asia Sovereignties from 1960 to 2022

| Geographic Area    | Year                                                 |                                                      |                                                      |                                                      |                                                  |                                                  |                                                  |                                                  |  |  |
|--------------------|------------------------------------------------------|------------------------------------------------------|------------------------------------------------------|------------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------------------------|--|--|
|                    | 1960                                                 | 1970                                                 | 1980                                                 | 1990                                                 | 2000                                             | 2010                                             | 2020                                             | 2022                                             |  |  |
| <b>Cambodia</b>    | Kingdom of Cambodia                                  | Khmer Republic                                       | People's Republic of Kampuchea                       | State of Cambodia                                    | Kingdom of Cambodia                              | Kingdom of Cambodia                              | Kingdom of Cambodia                              | Kingdom of Cambodia                              |  |  |
| <b>China</b>       | People's Republic of China                           | People's Republic of China                           | People's Republic of China                           | People's Republic of China                           | People's Republic of China                       | People's Republic of China                       | People's Republic of China                       | People's Republic of China                       |  |  |
| <b>Hong Kong</b>   | United Kingdom of Great Britain and Northern Ireland | United Kingdom of Great Britain and Northern Ireland | United Kingdom of Great Britain and Northern Ireland | United Kingdom of Great Britain and Northern Ireland | Hong Kong Special Administrative Region of China | Hong Kong Special Administrative Region of China | Hong Kong Special Administrative Region of China | Hong Kong Special Administrative Region of China |  |  |
| <b>Japan</b>       | Nihonkoku                                            | Nihonkoku                                            | Nihonkoku                                            | Nihonkoku                                            | Nihonkoku                                        | Nihonkoku                                        | Nihonkoku                                        | Nihonkoku                                        |  |  |
| <b>Macau</b>       | Portuguese Republic                                  | Portuguese Republic                                  | Portuguese Republic                                  | Portuguese Republic                                  | Macau Special Administrative Region of China     | Macau Special Administrative Region of China     | Macau Special Administrative Region of China     | Macau Special Administrative Region of China     |  |  |
| <b>North Korea</b> | Democratic People's Republic of Korea                | Democratic People's Republic of Korea                | Democratic People's Republic of Korea                | Democratic People's Republic of Korea                | Democratic People's Republic of Korea            | Democratic People's Republic of Korea            | Democratic People's Republic of Korea            | Democratic People's Republic of Korea            |  |  |
| <b>Philippines</b> | Republic of the Philippines                          | Republic of the Philippines                          | Republic of the Philippines                          | Republic of the Philippines                          | Republic of the Philippines                      | Republic of the Philippines                      | Republic of the Philippines                      | Republic of the Philippines                      |  |  |
| <b>South Korea</b> | Republic of Korea                                    | Republic of Korea                                    | Republic of Korea                                    | Republic of Korea                                    | Republic of Korea                                | Republic of Korea                                | Republic of Korea                                | Republic of Korea                                |  |  |
| <b>Taiwan</b>      | Republic of China                                    | Republic of China                                    | Republic of China                                    | Republic of China                                    | Republic of China                                | Republic of China                                | Republic of China                                | Republic of China                                |  |  |
| <b>Thailand</b>    | Kingdom of Thailand                                  | Kingdom of Thailand                                  | Kingdom of Thailand                                  | Kingdom of Thailand                                  | Kingdom of Thailand                              | Kingdom of Thailand                              | Kingdom of Thailand                              | Kingdom of Thailand                              |  |  |
| <b>Vietnam</b>     | Democratic Republic of Vietnam, Republic of Vietnam  | Democratic Republic of Vietnam, Republic of Vietnam  | Socialist Republic of Vietnam                        | Socialist Republic of Vietnam                        | Socialist Republic of Vietnam                    | Socialist Republic of Vietnam                    | Socialist Republic of Vietnam                    | Socialist Republic of Vietnam                    |  |  |

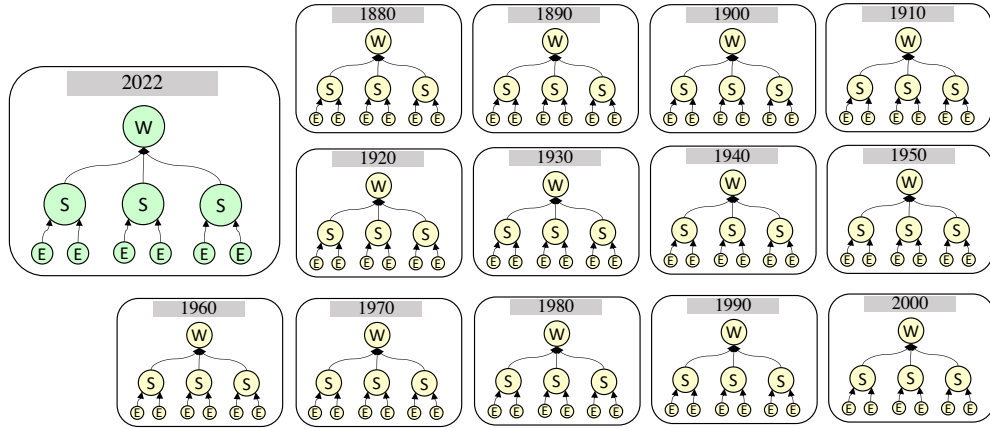


Figure 8.8: The 14 taxonomies constructed in this study

W:World; S:Sovereignty; E:Geographic Entity

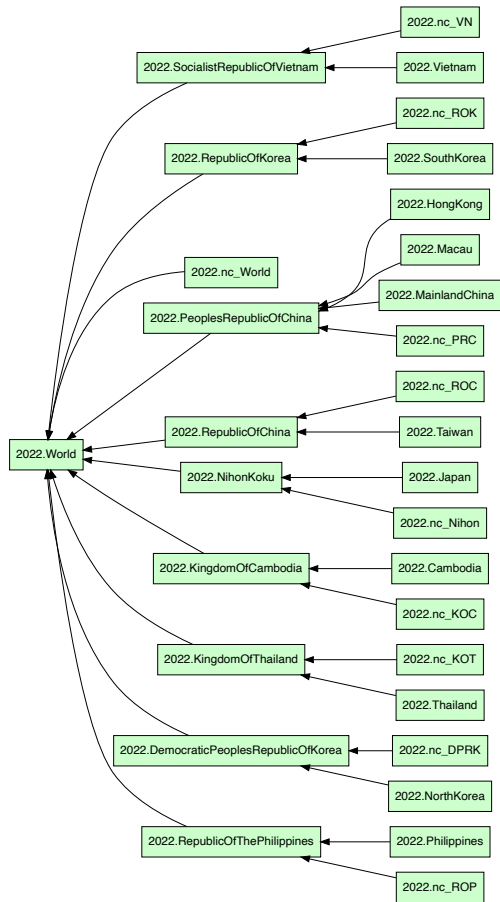


Figure 8.9: East Asia in 2022

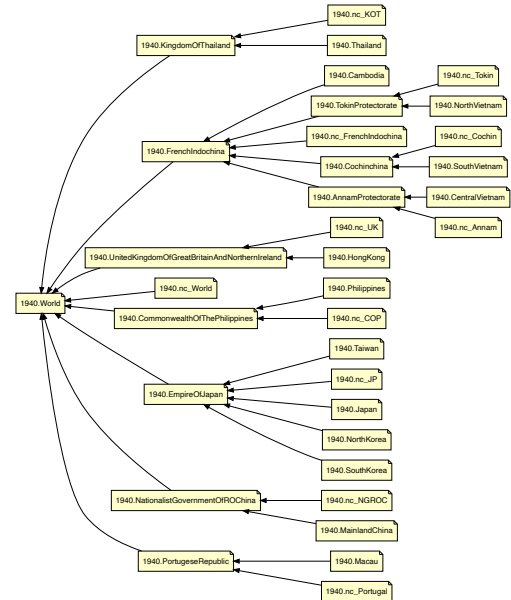


Figure 8.10: East Asia in 1940

### 8.4.3 Phase 3: Taxonomy Alignment Input

For this use case, we construct 13 input alignments (2022 vs 1880; 2022 vs 1890;...) in total. One example of the taxonomy alignment inputs is shown in Figure 8.12. In each alignment, the geographic entities (in the lowest level nodes) of one taxonomy are marked as equivalent to its counterpart in another taxonomy. The non-coverage nodes of a particular sovereignty from the “past year taxonomies” would be considered as part of the non-coverage nodes of the World in the current year, and vice versa.

We focus on comparing a taxonomy from the past year ( $T_P$ ) with the current year ( $T_C$ ). As illustrated in Figure 8.6, the taxonomies we constructed are linked by a set of RCC-5 relations in purple dotted lines. We connect the lowest level node in this case because these nodes represent the main-lands of a geographic entity, and are less ambiguous to determine how one entity relates to another entity from a different time. In total, we constructed 13 input alignments (2022 vs 1880; 2022 vs 1890; 2022 vs 1900; ...) (model alignment figure in Figure 8.11). In each alignment, the lowest level nodes in one taxonomy are usually marked as equivalent to its counterpart in another taxonomy, since the main geographic areas covered are the same in most cases. The non-coverage nodes of a particular sovereignty from the “past year taxonomies” would be considered as part of the non-coverage nodes of the World in the current year, and vice versa. An example of the 33 articulations made for the 2022-1940 alignment is as follows:

[2022.Taiwan equals 1940.Taiwan]

[2022.MainlandChina equals 1940.MainlandChina]

[2022.HongKong equals 1940.HongKong]

[2022.Macau equals 1940.Macau]

[2022.Japan equals 1940.Japan]

[2022.NorthKorea equals 1940.NorthKorea]

[2022.SouthKorea equals 1940.SouthKorea]

[2022.Cambodia equals 1940.Cambodia]  
[2022.Philippines equals 1940.Philippines]  
[2022.Thailand equals 1940.Thailand]  
[2022.Vietnam includes 1940.NorthVietnam]  
[2022.Vietnam includes 1940.CentralVietnam]  
[2022.Vietnam includes 1940.SouthVietnam]  
[2022.nc\_World includes 1940.nc\_NGROC]  
[2022.nc\_World includes 1940.nc\_JP]  
[2022.nc\_World includes 1940.nc\_COP]  
[2022.nc\_World includes 1940.nc\_KOT]  
[2022.nc\_World includes 1940.nc\_UK]  
[2022.nc\_World includes 1940.nc\_Portugal]  
[2022.nc\_World includes 1940.nc\_FrenchIndochina]  
[2022.nc\_World includes 1940.nc\_Tokin]  
[2022.nc\_World includes 1940.nc\_Annam]  
[2022.nc\_World includes 1940.nc\_Cochin]  
[2022.nc\_ROC is\_included\_in 1940.nc\_World]  
[2022.nc\_PRC is\_included\_in 1940.nc\_World]  
[2022.nc\_Nihon is\_included\_in 1940.nc\_World]  
[2022.nc\_DPRK is\_included\_in 1940.nc\_World]  
[2022.nc\_ROK is\_included\_in 1940.nc\_World]  
[2022.nc\_KOC is\_included\_in 1940.nc\_World]  
[2022.nc\_ROP is\_included\_in 1940.nc\_World]  
[2022.nc\_KOT is\_included\_in 1940.nc\_World]  
[2022.nc\_VN is\_included\_in 1940.nc\_World]

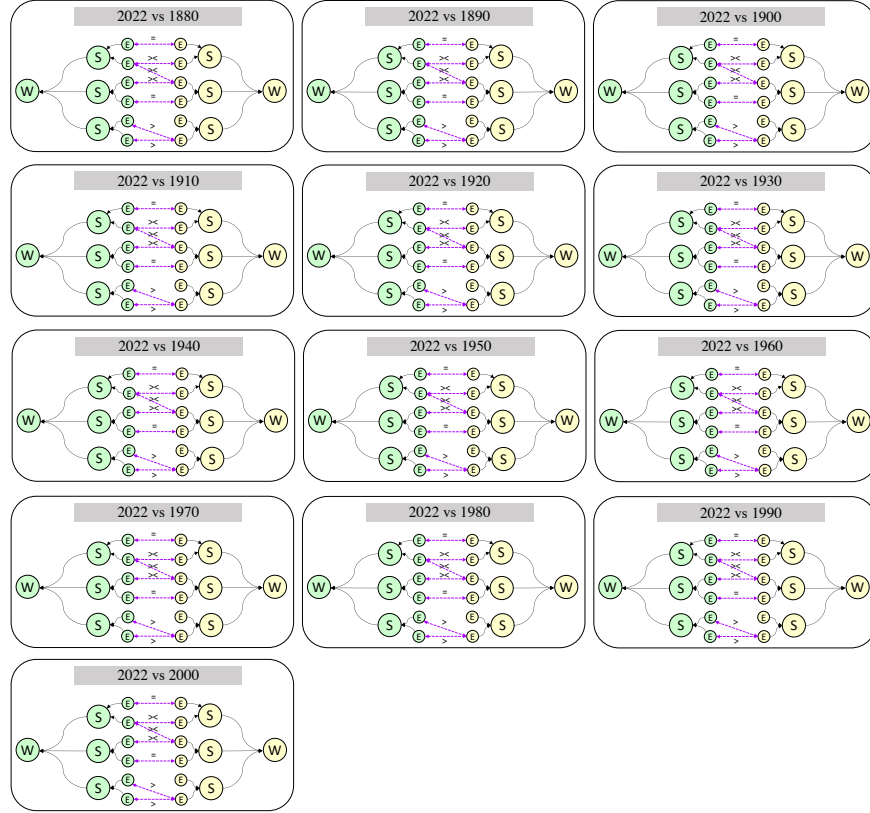


Figure 8.11: 13 input alignments for  $T_C$  and  $T_P$

Figures are shown for illustrative purposes: each input alignment's articulations (purple lines and the RCC-5 notations) would be different from one another. W: World; S: Sovereignty; E: Geographic Entity

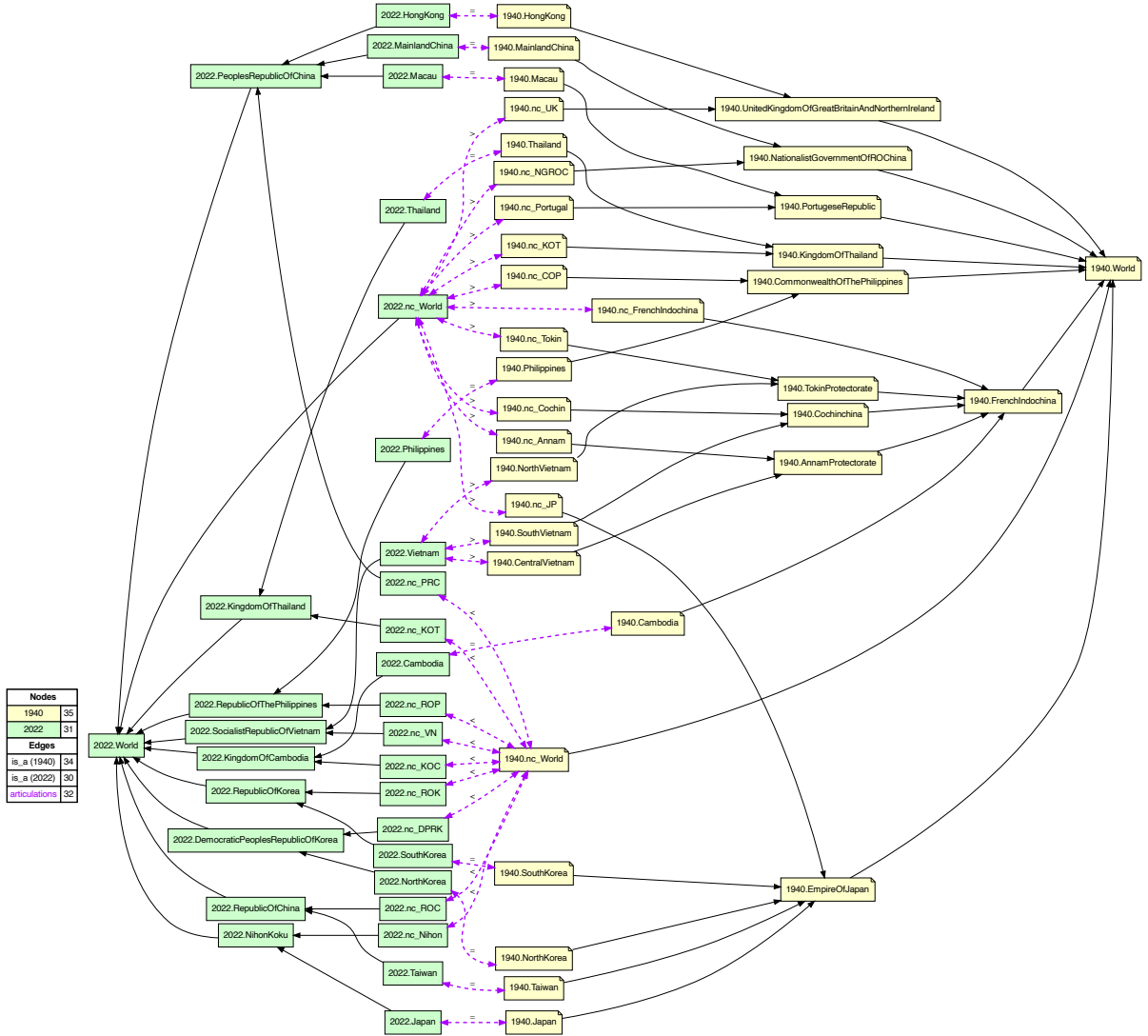


Figure 8.12: One of the 13 alignments:  $T_{2022}$  and  $T_{1940}$

#### 8.4.4 Phase 4: Output Possible Worlds

For this use case, we analyze all 13 alignments' output possible worlds in terms of the type of possible worlds each alignment produced. In Table 8.7, we present the number of possible worlds in each of the four types of PWs and the total number of the PWs for each alignment.

**Across Alignments.** Alignment 1 to 7 all yield the same results with a total of 10 possible worlds, and 2, 3, 3, 2 in Type I, II, III, IV PW, respectively. The sameness in these

seven alignments suggests a long-term structural stability of the seven taxonomies from 1880 to 1940. In other words, these seven taxonomies may consist of similar parent and children nodes structures, with the semantics (or the concept, geographic names) of the nodes as the only change throughout time.

From Alignment 7 to 8, we see a sudden decrease in the number of total possible worlds. In particular, all four types of PWs drop by one in the number of worlds from 1940 to 1950. This watershed moment from long-term structural stability to a decrease to six PWs is likely caused by a number of historical sovereignties changes after the World War II. The Empire of Japan no longer exist in 1950, but a couple new sovereignties started their rulings including the People’s Republic of China, the Republic of China, the Democratic People’s Republic of Korea, and the Republic of Korea. Ruling the North, Central, and South Vietnam, the sovereignties Tokin Protectorate, Annam Protectorate, and Cochinchina in 1940 became the State of Vietnam and the Democratic Republic of Vietnam in 1950. These newly-formed sovereignties in 1950 resemble the sovereignties of the current year (2022) more than the 1940’s sovereignties do, resulting in a decrease in the number of PWs .

In general, the alignments yield possible worlds in a downward trend – the closer it is to the current world, the less possible worlds and ambiguities there are. Alignment 11-13 only have four PWs, suggesting that the world from 1980, 1990, 2000 very much resemble that of 2022, even more so than the year 1950, 1960, and 1970.

***Within Alignments.*** To examine within each alignment, for example, for Type I PW, where the root node “World” is congruent in both taxonomies, there are seven alignments that resulted in two possible worlds that are of this type (see Table 8.7). The main difference between the two possible worlds within the alignment is the entities relating to the geographic entity “Vietnam”. The possible worlds showcase that it is difficult to discern whether the geographic entity of modern day Vietnam is entirely a part of the political sovereignty from the past, or whether the two are partially overlapping. For Type II, III, IV PW, if an alignment has more than one possible world, the ambiguity also lies in the entities relating

to Vietnam, generating two to three possible worlds.

Table 8.7: 13 Alignments and the corresponding counts of possible worlds in each type

| Alignment | $T_C/T_P$ | Type I PW<br>( $T_C == T_P$ ) | Type II PW<br>( $T_C > T_P$ ) | Type III PW<br>( $T_C > T_P$ ) | Type IV PW<br>( $T_C < T_P$ ) | Total PW |
|-----------|-----------|-------------------------------|-------------------------------|--------------------------------|-------------------------------|----------|
| 1         | 2022/1880 | 2                             | 3                             | 3                              | 2                             | 10       |
| 2         | 2022/1890 | 2                             | 3                             | 3                              | 2                             | 10       |
| 3         | 2022/1900 | 2                             | 3                             | 3                              | 2                             | 10       |
| 4         | 2022/1910 | 2                             | 3                             | 3                              | 2                             | 10       |
| 5         | 2022/1920 | 2                             | 3                             | 3                              | 2                             | 10       |
| 6         | 2022/1930 | 2                             | 3                             | 3                              | 2                             | 10       |
| 7         | 2022/1940 | 2                             | 3                             | 3                              | 2                             | 10       |
| 8         | 2022/1950 | 1                             | 2                             | 2                              | 1                             | 6        |
| 9         | 2022/1960 | 1                             | 2                             | 2                              | 1                             | 6        |
| 10        | 2022/1970 | 1                             | 2                             | 2                              | 1                             | 6        |
| 11        | 2022/1980 | 1                             | 1                             | 1                              | 1                             | 4        |
| 12        | 2022/1990 | 1                             | 1                             | 1                              | 1                             | 4        |
| 13        | 2022/2000 | 1                             | 1                             | 1                              | 1                             | 4        |

$T_C$ : Taxonomy of the most current year;  $T_P$ : Taxonomy of a past year

#### 8.4.5 Utilizing Possible Worlds

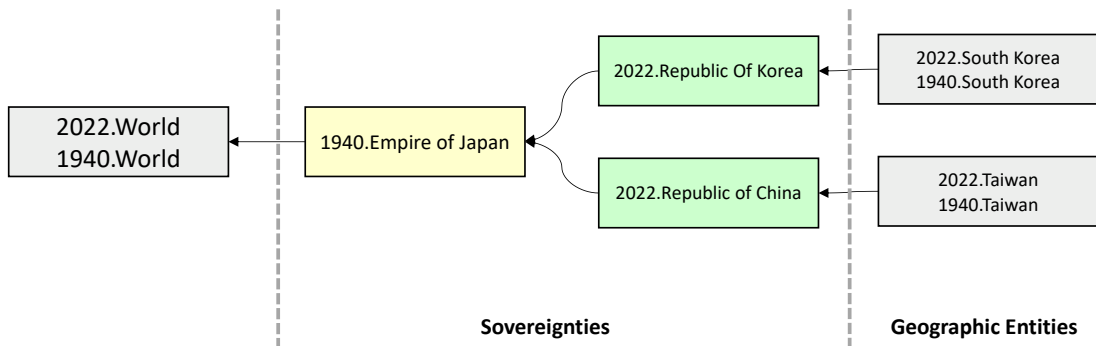


Figure 8.13: Illustration of a simplified example possible world

Geographic information, seemingly objective, is ambiguous and subjective when it comes to different perspectives and time periods. The proposed conceptual model has made the

ambiguous historical sovereignties salient through the process of aligning different data-centric taxonomies (both biodiversity data and sovereignty data). It is important to analyze across and within alignment results to make sense of the possible worlds, but what is more important is to utilize the possible worlds to enhance the original biodiversity data.

Figure 8.13 is an illustration of a possible world that may likely result from the alignment, with sovereignties of the past in yellow (1940.Empire of Japan), sovereignties of the current year in green (2022.Republic of Korea, 2022. Republic of China), and congruent geographic entities from both the past and the current year in grey (South Korea, Taiwan). Having such graphs from the alignment, we are able to infer that in what is now called as “Republic of Korea” is actually “Empire of Japan” in 1940. To enhance the species occurrence data based on this fact, if there is a occurrence record from the 1940 in “South Korea” or the “Republic of Korea”, we should provide an value-added information in a separate column denoting that the historical sovereignty is “Empire of Japan”.

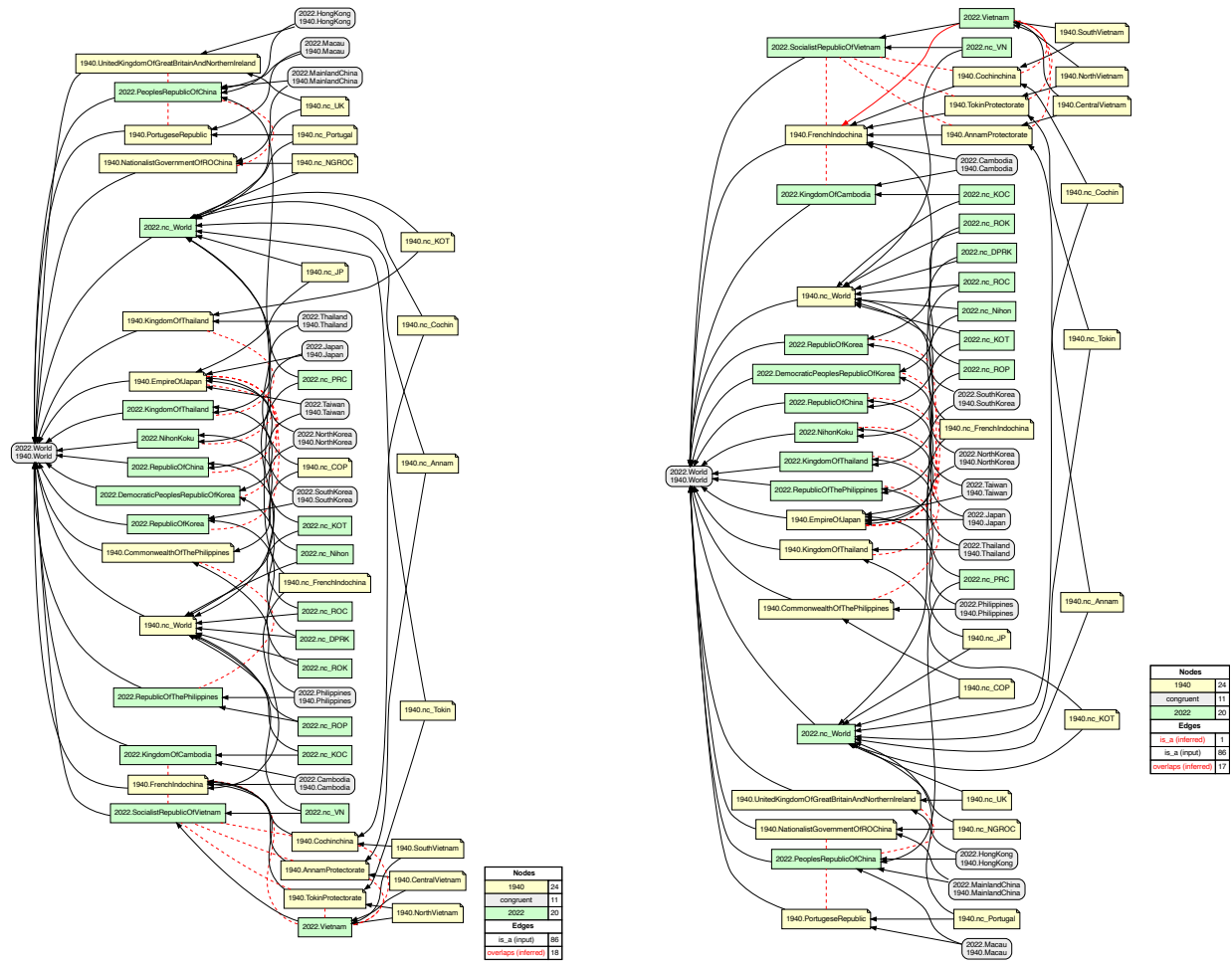


Figure 8.14: Example of two Type I possible worlds for  $T_{2022}$  and  $T_{1940}$

This also brings out the value in the proposed conceptual model with taxonomy alignment process – providing a hierarchical structure with both geographic entities covered and sovereignties in one view. Instances when there are multiple possible worlds, such as in Figure 8.14, also demonstrate the value of having diverse perspectives that users are able to peruse and select from. After all, when it comes to historical information, there is no one ground truth.

## 8.5 Modeling Challenges

While our proposed model with aligning taxonomies aims at providing multiple perspectives through the merged possible world approach, we are cognizant that the specific modeling challenges we faced may hinder the representation of other perspectives in this study. Here we highlight the challenges we encountered in this research.

### 8.5.1 Reconciling Geographic Ambiguity

It is challenging to reconcile geographic information when the reconciliation process is contingent upon unsettled sovereignty data. In our study, we view the sovereignty in East Asia as of now as unambiguous, but there are many contested regions and maritime zones in this modern day. We highlighted one of these cases in our prior work with Taiwan and China relations (Cheng & Ludäscher, 2019b). Other geopolitical or civil unrest also exists, such as the Hong Kong protest (Cheung & Hughes, 2019), or the Diaoyutai (Senkaku) islands sovereignty dispute (Shaw, 1999). If we were to align the ambiguity of *the now* with the ambiguity of *the past*, it would be nearly impossible to model the situations, given it would produce an exponential amount of pairwise alignments. In our attempts, we realize that the best practice to model critical cartography, thereof, is to keep one variable constant – to keep the *now* definite.

Further, it is also difficult to determine the geographic entities’ boundaries and the historical sovereignties’ reign. This challenge arises mainly because there is no systematic, unbiased production of maps throughout time that depicts every geographic entity there ever was, has been, or still is. There is also no telling from historical and current maps what all the areas a historical sovereignty covers. Echoing our earlier point in the Related Work section that maps produced by different agencies or institutions may have their own motives to promote elite groups and silence the marginalized (Crampton, 2001; Harley, 1988, 1992). In our experience, we could only construct the taxonomy to approximate where a

sovereignty might cover to the best of our knowledge from the data sources, rather than account for every little islands it might actually own.

### 8.5.2 Modeling Time

Modeling time, or periods of time, is another challenge we encountered in this research. Resonating the recommendations posed by Tennis and colleagues (Tennis, 2010, 2013; Tennis et al., 2012), modeling the units and changes of time and the decision on which units to use could have future repercussions. In our case to model historical sovereignties, the challenge is mainly due to a lack of an unequivocal, longitudinal sovereignty data that encompass every geographic entity of the world on the same *timescale*. We attempt to model time in a long term, 10-year interval, that is, to slice sovereignty changes for all the geographic entities involved by a decade, much like how Tennis work with the case of Eugenics in bibliographic classifications (Tennis, 2012).

Using time slicing to approximate time periods for change is certainly useful to put every entry under the same scale, but continuous as time is, we find that sometimes two sovereignties may reign in the same geographic entities in overlapping years. Such case is shown in the Wikipedia’s list of sovereignties <sup>20</sup> for the geographic entity Cambodia, when sovereignties overlap in the time periods of their reign: Democratic Kampuchea (1975–1982), People’s Republic of Kampuchea (1979–1989), Coalition Government of Democratic Kampuchea (1982–1990), and State of Cambodia (1989–1993). In our study, we model such case to prefer the sovereignty of the successors who reign longer following that year. For this case, for the year 1980, we would model it as People’s Republic of Kampuchea rather than the Democratic Kampuchea. For cases that may have almost equal overlapping years in their reign, or unclear geographic coverage of their reign, such as Democratic Republic of Vietnam and Republic of Vietnam, we keep both entities in our model. The best practice

---

<sup>20</sup>Wikipedia sovereignty list: [https://en.wikipedia.org/wiki/List\\_of\\_predecessors\\_of\\_sovereign\\_states\\_in\\_Asia](https://en.wikipedia.org/wiki/List_of_predecessors_of_sovereign_states_in_Asia)

will be to include historians or East Asia experts who are familiar with East Asia history and geographic changes, to provide a best optimal modeling situation.

### 8.5.3 The “Residual”, or Other Category

In constructing the taxonomies, we implemented a placeholder node for the “other” categories to express the otherwise unknown geographic entities that a sovereignty covers. We acknowledge three of the eight aspects of “residual categorizing” mentioned by Star and Bowker (2007) which we conceive as modeling challenges for our study. These three contexts when we place a “other” node are when:

- The history of geographic entities is ambiguous, and the boundary of what their sovereignties cover are sometimes “unknown” or “unknowable” (e.g., Did the Empire of Japan cover more geographic entities other than Japan, Taiwan, North Korea, South Korea in 1940?);
- Some sovereignties “embodies two or more of the categories” that transcends the geographic region of interest in this study (e.g., Portuguese Republic concerns Macau in the East Asia region; other geographic entities that Portuguese Republic covers in Europe are placed under the other category);
- Geographic entities are border-less across time; if we attempt to model all geographic entities cover by all sovereignties across time, we will have to model all the world’s entities – causing “strains on the technical capability”, or computational power of the alignment process.

We understand that by modeling residual categories, it might hinder the accessibility of certain entities (Bowker & Star, 2000; Star & Bowker, 2007). We maintain that the modeling of these so-called residual categories in the context of historical sovereignties in this research

is not to impose invisibility, or silence the multiplicity of any entities, but constraints that warrants further investigations.

#### 8.5.4 What's in a name?

As mentioned in Bishop et al. (2015), geographic names are politically charged at times. This presents a challenge to modeling geographic-related taxonomies, for it is difficult to obtain geographic information that has a neutral stance and free of any perspectives. For instance, in maritime zone, is it called the Taiwan Strait (Taiwanese perspective), or the Chinese Formosan Strait (Chinese perspective)? Sometimes, the naming is not only political, but also relative to the geographic directions. For example, South China Sea is called East Sea from the Vietnamese perspective.

In addition, while Buckland and Lancaster believe that measures such as latitude and longitude are the *lingua franca* for geographic information Buckland and Lancaster (2004), *lingua franca* for geographic names varies upon sovereignties. In an ideal modeling situation, the geographic entities should be presented in whichever languages their *lingua franca* were during the time of a certain sovereignty reign. In reality, it is difficult to incorporate multilingual content in a taxonomy without raising confusion, not to mention the extra layer of difficulty to trace back in history on how one sovereignty calls its sovereign states. The usual resolution is to model taxonomies in English, such as this study, with the risk of spelling differences (e.g., is it *Macau* or *Macao*?). Our stance in presenting the taxonomies in English is not to exclude any other languages or perspectives, but rather a limitation on “naming entities in interdisciplinary (in our case, international) communication” as Bowker addressed (Bowker, 2000b). After all, not all species occurrence record or legacy literature has latitude and longitude information, and would rely on historical, often English, geographic names.

## 8.6 Chapter Summary

In this work, we propose a conceptual model for historical sovereignties in biodiversity data. The conceptual model consists of four phases: (1) Collecting and preprocessing of species and historical sovereignties data to determine the geographic point of interests; (2) Constructing taxonomies of the geographic entities and sovereignties based on certain time slices; (3) Aligning taxonomies in a pairwise manner to compare each past year taxonomy with the most current year taxonomy; (4) Producing possible worlds based on the alignments to demonstrate in a given past year, what was the historical sovereignty of a modern geographic entity. This chapter also address how multiple possible worlds can be compared by categorizing them into four types.

The practical contribution for a conceptual model of sovereignties is to provide additional information in biodiversity data relating to location. In our observation, early year occurrence data usually lacks absolute location, that is, latitude and longitude information. These occurrences are then highly reliant on *Country*, *Country Code*, or *Locality* information in a Darwin Core formatted dataset, which do not reflect its historical geographic information. The Darwin Core metadata provides substantial elements for “location” information<sup>21</sup>, including *higherGeography*, *continent*, *waterBody*, *islandGroup*, etc, but the elements are again, referring to geography of current times. Plus, Darwin Core suggests the use of Getty Thesaurus of Geographic Names as controlled vocabulary, which can also be subjected to ambiguity. Additionally, occurrence datasets from different institutions may implement Simple Darwin Core<sup>22</sup>, which gives them flexibility to choose which metadata elements and attributes that most fit with their collections, and not all elements in the “location” section will be implemented. Were additional elements to be implemented in these biodiversity data, we advocate that the findings from our possible worlds could contribute to the intricacies of the constant-evolving geographic information.

---

<sup>21</sup>Darwin Core location: <https://dwc.tdwg.org/terms/#location>

<sup>22</sup>Simple Darwin Core: <https://dwc.tdwg.org/simple/>

The theoretical contribution of our work is to raise awareness that implicit assumptions on geographic information exist in biodiversity data. Much like Bowker (Bowker, 2000b) had mentioned, the diversity of biodiversity data makes it challenging for long-term curation, we should devise multiple representational forms that could empower the “multiple data diversities of the field of biodiversity”. By discussing critical cartography problems in biodiversity data, we acknowledge that there may be inherent biases that are previously overlooked; and by providing a taxonomic view and a resulting world view in multiple possible solutions, we believe multiple perspectives could be supported.

The comparison among possible worlds in this chapter is rather topic-dependent that may only be salient towards modeling historical sovereignties. In future work, we anticipate verifying whether the comparisons of possible worlds used in this conceptual model can be generalized towards other socio-geographic cases.

## CHAPTER 9

## CONCLUSION

This dissertation applied a logic-based alignment approach to different contexts, namely, contexts that involves geographic information and social perceptions. I call these socio-geographic contexts. To reiterate, the use cases in this dissertation are those that contain more than two taxonomies about the same subject matter in need of reconciliation.

In this chapter, I first discuss the findings and research questions of this dissertation based on the previous four chapters on use cases of taxonomy alignment problems. I then discuss the methodological and empirical contributions of this dissertation. Finally, I conclude this dissertation with proposed directions for future work.

### 9.1 Dissertation Findings

#### 9.1.1 Research Question 1

In Research Question 1, I ask that given a pair of taxonomies, how do we align them so that the information from both taxonomies is preserved?

Throughout Chapters 5, 6, 7, 8, I have shown that employing the logic-based taxonomy alignments in RCC-5 can align distinct taxonomies to resolve disagreements, while *preserving the internal coherence and context of each taxonomy*. Here I highlight the findings from two of the chapters to address this question.

In Chapter 5, I align two input taxonomies that classify U.S. regions: the taxonomy of the Census Bureau regions, and the taxonomy of the National Diversity Council regions. The

process involves relating 49 leaf-level (lower 48 states and D.C.) RCC-5 articulations that map the state in one taxonomy congruent with the state-level nodes in another taxonomy. The output is a single possible world with all state-level nodes congruent and new articulations for region-level alignments on the upper-level classifications. I then align another input taxonomy, the U.S. time zone map, with the Census Bureau taxonomy. This time on the region-level, with 12 region-level RCC-5 articulations (none of them equivalent) that show how each higher-level input region can be spatially related with its respective counterparts. The resulting merged solution is also a single merged solution.

In Chapter 7, five input taxonomies about the indigenous tribes in Taiwan are provided. These include the electoral constituency voting group, the past, current, and possibly future number of recognized tribes, and a re-scaled grouping based on population. The alignment involves four pairs of taxonomies, each taxonomy is aligned with the bipartite voting group. All alignments in this chapter yield a single merged solution. However, to further examine which regions are overlapping with the other, I extrapolate the overlaps relations by using a “combined concept” view of the merged solutions (or possible worlds). That is, a more granular, “zoom-in view” that demonstrates what nodes are in between “the middle layer” of the two taxonomies.

In the merged taxonomies in both chapters, though new nodes or new inferred relations among the nodes are formed, the original information about the input taxonomies is still present. This “preserved information” includes the names of the concepts in either taxonomies and the is-a relations within each taxonomy. While in other alignment approaches, new names would be generated in a merged solution and the original information would be overwritten and lost.

Having information from both taxonomies in a merged solution allows for a more transparent and reproducible pipeline in taxonomy alignment work. For instance, in Chapter 5, from the merged solution, we can trace back to what  $T_{CEN}$  and  $T_{NDC}$  originally were. We can also come up with new taxonomy alignment problems (e.g., a new case aligning the

regions of  $T_{CEN}$  and  $T_{NDC}$ ) based on the merged solution of  $T_{CEN}$  and  $T_{NDC}$ , and can be able to distinguish which nodes should be in the original  $T_{CEN}$ , which ones should be in the original  $T_{NDC}$ .

Furthermore, the original information in the input taxonomies serves as a gateway to retain the identity of each taxonomy. For example, for the cases discussed in Chapter 7, a clear tribal name distinction is still present in the merged solution along with the coarse geo-based voting groups – this merged view with all the information preserved permits a more diverse perspective in reconciling different taxonomies. In contrast, if one simply uses a geographic information system (GIS) to overlay maps of different taxonomies, it will be difficult to recognize from the layered maps whether regions are from the original taxonomies, or if they are newly-formed regions from intersecting two (or more) taxonomies.

### 9.1.2 Research Question 2

In Research Question 2, I ask that given a pair of taxonomies, how, if at all, do we get multiple merged solutions when we align the taxonomies? This question is addressed particularly in Chapters 6, and 8, when the taxonomy alignment problems yield multiple merged solutions.

Chapter 6 is a first attempt at looking into the disputed geographic regions and sovereignties in the world, focusing on the case of Taiwan. I align three input taxonomies of the world’s countries (or geographic entities): the ISO country code, the U.S. Department of Homeland Security country taxonomy (DHS), and the CIA World Fact Book. I focus specifically on the Taiwan and China relation in this chapter, so the input taxonomies are “zooming in” on East Asian entities of the world. The three taxonomies are aligned in a pairwise fashion throughout the three use cases:  $T_{ISO}$  and  $T_{DHS}$ ,  $T_{ISO}$  and  $T_{CIA}$ ,  $T_{CIA}$  and  $T_{DHS}$ . In the third alignment,  $T_{CIA}$  and  $T_{DHS}$ , the two taxonomies have more similar taxonomic structures (both with two levels of nodes, and the same number of nodes in each taxonomy). One congruent node is made at the leaf-level, and one specifying that Asia is congruent in both taxonomies.

This particular case yield *seven* possible worlds, and I compare these seven merged solutions in terms of how they reflect the real-world social perspectives. The seven possible worlds interestingly mirror different viewpoints around China-Taiwan relations, and could possibly be interpreted as whether a world is more similar to the Chinese government’s perspective, the Taiwanese government’s perspective, the neighboring countries’ perspectives, or political extremists’ perspectives, among several others.

Chapter 8 moves beyond looking solely at geographic entities and sovereignties by incorporating the ideas of time and change in datasets. I discuss socio-geographic concerns in biodiversity data that previously may have been viewed as objective or have been overlooked as a potential data quality problem. To untangle the many layers of the geographic information within biodiversity data, I first propose a conceptual model for specifically looking at the *historical sovereignties* that govern the species over time. The conceptual model involves the construction of taxonomies from the data, and the alignment to compare each past year taxonomy with the most current year taxonomy.

To demonstrate the utility of the conceptual model, I discuss a specific use case with the Black-faced Spoonbills and construct 14 input taxonomies that include the current sovereignties in East Asia, and each decade’s taxonomy since 1880. In total, there are 13 alignments in this chapter, and each alignment has about 30 to 35 articulations with the RCC-5 relations (equivalent, proper-part, inverse proper-part, overlap, disjoint). Each alignment yields four to ten possible worlds with many inferred relations. To compare different merged solutions (possible worlds) in modeling historical sovereignties in datasets, I organize the worlds into four types ( $T_C == T_P$ ;  $T_C >< T_P$ ;  $T_C > T_P$ ;  $T_C < T_P$ ), mainly whether a past year’s world is congruent with, overlapping, subsumed in, or subsuming the current year’s world.

The use cases have shown how and why multiple solutions are provided and needed in a taxonomy alignment problem. As to further exploration on what these possible worlds might mean, one of our prior studies (Gupta, Cheng, & Ludäscher, 2019) attempts to quantitatively compare the merged solutions using the *Possible World Explorer* tool by querying a series of

distance similarity measures. The comparison could be helpful to explore systematically the node and edge structure of a given possible world, and rank the worlds based on user-defined measures (e.g., the number of congruent nodes). However, given there are no apparent correlations among the characteristics (nodes and edges) of the input taxonomies, number and types of articulations, and the number of output (possible worlds), a quantitatively comparison could only do so much to rank plausible solutions.

A combination of variables that are *case-specific* and *human-decision-dependent* may contribute to the complexity of a taxonomy alignment problem and result in many possible worlds. For instance, in the alignments in Chapter 8, the root nodes are intentionally not mapped because the historical context is unknown. We cannot say for sure that the boundary of East Asia back then is exactly the same as the East Asia now. This mapping decision due to *historical ambiguity* can be one of the causes that lead to the multiple possible worlds in all the cases in that chapter. For the  $T_{CIA}$  and  $T_{DHS}$  case, other factor may interplay with the alignment, such as *the number of unmapped leaf nodes*. Almost all use cases in this dissertation align all leaf nodes with its counterpart in the other taxonomy. In the  $T_{CIA}$  and  $T_{DHS}$ , the unmapped leaf nodes is a proclamation of “not making any assumptions” between the two geographic entities discussed. Regardless of how we may rank the possible worlds based on some quantitative measures, the four to ten possible worlds generated in these two cases can all be plausible and optimal merged solutions – depending on whose perspective it is or depending on experts’ opinions on the subject matter at hand.

## 9.2 Contributions

Through aligning taxonomies in a logic-based approach, this dissertation achieves the goals proposed in the beginning chapter: to preserve the information in both taxonomies in the merged solution(s); to provide multiple merged solutions when the input taxonomy is ambiguous.

### 9.2.1 Methodological Contributions

The methodological contributions of this dissertation lie in the uniqueness of the RCC-5 logic-based alignment approach. First, its capability to provide multiple solutions is a gateway to other systems of thoughts. Echoing back to the quote by Foucault in Chapter 1 on the classification of animals, even though the classification in the passage seemed too exotic to be real, it is still “another system of thought”. Much like Foucault’s animal classification, there are many taxonomies in this world, and no one is more special or more correct than the other. By preserving information from the original taxonomies, and by providing multiple merged solutions, we advocate that each taxonomy within a possible world is equally important.

Relating to its ability to provide multiple perspectives in one merged view, second, the logic-based taxonomy alignment approach serves as a mean to make hidden structures visible. An objective dataset, a frequently-used bibliographic classification, or a drop-down menu list of countries can hold inherent assumptions that go unnoticed – and intentionally or unintentionally suppress other perspectives. Raising awareness that such embedded structure exists is a good deed; presenting possibilities of how these perspectives can simultaneously and consistently co-exist is that olive branch that helps reaching a consensus.

Third, this dissertation demonstrates that taxonomy alignment is an ongoing, iterative process. I maintain that Taxonomy Alignment should not be a one-stop solution, but rather, an alignment life cycle that involves re-designing and revising of the input taxonomies and the mapping relations to reach logically consistent and coherent merged solution(s). A merged solution from a taxonomy alignment problem can then in return become another alignment problem’s input taxonomies, starting another iterative life cycle.

Finally, the artifacts this dissertation provides in Chapters 3 and 4, including a thorough review of existing taxonomy alignment methods, and a representation of the RCC-5 composition table in Euler diagrams, are original contributions to the taxonomy alignment approach literature at large.

### 9.2.2 Empirical Contributions

The alignment mechanism in the logic-based approach used in this dissertation, Region Connection Calculus (RCC-5), is a qualitative *spatial* reasoning measure suitable for comparing spatial or geographical artifacts. Motivated by this initial intention of RCC-5, this dissertation empirically applies our RCC-5 based alignment approach to socio-*geographic* contexts. Previously only applied to biodiversity contexts, the use cases in this dissertation are the first of this RCC-5 logic-based approach applied to *geographic* contexts. Other existing taxonomy alignment approaches draw inspirations from domains such as e-commerce, biology, or linguistics. By applying the logic-based approach to multiple taxonomy alignment problems in socio-geographic contexts that are closely related to everyone’s lives, we may be able to better articulate the needs of reconciliation to the general public.

The practical implications of this dissertation are three-fold, relating to the utility of the merged solutions of a taxonomy alignment problem. First, the multiple merged solution(s) from the RCC-5 logic-based approach can serve as minimum viable products (MVPs) towards reaching a consensus on certain matters. In a field-specific taxonomy (e.g., biodiversity), the possible world(s) can be a first artifact used to communicate with experts, and to further involve the experts in the iterative design and alignment process.

Second, for some topics (e.g., Indigenous peoples’ tribes and voting groups in Chapter 7), the merged solutions can serve as indicators for decision-making. The possible worlds not only show the two sides of a coin (i.e., the two perspectives of the two taxonomies), but they also show *the metal core* that the two sides are engraved on (i.e., the relation between the two sides). The middle-layer nodes that both sides can partially agree upon is that metal core; and this common ground that both sides can agree with may be further leveraged to reach consensus (e.g., assigning seats to the voting groups based on this middle-layer).

Third, the possible worlds can be used as a guideline for long-term data curation. It can be especially helpful in dataset merging, data migration, and data versioning control – in a

way that the possible worlds showcase possible curation scenarios that a researcher (or data curator) can benefit from.

### 9.3 Future Work

This dissertation lays a foundation of understanding the RCC-5 logic-based approach in aligning taxonomies. The findings from each chapter set out several new methodological and empirical directions in which my future work will be based on.

In terms of the method of the RCC-5 taxonomy alignment approach, there is room for additional analysis to evaluate the different matching dimensions (Chapter 3.6.6). A possible future study is to conduct an experimental procedure with one use case and control for the number and types of input taxonomies and articulation. This may enable us to further observe how different variables in a taxonomy alignment problem affect the number of possible worlds produced.

Another direction relating to the method is to explore the multi-way taxonomy alignment. The RCC-5 taxonomy alignment approach can be used to align and merge more than two taxonomies at a time. In this dissertation, the use cases are only pairwise taxonomy alignments due to computational resource limitations. When computation resources and scalability issues are resolved, it will be interesting to further explore multi-way alignment problems.

As for empirical directions, this dissertation has shown the feasibility of the RCC-5 logic-based approach in socio-geographic contexts. I plan to further extend the application of this approach to other areas, specifically, explore whether the approach can be applicable to other knowledge organization systems such as metadata standards or bibliographic classification schemes.

The findings and contributions of this dissertation suggest that the merged solutions of a taxonomy alignment problem could be useful for decision making, or for creating long-term

data curation goals. I plan to seek other extended uses of the merged solutions, including how these possible worlds might be turned into knowledge graphs, and how experts might be involved in specifying input articulations.

Finally, I have demonstrated how taxonomy alignment can uncover socio-geographic concerns in biodiversity data in this dissertation. In future work, I envision a *multi-layered* taxonomy alignment problem in the field of biodiversity: we can investigate different kinds of taxonomies (e.g., geographic taxonomies, species taxonomies, and other taxonomies) within biodiversity data, literature, and information systems; and we can *agree to disagree* on how to reconcile these taxonomies.

## REFERENCES

- Aanen, S. S., Nederstigt, L. J., Vandić, D., & Fräsincar, F. (2012). Schema-an algorithm for automated product taxonomy mapping in e-commerce. In *Extended semantic web conference* (pp. 300–314).
- Adler, M. (2016). The case for taxonomic reparations. *Knowledge Organization*, 43(8), 630–640.
- Adler, M., & Tennis, J. (2013). Toward a taxonomy of harm in knowledge organization systems. *Knowledge Organization*, 40(4).
- Adler, M., Tennis, J., Milojević, S., van Hooland, S., Rogers, C., & West, J. D. (2013). The temporal dimension in the study of knowledge bases: approaches to understanding knowledge creation and representation over time. *Proceedings of the American Society for Information Science and Technology*, 50(1), 1–3.
- 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.
- Bachmair, L., Ganzinger, H., & Waldmann, U. (1993). Set constraints are the monadic class. In *[1993] proceedings eighth annual ieee symposium on logic in computer science* (pp. 75–83).
- BBC. (2021). *What’s behind the China-Taiwan divide?* <https://www.bbc.com/news/world-asia-34729538>. BBC News.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, 284(5), 34–43.
- Biermann, C., & Mansfield, B. (2014). Biodiversity, purity, and death: conservation biology as biopolitics. *Environment and Planning D: Society and Space*, 32(2), 257–273.
- Biho, M. (n.d.). What is plain indigenous people? *mayawbiho*. <http://www.mayawbiho.tw/plain>.
- Bird, K. (2014). Ethnic quotas and ethnic representation worldwide. *International Political Science Review*, 35(1), 12–26.
- Bishop, W., & Grubestic, T. H. (2016a). Geographic Information, Maps, and GIS. In W. Bishop & T. H. Grubestic (Eds.), *Geographic Information: Organization, Access, and Use* (pp. 11–25). Cham: Springer International Publishing. doi: [https://doi.org/10.1007/978-3-319-22789-4\\_2](https://doi.org/10.1007/978-3-319-22789-4_2)
- Bishop, W., & Grubestic, T. H. (2016b). Metadata. In W. Bishop & T. H. Grubestic (Eds.), *Geographic Information: Organization, Access, and Use* (pp. 79–103). Cham: Springer International Publishing. doi: [10.1007/978-3-319-22789-4\\_5](https://doi.org/10.1007/978-3-319-22789-4_5)

- Bishop, W., Moulaison, H. L., & Burwell, C. L. (2015, August). Geographic Knowledge Organization: Critical Cartographic Cataloging and Place-Names in the Geoweb. *Knowledge Organization*, 42(4), 199–210. doi: 10.5771/0943-7444-2015-4-199
- Blust, R., et al. (2013). *The Austronesian languages*. Asia-Pacific Linguistics, School of Culture, History and Language, College . . .
- Boakes, E. H., McGowan, P. J., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol*, 8(6), e1000385.
- Bodirsky, M., & Hils, M. (2012). Tractable set constraints. *Journal of Artificial Intelligence Research*, 45, 731–759.
- Bowker, G. (2000a). Biodiversity datadiversity. *Social studies of science*, 30(5), 643–683.
- Bowker, G. (2000b, December). Mapping biodiversity. *International Journal of Geographical Information Science*, 14(8), 739–754. (Publisher: Taylor & Francis \_eprint: <https://doi.org/10.1080/136588100750022769>) doi: 10.1080/136588100750022769
- Bowker, G., & Star, S. L. (2000). *Sorting things out: Classification and its consequences*. MIT press.
- Bowman, E. (2021). *She struggled to reclaim her indigenous name. she hopes others have it easier*. <https://www.npr.org/2021/07/04/1007737459/canada-indigenous-names-reconciliation-schools>. NPR.
- Brachman, R. J. (1983). What is-a is and isn't: An analysis of taxonomic links in semantic networks. *Computer*(10), 30–36.
- Brown, D. (2022). *China and taiwan: A really simple guide*. <https://www.bbc.com/news/world-asia-china-59900139>. BBC News.
- Buckland, M., & Lancaster, L. (2004). Combining Place, Time, and Topic: The Electronic Cultural Atlas Initiative. *D-Lib Magazine*, 10(5). Retrieved from <http://mirror.dlib.org/dlib/may04/buckland/05buckland.html>
- Buckley, C., & Horton, C. (2019). *Xi jingping warns taiwan that unification is the goal and force is an option*. <https://www.nytimes.com/2019/01/01/world/asia/xi-jinping-taiwan-china.html>. New York Times.
- Callahan, W. A. (2009). The cartography of national humiliation and the emergence of China's geobody. *Public Culture*, 21(1), 141–173.
- 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. M., & Zeng, M. L. (2006). Metadata interoperability and standardization—a study of methodology part i. *D-Lib magazine*, 12(6), 1082–9873.
- Charney, J. I., & Prescott, J. R. (2000). Resolving cross-strait relations between China and Taiwan. *American Journal of International Law*, 94(3), 453–477.
- Chen, M. (2014). *Reasoning about queries and constraints* (Doctoral Dissertation). University of California Davis.
- 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*.
- Chen, M., Yu, S., Franz, N., Bowers, S., et al. (2014). Euler/X: a toolkit for logic-based taxonomy integration. *arXiv preprint arXiv:1402.1992*.

- Chen, M., Yu, S., Kianmajd, P., Franz, N., Bowers, S., & Ludäscher, B. (2014). Provenance for explaining taxonomy alignments. In *International provenance and annotation workshop* (pp. 258–260).
- Cheng, Y.-Y. (2021). Systematic comparison of data models used in mapping knowledge organization systems. *Proceedings of the Association for Information Science and Technology*, 58(1), 685–687.
- Cheng, Y.-Y. (2022). *Data artifacts for reconciling the socio-geographic in biodiversity data and information systems*. doi: <https://doi.org/10.5281/zenodo.6632129>
- Cheng, Y.-Y., Dilliplane, S., & Ludäscher, B. (2020). Snailed it! Merging taxonomically organized biodiversity datasets with shifting geopolitical realities. *IDCC*.
- Cheng, Y.-Y., Franz, N., Schneider, J., Yu, S., Rodenhausen, T., & Ludäscher, B. (2017a). 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., Franz, N., Schneider, J., Yu, S., Rodenhausen, T., & Ludäscher, B. (2017b). *Data artifacts for reconciling the socio-geographic in the united states maps*. doi: <https://doi.org/10.5281/zenodo.6625833>
- Cheng, Y.-Y., Hoang, L., & Ludäscher, B. (2020). Cacao, Cocoa, or Cocoa? Reconciliation of taxonomic names in Biodiversity Heritage Library. *ISKO*.
- Cheng, Y.-Y., & Ludäscher, B. (2018). Full of beans: a study on the alignment of two flowering plants classification systems. *Proceedings of the 18th European Networked Knowledge Organization Systems Workshop (NKOS)*.
- Cheng, Y.-Y., & Ludäscher, B. (2019a). *Data artifacts for reconciling the socio-geographic in country taxonomies*. doi: <https://doi.org/10.5281/zenodo.6625835>
- Cheng, Y.-Y., & Ludäscher, B. (2019b). Exploring geopolitical realities through taxonomies: The case of Taiwan. *NASKO*, 7(1), 77–93.
- Cheng, Y.-Y., & Ludäscher, B. (2020a). *Data artifacts for reconciling the socio-geographic in indigenous peoples' tribes*. doi: <https://doi.org/10.5281/zenodo.6625839>
- Cheng, Y.-Y., & Ludäscher, B. (2020b). Reconciling taxonomies of electoral constituencies and recognized tribes of indigenous Taiwan. *Proceedings of the Association for Information Science and Technology*, 57(1), e248.
- Cheung, H., & Hughes, R. (2019). Why are there protests in Hong Kong? all the context you need. *BBC News*, 4.
- Chiu, S.-h., & Chiang, W.-y. (2012). Representations of the name rectification movement of Taiwan indigenous people: Through whose historical lens? *Language and Linguistics*, 13(3), 523.
- Clarke, R. I., & Schoonmaker, S. (2019). Metadata for diversity: Identification and implications of potential access points for diverse library resources. *Journal of Documentation*.
- Cohn, A. G., & Renz, J. (2008). Qualitative spatial representation and reasoning. *Foundations of Artificial Intelligence*, 3, 551–596.
- Coombes, B., Gombay, N., Johnson, J., & Shaw, W. (2011). A companion to social geography. In V. J. Del Casino Jr, M. Thomas, P. Cloke, & R. Panelli (Eds.), (chap. The Challenges of and from Indigenous Geographies). John Wiley & Sons.
- Cragin, M. H. (2003). Toward integrative science: Organizing biodiversity and neuroscience

- data. *Bulletin of the American Society for Information Science and Technology*, 30(1), 14–17.
- Crampton, J. W. (2001). Maps as social constructions: power, communication and visualization. *Progress in human Geography*, 25(2), 235–252.
- 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. *Knowledge Organization*, 43(2), 107–112.
- Del Casino Jr, V. J. (2009). *Social geography: a critical introduction*. John Wiley & Sons.
- Del Casino Jr, V. J., Thomas, M., Cloke, P., & Panelli, R. (2011). *A companion to social geography*. John Wiley & Sons.
- Denecker, M., Lierler, Y., Truszczyński, M., & Vennekens, J. (2012). A Tarskian informal semantics for answer set programming. In *Technical communications of the 28th international conference on logic programming* (Vol. 17, pp. 277–289).
- Dou, D., McDermott, D., & Qi, P. (2005). Ontology translation on the semantic web. In *Journal on data semantics ii* (pp. 35–57). Springer.
- Doyle, A. M. (2013). *Naming, claiming and (re) creating: Indigenous knowledge organization at the cultural interface* (Doctoral Dissertation). University of British Columbia.
- Duarte, M. E., & Belarde-Lewis, M. (2015). Imagining: Creating spaces for indigenous ontologies. *Cataloging & Classification Quarterly*, 53(5-6), 677–702.
- Enos, R. D. (2017). *The space between us: Social geography and politics*. Cambridge University Press.
- 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.
- Ericsson, N. S. (2004). Creating “Indian country” in Taiwan? *Harvard Asia Quarterly*, 8, 33–44.
- Euzenat, J., Shvaiko, P., et al. (2007). *Ontology matching* (Vol. 18). Springer.
- Ewers-Saucedo, C., Allspach, A., Barilaro, C., Bick, A., Brandt, A., Fiege, D., ... others (2021). Natural history collections recapitulate 200 years of faunal change. *Royal Society open science*, 8(4), 201983.
- Fettke, P. (2009). How conceptual modeling is used. *Communications of the Association for Information Systems*, 25(1), 43.
- Fox, M. J. (2016). “Priorities of arrangement” or a “hierarchy of oppressions?”: Perspectives on intersectionality in knowledge organization. *Knowledge Organization*, 43(5), 373–383.
- 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.
- Franz, N. M., Pier, N. M., Reeder, D. M., Chen, M., Yu, S., Kianmajd, P., ... Ludäscher,

- B. (2016). Two influential primate classifications logically aligned. *Systematic biology*, 65(4), 561–582.
- Franz, N. M., & Sterner, B. W. (2018). To increase trust, change the social design behind aggregated biodiversity data. *Database*, 2018.
- GBIF.org. (2022). *Platalea minor Temminck & Schlegel, 1849 – GBIF occurrence download*. doi: 10.15468/dl.qtvbt7
- 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.
- 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., Mercurio, 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*.
- Girardello, M., Chapman, A., Dennis, R., Kaila, L., Borges, P. A., & Santangeli, A. (2019). Gaps in butterfly inventory data: a global analysis. *Biological Conservation*, 236, 289–295.
- Green, R. (2015). Indigenous peoples in the US, sovereign nations, and the DDC. *Knowledge Organization*, 42(4), 211–221.
- Gupta, S., Cheng, Y.-Y., & Ludäscher, B. (2019). Possible worlds explorer: Datalog & answer set programming for the rest of us. In *Datalog 2.0*.
- Gupta, S., & Ludäscher, B. (2019). *Leaneuler*. <https://github.com/idaks/LeanEuler>. Github Repository.
- Hajibayova, L., & Buente, W. (2017). Representation of indigenous cultures: Considering the Hawaiian hula. *Journal of Documentation*.
- Hajibayova, L., Buente, W., Quiroga, L., & Valeho-Novikoff, S. (2016). Representation of Kanaka Maoli (Hawaiian) culture: A case of hula dance. In *Proceedings of the 79th asis&t annual meeting: Creating knowledge, enhancing lives through information & technology* (p. 128).
- Harley, J. B. (1988). Silences and secrecy: the hidden agenda of cartography in early modern Europe. *Imago mundi*, 40(1), 57–76.
- Harley, J. B. (1992). Rereading the maps of the Columbian encounter. *Annals of the Association of American Geographers*, 82(3), 522–542.
- Hedden, H. (2016). *The accidental taxonomist*. Information Today, Inc.
- Hervitz, H. M. (1985). Selectivity, adaptation, or disruption? a comparison of alternative hypotheses on the effects of migration on fertility: The case of Brazil. *International Migration Review*, 19(2), 293–317.
- Hiemstra, N., & Mountz, A. (2011). A companion to social geography. In V. J. Del Casino Jr, M. Thomas, P. Cloke, & R. Panelli (Eds.), (chap. Geopolitics). John Wiley & Sons.
- Higgins, M. (2016). Totally invisible: Asian American representation in the Dewey Decimal Classification, 1876-1996. *Knowledge Organization*, 43(8), 609–621.
- Hobern, D., Baptiste, B., Copas, K., Guralnick, R., Hahn, A., van Huis, E., ... others

- (2019). Connecting data and expertise: a new alliance for biodiversity knowledge. *Biodiversity data journal*, 7.
- Hodge, G. (2000). *Systems of knowledge organization for digital libraries: Beyond traditional authority files*. ERIC.
- Horton, C. (2019). *Taiwan's president, defying xi jinping, calls unification offer 'impossible'*. <https://www.nytimes.com/2019/01/05/world/asia/taiwan-xi-jinping-tsai-ing-wen.html>. New York Times.
- Humpage, L. (2008). Revision required: reconciling New Zealand citizenship with Māori nationalisms. *National Identities*, 10(3), 247–261.
- Inants, A., & Euzenat, J. (2015). An algebra of qualitative taxonomical relations for ontology alignments. In *International semantic web conference* (pp. 253–268).
- Jabareen, Y. (2009). Building a conceptual framework: philosophy, definitions, and procedure. *International journal of qualitative methods*, 8(4), 49–62.
- Jarvis, B. (2017). Who Decides Who Counts as Native American? *The New York Times*. <https://www.nytimes.com/2017/01/18/magazine/who-decides-who-counts-as-native-american.html>.
- Jett, J., Sacchi, S., Lee, J. H., & Clarke, R. I. (2016). A conceptual model for video games and interactive media. *Journal of the Association for Information Science and Technology*, 67(3), 505–517.
- Johnston, M. A. (2016). Redefinition of the Eleodes Eschscholtz subgenera Tricheleodes Blaisdell and Pseudeleodes Blaisdell, with the description of a new species (Coleoptera: Tenebrionidae). In *Annales zoologici* (Vol. 66, pp. 665–679).
- Jung, J. J. (2006). Taxonomy alignment for interoperability between heterogeneous digital libraries. In *International conference on asian digital libraries* (pp. 274–282).
- Jung, J. J. (2008). Taxonomy alignment for interoperability between heterogeneous virtual organizations. *Expert Systems with Applications*, 34(4), 2721–2731.
- Kam, D. V. (2007). Subject headings for aboriginals: The power of naming. *Art Documentation: Journal of the Art Libraries Society of North America*, 26(2), 18–22.
- 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.
- 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.
- Koenig, A., & Stein, J. (2008). Federalism and the state recognition of Native American tribes: a survey of state-recognized tribes and state recognition processes across the United States. *Santa Clara L. Rev.*, 48, 79.
- Lampland, M., & Star, S. L. (2009). *Standards and Their Stories: How Quantifying, Classifying, and Formalizing Practices Shape Everyday Life*. Cornell University Press.
- Lapp, H., Morris, R. A., Catapano, T., Hobern, D., & Morrison, N. (2011). Organizing our knowledge of biodiversity. *Bulletin of the American Society for Information Science and Technology*, 37(4), 38–42.

- Lee, D. (2011). Indigenous knowledge organization: A study of concepts, terminology, structure and (mostly) indigenous voices. *Partnership: The Canadian Journal of Library and Information Practice and Research*, 6(1).
- Lee, K. J., Roldan, W., Zhu, T. Q., Kaur Saluja, H., Na, S., Chin, B., . . . Yip, J. (2021). The show must go on: A conceptual model of conducting synchronous participatory design with children online. In *Proceedings of the 2021 chi conference on human factors in computing systems* (pp. 1–16).
- Lee, L.-H., Yu, Y.-T., & Huang, C.-R. (2009). Chinese wordnet domains: Bootstrapping chinese wordnet with semantic domain labels. In *Proceedings of the 23rd pacific asia conference on language, information and computation, volume 1* (pp. 288–296).
- Lehmann, F. (1992). Semantic networks. *Computers & Mathematics with Applications*, 23(2-5), 1–50.
- Lemke, T. (2001). “The birth of bio-politics”: Michel Foucault’s lecture at the Collège de France on neo-liberal governmentality. *Economy and society*, 30(2), 190–207.
- Li, J.-m., Zhao, H.-z., & Du, M.-j. (2011). Detection and resolution of structural conflictions in heterogeneous xbrl taxonomies. In *The 5th international conference on new trends in information science and service science* (Vol. 2, pp. 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.
- Lilley, S. C. (2015). Ka Pō, Ka Ao, Ka Awatea: The interface between epistemology and Māori subject headings. *Cataloging & Classification Quarterly*, 53(5-6), 479–495.
- Lin, H., Liu, Y., Zhang, P., & Wang, J. (2019). Representation learning of taxonomies for taxonomy matching. In *International conference on computational science* (pp. 383–397).
- Littletree, S., Belarde-Lewis, M., & Duarte, M. (2020, July). Centering Relationality: A conceptual model to advance indigenous knowledge organization practices. *Knowledge Organization*, 47(5), 410–426. doi: 10.5771/0943-7444-2020-5-410
- Littletree, S., & Metoyer, C. A. (2015). Knowledge organization from an indigenous perspective: The Mashantucket Pequot thesaurus of American Indian terminology project. *Cataloging & Classification Quarterly*, 53(5-6), 640–657.
- Lu, M.-T. (2019). *Why do we classify indigenous peoples as mountain people and plain people, when more than half of them live in the cities [translated title from traditional chinese]*. <https://www.storm.mg/lifestyle/1831008>. The Storm Media.
- Mai, J.-E. (2016). Marginalization and exclusion: Unraveling systemic bias in classification. *Knowledge Organization*, 43(5), 324–330.
- Malavasi, M. (2020, December). The map of biodiversity mapping. *Biological Conservation*, 252, 108843. doi: 10.1016/j.biocon.2020.108843
- 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.
- McCulloch, A. M., & Wilkins, D. E. (1995). "Constructing" Nations within States: The

- Quest for Federal Recognition by the Catawba and Lumbee Tribes. *American Indian Quarterly*, 19(3), 361.
- McLaughlin, K. (2019). A big moment finally comes for the Little Shell: Federal recognition of their tribe. *The Washington Post*. [https://www.washingtonpost.com/national/a-big-moment-finally-comes-for-the-little-shell-federal-recognition-of-their-tribe/2019/12/20/72f5ee86-204d-11ea-bed5-880264cc91a9\\_story.html](https://www.washingtonpost.com/national/a-big-moment-finally-comes-for-the-little-shell-federal-recognition-of-their-tribe/2019/12/20/72f5ee86-204d-11ea-bed5-880264cc91a9_story.html).
- Montenegro, M. (2019). Subverting the universality of metadata standards: The TK labels as a tool to promote indigenous data sovereignty. *Journal of Documentation*.
- Mouriño-García, M., Pérez-Rodríguez, 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. In *Ontology matching* (p. 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.
- News, B. (2018). *U.s. airlines alter websites to meet china's demand on taiwan's name*. <https://www.bloomberg.com/news/articles/2018-07-25/u-s-airlines-alter-websites-to-meet-china-demand-on-taiwan-name>. Bloomberg.
- 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.
- of Congress, L. (2008). *Marc to dublin core crosswalk*. <https://www.loc.gov/marc/marc2dc.html>. The Library of Congress.
- Oksanen, M., & Vuorisalo, T. (2019). Conservation sovereignty and biodiversity. In *From assessing to conserving biodiversity* (pp. 435–452). Springer, Cham.
- Olson, H. A. (2002). *The power to name: locating the limits of subject representation in libraries*. Springer Science & Business Media.
- Olson, H. A. (2017). Social influences on classification. In *Encyclopedia of library and information sciences* (pp. 4204–4211). CRC Press.
- Otero-Cerdeira, L., Rodríguez-Martínez, F. J., & Gómez-Rodríguez, 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. In *International joint conference on knowledge discovery, knowledge engineering, and knowledge management* (pp. 39–55).
- Pfeifer, K., & Peukert, E. (2013b). Mapping text mining taxonomies. In *Kdir/kmis* (pp. 5–16).
- Plumer, B., Popovich, N., & Renault, M. (2020). *How racist urban planning left some neighborhoods to swelter*. <https://www.nytimes.com/2020/08/26/climate/racist-urban-planning.html>. New York Times.
- Ponzetto, S. P., & Navigli, R. (2009). Large-scale taxonomy mapping for restructuring and

- integrating wikipedia. In *Ijcai* (Vol. 9, pp. 2083–2088).
- Purves, R. S., Winter, S., & Kuhn, W. (2019). Places in information science. *Journal of the Association for Information Science and Technology*, 70(11), 1173–1182.
- Quinn, S. D., & Tucker, D. A. (2017). How geopolitical conflict shapes the mass-produced online map. *First Monday*.
- Rabinowitz, A., Shaw, R., Buchanan, S., Golden, P., & Kansa, E. (2016). Making sense of the ways we make sense of the past: The PeriodO project. *Bulletin of the Institute of Classical Studies*, 59(2), 42–55.
- Ragni, M., & Wöfl, S. (2005). Temporalizing spatial calculi: On generalized neighborhood graphs. In *Annual conference on artificial intelligence* (pp. 64–78).
- Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4), 334–350.
- Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. *KR*, 92, 165–176.
- Raunich, S., & Rahm, E. (2014). Target-driven merging of taxonomies with ATOM. *Information Systems*, 42, 1–14.
- Remsen, D. (2016). The use and limits of scientific names in biological informatics. *ZooKeys*, 550, 207.
- Renz, J. (2002). *Qualitative spatial reasoning with topological information*. Springer.
- Reynaud, C., & Safar, B. (2006). When usual structural alignment techniques don't apply. In *Ontology matching* (p. 191).
- Robertson, T., Döring, M., Guralnick, R., Bloom, D., Wieczorek, J., Braak, K., . . . Desmet, P. (2014). The GBIF integrated publishing toolkit: facilitating the efficient publishing of biodiversity data on the internet. *PloS one*, 9(8), e102623.
- Ross, R. S. (2002). Navigating the Taiwan Strait: Deterrence, escalation dominance, and US-China relations. *International Security*, 27(2), 48–85.
- Rundstrom, R. A. (1995). GIS, indigenous peoples, and epistemological diversity. *Cartography and geographic information systems*, 22(1), 45–57.
- Shaw, H.-Y. (1999). The Diaoyutai/Senkaku Islands dispute: its history and an analysis of the ownership claims of the PRC, ROC, and Japan. *Maryland Series in Contemporary Asian Studies*, 1999(3), 1.
- 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.
- Simon, S. (2010). Negotiating power: Elections and the constitution of indigenous taiwan. *American Ethnologist*, 37(4), 726–740.
- 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.
- Smith, J. A., Benson, A. L., Chen, Y., Yamada, S. A., & Mims, M. C. (2020). The power, potential, and pitfalls of open access biodiversity data in range size assessments: Lessons from the fishes. *Ecological Indicators*, 110, 105896.
- Soeller, G., Karahalios, K., Sandvig, C., & Wilson, C. (2016). Mapwatch: Detecting and

- monitoring international border personalization on online maps. In *Proceedings of the 25th international conference on world wide web* (pp. 867–878).
- 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*. Recuperado de <http://citeseerx.ist.psu.edu/viewdoc/summary>.
- Sowa, J. F. (1992). Conceptual graphs as a universal knowledge representation. *Computers & Mathematics with Applications*, 23(2-5), 75–93.
- Star, S. L., & Bowker, G. (2007). Enacting silence: Residual categories as a challenge for ethics, information systems, and communication. *Ethics and Information Technology*, 9(4), 273–280.
- Stewart, R. N., Piburn, J. O., Sorokine, A., Myers, A. T., & White, D. A. (2015). World Spatiotemporal Analytics and Mapping Project (WSTAMP): Discovering, exploring, and mapping spatiotemporal patterns across the world’s largest open source geographic data sets. *ISPRS Annals of Photogrammetry*.
- 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.
- Sung, Y.-H., Tse, I. W.-L., & Yu, Y.-T. (2018). Population trends of the Black-faced Spoonbill *Platalea minor*: analysis of data from international synchronised censuses. *Bird Conservation International*, 28(1), 157–167.
- Templeman, K. (2018). When do electoral quotas advance indigenous representation?: Evidence from the Taiwanese legislature. *Ethnopolitics*, 17(5), 461–484.
- Tennis, J. (2007, April). Scheme Versioning in the Semantic Web. *Cataloging & Classification Quarterly*, 43(3-4), 85–104. (Publisher: Routledge \_eprint: [https://doi.org/10.1300/J104v43n03\\_05](https://doi.org/10.1300/J104v43n03_05)) doi: 10.1300/J104v43n03\_05
- Tennis, J. (2010). *Measured Time: Imposing a Temporal Metric to Classificatory Structures* (SSRN Scholarly Paper No. ID 2879000). Rochester, NY: Social Science Research Network.
- Tennis, J. (2012). The strange case of eugenics: A subject’s ontogeny in a long-lived classification scheme and the question of collocative integrity. *Journal of the American Society for Information Science and Technology*, 63(7), 1350–1359.
- Tennis, J. (2013, September). *Metaphors of Time and Installed Knowledge Organization Systems: Ouroboros, Architectonics, or Lachesis?* (SSRN Scholarly Paper No. ID 2879089). Rochester, NY: Social Science Research Network.
- Tennis, J., Thornton, K., & Filer, A. (2012, February). Some temporal aspects of indexing and classification: toward a metrics for measuring scheme change. In *Proceedings of the 2012 iConference* (pp. 311–316). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2132176.2132216
- Thau, D. (2010). *Reasoning about taxonomies* (Doctoral Dissertation). University of California Davis.
- Thau, D., Bowers, S., & Ludäscher, B. (2008). Merging taxonomies under RCC-5 algebraic articulations. In *Proceedings of the 2nd international workshop on ontologies and information systems for the semantic web* (pp. 47–54).

- Thau, D., Bowers, S., & Ludäscher, B. (2009). CleanTax: A framework for reasoning about taxonomies. In *Aaai spring symposium: Benchmarking of qualitative spatial and temporal reasoning systems* (pp. 49–50).
- Thau, D., & Ludäscher, B. (2007). Reasoning about taxonomies in first-order logic. *Ecological Informatics*, 2(3), 195–209.
- Thomer, A., Cheng, Y.-Y., Schneider, J., Twidale, M., & Ludäscher, B. (2017). Logic-based schema alignment for natural history museum databases. *Knowledge Organization*, 44(7), 545–558.
- Trist, E. L. (1981). *The evolution of socio-technical systems* (Vol. 2). Ontario Quality of Working Life Centre Toronto.
- Tsai, Y.-Y. (2017). Becoming “Thao” : Genetic science, identity construction and the thao aboriginal movement. *Taiwanese Journal of Sociology*(62), 113–175.
- Van Hooland, S., & Verborgh, R. (2014). *Linked data for libraries, archives and museums: How to clean, link and publish your metadata*. Facet publishing.
- Wand, Y., & Weber, R. (2002). Research commentary: information systems and conceptual modeling—a research agenda. *Information systems research*, 13(4), 363–376.
- Wang, Z., Lv, Q., Lan, X., & Zhang, Y. (2018). Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 349–357).
- Webster, K., & Doyle, A. (2008). Don’t class me in antiquities! giving voice to Native American materials. *Radical cataloging: Essays at the front*, 189–197.
- Wee, S.-L. (2018). *Giving in to china, u.s. airlines drop taiwan (in name at least)*. <https://www.nytimes.com/2018/07/25/business/taiwan-american-airlines-china.html>. New York Times.
- Woods, W. A. (1975). What’s in a link: Foundations for semantic networks. In *Representation and understanding* (pp. 35–82). Elsevier.
- Xanthaki, A., & O’Sullivan, D. (2009). Indigenous participation in elective bodies: The Maori in New Zealand. *International Journal on Minority and Group Rights*, 16(2), 181–207.
- Xu, S., & Sun, M. (2007). Leveraging world knowledge in chinese text classification. In *Sixth international conference on advanced language processing and web information technology (alpit 2007)* (pp. 33–38).
- Yeung, C.-L., Yao, C.-T., Hsu, Y.-C., Wang, J.-P., & Li, S.-H. (2006). Assessment of the historical population size of an endangered bird, the black-faced spoonbill (*Platalea minor*) by analysis of mitochondrial DNA diversity. *Animal Conservation*, 9(1), 1–10.
- Yip, J. C., Lee, K. J., & Lee, J. H. (2020). Design partnerships for participatory librarianship: A conceptual model for understanding librarians co designing with digital youth. *Journal of the Association for Information Science and Technology*, 71(10), 1242–1256.
- Yu, S. (2017). *Enhancing taxonomy alignment with region connection calculus reasoning approach* (Doctoral Dissertation). University of California Davis.
- Zeng, M. L. (2008). Knowledge organization systems (KOS). *Knowledge Organization*, 35(2-3), 160–182.
- Zeng, M. L., & Chan, L. M. (2006). Metadata interoperability and standardization-a study

- of methodology, part II. *D-Lib Magazine*, 12(6), 1082–9873.
- Zeng, M. L., & Chan, L. M. (2010). Semantic interoperability. In *Encyclopedia of library and information sciences* (pp. 4645–4662).
- Zhitomirsky-Geffet, M. (2019). Towards a diversified knowledge organization system: An open network of inter-linked subsystems with multiple validity scopes. *Journal of Documentation*.

# APPENDIX A

All the data artifacts from this dissertation can be retrieved at the following online links:

- Chapter 5 Data artifacts for Reconciling the Socio-Geographic in the United States Maps: <https://doi.org/10.5281/zenodo.6625833>
- Chapter 6 Data artifacts for Reconciling the Socio-Geographic in Country Taxonomies: <https://doi.org/10.5281/zenodo.6625835>
- Chapter 7 Data artifacts for Reconciling the Socio-Geographic in Indigenous Peoples' Tribes: <https://doi.org/10.5281/zenodo.6625839>
- Chapter 8 Data artifacts for Reconciling the Socio-Geographic in Biodiversity Data and Information Systems: <https://doi.org/10.5281/zenodo.6632129>