Snailed it! Merging Taxonomically Organized Biodiversity Datasets with Shifting Geopolitical Realities

Yi-Yun Cheng[1], Steven Dilliplane[2], Bertram Ludäscher[1]

[1]School of Information Sciences, University of Illinois at Urbana-Champaign; [2] The Academy of Natural Sciences of Drexel University

Taxonomy; Biodiversity Informatics; Geopolitical Realities

PROBLEMS

- Difficulties in maintaining a seamless and explicit navigation among biodiversity, taxonomically organized datasets & Natural History Museum Literature (NHM) (Page, 2013, 2019)
- Concerns about the quality of aggregated biodiversity data from data integration services such as GBIF (Franz & Sterner, 2017)
- Dataset distributors usually provide a more Westernized view of documentation that has overlooked some of the geopolitical realities in other regions of the world (Boakes et al., 2010; Harris & Froufe, 2005; Karl & Bowen, 1999)

OUR GOALS

- Provides a more precise approach to merge taxonomically organized datasets that contain *region sovereignty* changes over time
- Serves as a first step towards bridging NHM literature and biodiversity datasets

USE CASE

GEOGRAPHIC POINT OF INTEREST: TAIWAN

- Taiwan has been historically complex in terms of sovereignty, or geopolitical realities

SPECIES: *Pupinella swinhoei* sec. H. Adams 1866

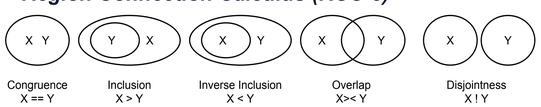
- A land snails species endemic to Taiwan and Japan

WHAT QUESTIONS TO ANSWER?

- What is the historical distributions of species such as the land snails?
- Are they endemic to Taiwan, Japan, or other locations?
- What is different from the 1905 historical text on such species and now?
- Can we leverage the 1905 historical texts to enrich species descriptions?

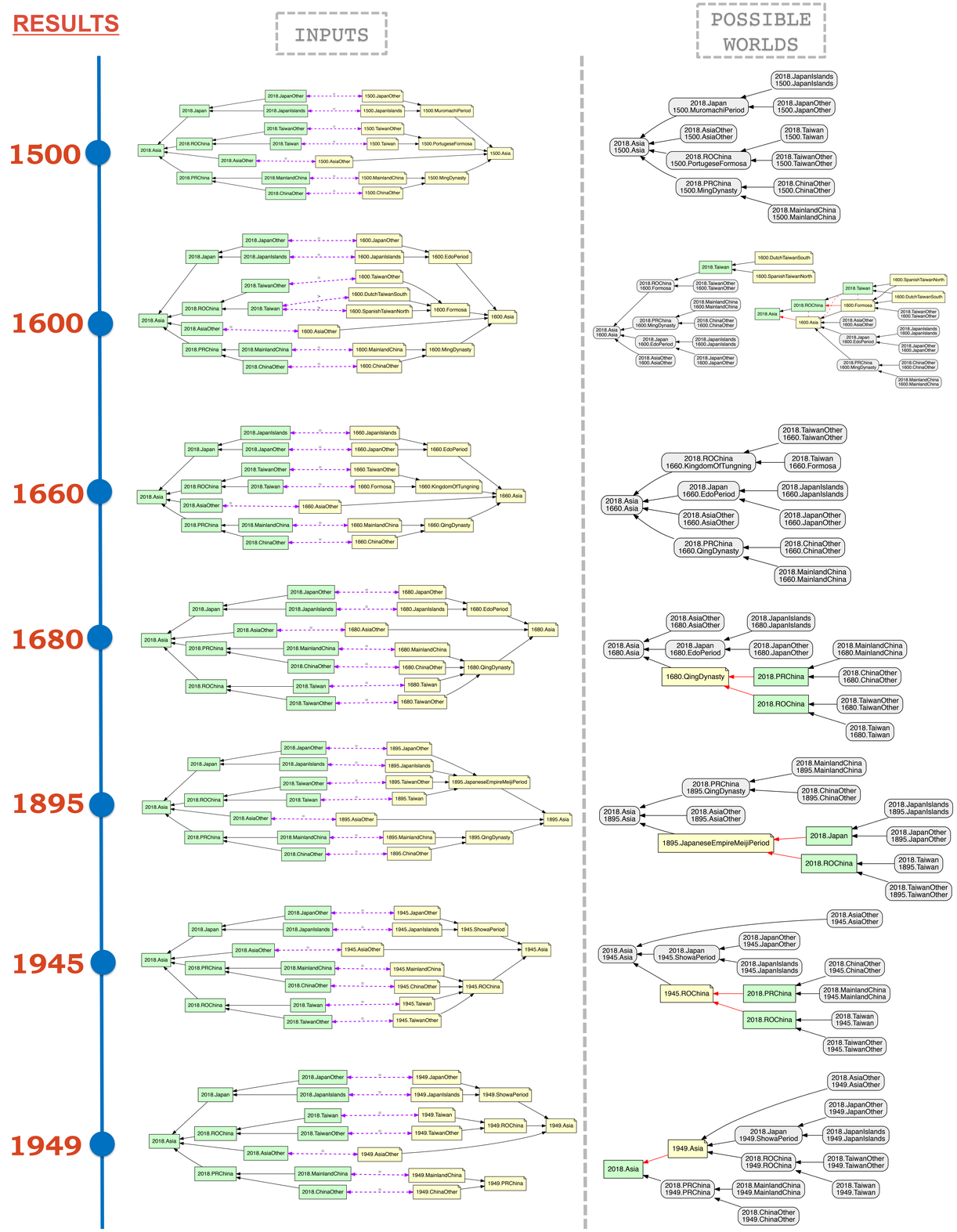
METHOD: TAXONOMY ALIGNMENT

- **Taxonomy Alignment Problems (TAP) :** Taxonomies T_1, T_2 are inter-linked via a set of input *articulations A* to yield a "merged" taxonomy T_3
- ***Articulations****:* a constraint or rule that defines a relationship (a set constraint) between two concepts from different taxonomies
- ***Region Connection Calculus (RCC-5)***



- ***Possible Worlds*** – When encoding and solving TAPs via ASP, the different answer sets represent alternative taxonomy merge solutions or possible worlds (PWs).

RESULTS



DATA SOURCES

NATURAL HISTORY MUSEUM LITERATURE

- From the Biodiversity Heritage Library (BHL)
- *Proceedings of the Academy of Natural Sciences of Philadelphia*
- 1905 articles on "*Catalogue of the land and fresh-water Mollusca of Taiwan (Formosa*), *with descriptions of new species"*

SPECIES OCCURRENCES DATASETS

- From Global Biodiversity Information Facility (GBIF)
- On *Pupinella swinhoei* : 50 occurrences from 18 datasets of different sources, ranging from the year 1700 to now

Data Source Key*	Institution Code	Scientific Name	Country Code	Locality	Year
1	MCZ	*Pupinella swinhoei* H.Adams, 1866	TW	Formosa	1700
1	MCZ	*Pupinella swinhoei* H.Adams, 1866	TW	Hotawa	1700
2	NSSM	*Pupinella swinhoei* H.Adams, 1866	TW	Hualien	1939
2	NSSM	*Pupinella swinhoei* H.Adams, 1866	TW	DawuDahu	1928
...					
18	TOYA	*Pupina adamsi* Sowerby, 1878	JP		1991

*Dataset key was modified from a 32 digit/letter combination code to a number.

Example of our target modified datasets

After the taxonomy alignment approach on the 'countries' with geopolitical realities, we can provide another value-added column showing the true historical sovereignties

Data Source Key*	Institution Code	Scientific Name	Sec. Author	Country Code	Year	Historical Sovereignty
1	MCZ	*Pupinella swinhoei*	H.Adams 1866	TW	1700	Qing Dynasty China
1	MCZ	*Pupinella swinhoei*	H.Adams 1866	TW	1700	Qing Dynasty China
2	NSSM	*Pupinella swinhoei*	H.Adams 1866	TW	1939	Japan
2	NSSM	*Pupinella swinhoei*	H.Adams 1866	TW	1928	Japan
...						
18	TOYA	*Pupina adamsi*	Sowerby 1878	JP	1991	Japan

PRACTICAL IMPLICATIONS

- The alignments of species name and historical sovereignties may aid the creation of a data-driven *knowledge graph* for a particular species
- Species phenotypes, traits, habitat information can then be added to enrich the data that were not included in a Darwin-Core formatted occurrence dataset

ACKNOLWEDGEMENT

- This project is the outcome of **the 2019 LEADS-4-NDP fellowship program**. The authors wish to thank the program organizers, specifically Dr. Jane Greenberg, and Sam Grabus for their continuous support.

CONTACTS

yiyunyc2@illinois.edu
https://github.com/EulerProject/IDCC20