Systematic Comparison of Data Models used in Mapping Knowledge Organization Systems

ILLINOIS
School of Information Sciences

Yi-Yun Cheng
School of Information Sciences, University of Illinois at Urbana-Champaign

yiyunyc2 [at] illinois.edu
@yiyunjessica

DATA MODEL

- The definition of data models varies across disciplines, but the commonality is that they represent raw data in a structured manner.
- Established data models includes tabular model, tree model, relational model, and RDF model.
- Here we follow Van Hooland & Verborgh's (2014) definition of data models in the context of metadata research:

" Data Models embodies the meaning of the data in its most essential and stripped down form."

PROBLEM STATEMENT

- Data models are abstract models used in the context of designing Knowledge Organization Systems (KOS) such as metadata schemas, taxonomies, and ontologies.
- Efforts have been made to examine various mapping methods (Chan & Zeng, 2006; Shvaiko, & Euzenat, 2005), or evaluate the effectiveness of these methods (Avesani, Giunchiglia, & Yatskevich, 2005; Euzenat et al., 2001).
- The impact of **choosing data models in the context of mapping KOS** has not been explicitly discussed.

AIM AND RESEARCH QUESTIONS

In this exploratory study, we **examine six KOS mapping projects and the data models used** in their studies. The six different projects are selected to represent a diverse set of KOS:

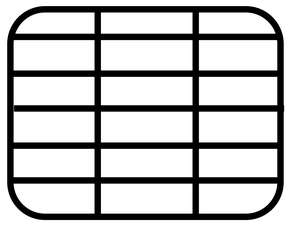
- mapping **metadata schemas**: Harping et al., 2000; Library of Congress, 2008
- mapping **taxonomies**: Franz et al., 2016; Cheng and Ludäscher, 2020
- mapping **ontologies**: Raunich & Rahm, 2014; Jung, 2008

Aim of research: compare the data models used across mapping projects.

RQ1: What kinds of information are made explicit/implicit in the data model used in these projects?
RQ2: What are the potential information gains/losses from the data model used in these projects?

Contributions: We hope the questions asked in this study can contribute to future mapping projects in the Information Science community and aid the researchers to choose the optimal data models fitting their goals.

TABULAR MODEL



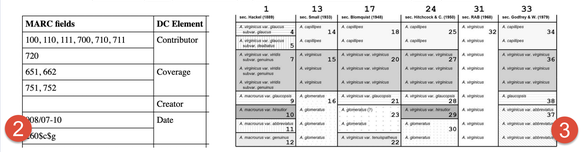
Tables is a direct and effective way to map different schemas or taxonomies, especially when there are numerous KOS being compared.

Explicit/implicit information from Tabular Models

- **All-in-one:** tables explicitly aggregate and present all related standards and taxonomies about the same subject in one place.
- **Ambiguity on what the relations are:** side-by-side representation of elements in columns implies that two elements are equivalent, but they may be overlapping, one subsuming the other, or near-equivalent
- **Ambiguity on what is being compared**: whether it is (1) a pair-wise comparison between schemas of adjacent columns; or (2) a comparison between the main schema in the first column versus the rest.

Information gains/losses from Tabular Models

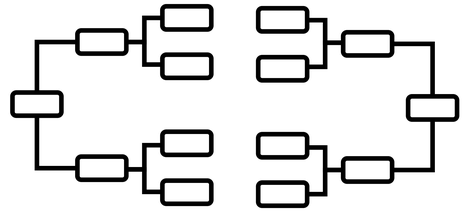
- **Information on the sub-elements:** the sub-elements (or children nodes) of the original schemas and the nested elements are not shown directly on the crosswalking table. Further labels needed to be included to indicate the structure of the original schemas
- **Granularity differences results in redundancies:** mapping a more granular schema (e.g., CDWA) to a more general schema (e.g., Dublin Core), redundancies would occur in the table where multiple elements are mapped to one same element



MARC fields	DC Element
100, 110, 111, 700, 710, 711	Contributor
720	
651, 662	Coverage
751, 752	
	Creator
008/07-10	Date
260cg	

Screenshots of:
1. Getty Research Institute project to crosswalk 15 different art-related metadata standards
2. Library of Congress's MARC to Dublin Core Crosswalk
3. Franz et al. (2016)'s study on aligning 11 biological taxonomies of Andropogon complex across 126 years

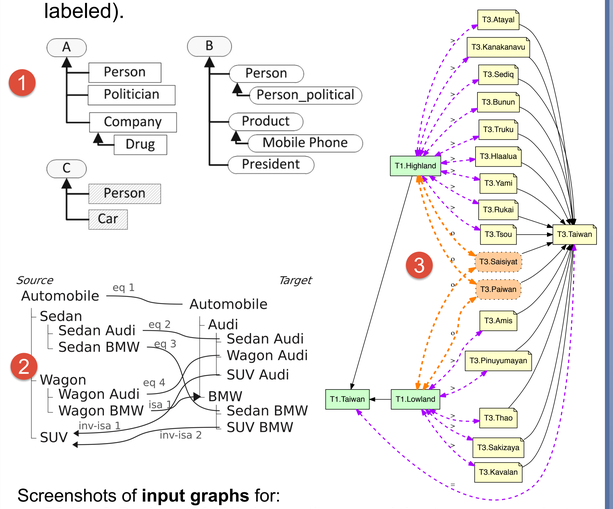
TREE MODEL



Tree models are often used to place hierarchies beside each other and connect the links between the two hierarchies to depict the **'before-merging' view (input)** of the project.

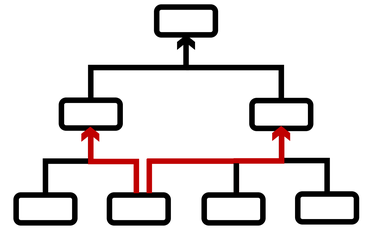
Explicit/implicit information from Tree Models

- **Maintaining structure:** tree models explicitly show how many levels of nodes there are, what the root node, parent nodes, sibling nodes, and children nodes are.
- **Manifesting the relations**: it is clear which level of nodes are we comparing, and what the relations between the concepts are.
- **Symmetry**: it is implicitly that if a tree structure is used, the two taxonomies being compared are symmetrically mapped. There're no source or target taxonomies unless further labeled.
- **Node and Edge type:** it is also assumed to the nodes and edges in both taxonomies should be of exactly same type (e.g. all the lines are *is-a* edges, though not explicitly labeled).



Screenshots of **input graphs** for:
1. Pfeifer & Peukert (2013)'s integrating text mining taxonomy project
2. Raunich & Rahm (2014)'s ATOM project
3. Cheng & Ludäscher (2020)'s Indigenous Taiwan project

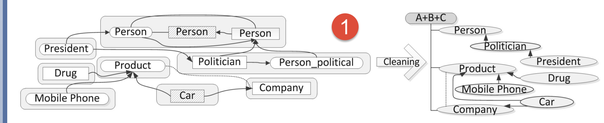
DIRECTED ACYCLIC GRAPHS (DAGs)



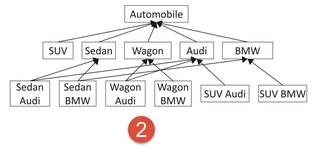
The **'merged'/'mapped'/output KOS** are usually in a special form of tree models – Directed Acyclic Graphs (DAGs) – which a node can have more than one parent nodes.

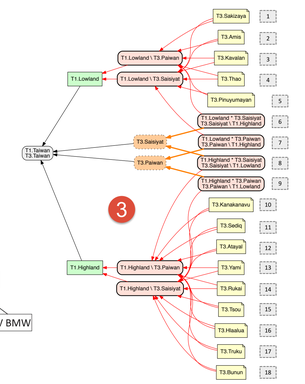
Information gains/losses from Tree Models (DAGs)

- **Information on the original KOS:** output DAGs can blur the structure and content of the original KOS, so it would be difficult to trace back how the input graphs look like.



Screenshots of **output DAGs** for:
1. Pfeifer & Peukert (2013)'s integrating text mining taxonomy project
2. Raunich & Rahm (2014)'s ATOM project
3. Cheng & Ludäscher (2020)'s Indigenous Taiwan project





DISCUSSION AND CONCLUSION

- The choice of using a certain data model over another in KOS mapping project may not be a random act, but one could choose it without realizing the implicit and explicit information a data model can signify.
- Both models can potentially obscure the information of the original KOS. This makes it problematic if we want to reconstruct a mapping project based on the outcomes manifested in either model.
- The list of data models is not exhaustive. Other alternative data models such as relational model or semantic model may exist in other mapping projects –which means information losses can be amplified.
- Data models are abstractions of real-world entities, so they may not represent what the actual mapping tools look like, but rather a way to help us conceptualize how two or more KOS relate to each other.

ACKNOWLEDGEMENT

The author would like to thank Dr. Bertram Ludäscher and Ly Dinh for their feedback on this study. The author would also like to extend their gratitude to the Center for Informatics Research in Science and Scholarship (CIRSS) at U of I for their generous travel support and funding for this work.